Volume 23, Number 1, March 2022 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

Sen A., Lahiri P., Estimation of mask effectiveness perception for small domains using multiple data sources

Yabaci A., Sigirli D., Comparison of tree-based methods used in survival data

Chaudhuri A., Samaddar S., Estimating the population mean using a complex sampling design dependent on an auxiliary variable

Goldoust M., Mohammadpour A., Generalized extended Marshall-Olkin family of lifetime distributions

Kowalczyk B., Wieczorkowski R., New improved Poisson and negative binomial item count techniques for eliciting truthful answers to sensitive questions

Oluyede B., Moako T., Chipepa F., The odd power generalized Weibull-G power series class of distributions: properties and applications

Abu-Shawiesh M. O. A., Sinsomboonthong J., Golam Kibria B. M., A modified robust confidence interval for the population mean of distribution based on deciles

Chaturvedi A., Kumar S., Estimation procedures for reliability functions of Kumaraswamy-G Distributions based on Type II Censoring and the sampling scheme of Bartholomew

Janus J., Long-term sovereign interest rates in Czechia, Hungary and Poland: a comparative assessment with an affine term structure model

Yasmeen U., Noor-ul-Amin M., Hanif M., Variance estimation in stratified adaptive cluster sampling

Akeem Ajibola Adepoju A. A., Abdulkadir S. S., Danjuma J., Chiroma H., Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart

Islamiyati A., Raupong , Anisa K., Sari U., Estimation the confidence interval of the regression coefficient of the blood sugar model through multivariable linear spline with known variance

EDITOR

Włodzimierz Okrasa — *University of Cardinal Stefan Wyszyński, Warsaw and Statistics Poland, Warsaw, Poland*
e-mail: w.okrasa@stat.gov.pl; phone number +48 22 – 608 30 66

EDITORIAL BOARD

Dominik Rozkrut (Co-Chairman) — *Statistics Poland, Warsaw, Poland*
Waldemar Tarczyński (Co-Chairman) — *University of Szczecin, Szczecin, Poland*
Czesław Domański — *University of Łódź, Łódź, Poland*
Malay Ghosh — *University of Florida, Gainesville, USA*
Graham Kalton — *University of Maryland, College Park, USA*
Mirosław Krzyśko — *Adam Mickiewicz University in Poznań, Poznań, Poland*
Partha Lahiri — *University of Maryland, College Park, USA*
Danny Pfeffermann — *Central Bureau of Statistics, Jerusalem, Israel*
Carl-Erik Särndal — *Statistics Sweden, Stockholm, Sweden*
Jacek Wesołowski — *Statistics Poland, and Warsaw University of Technology, Warsaw, Poland*
Janusz L. Wywiał — *University of Economics in Katowice, Katowice, Poland*

ASSOCIATE EDITORS

Arup Banerji	*The World Bank, Washington, USA*	Andrzej Młodak	*Statistical Office Poznań, Poznań, Poland*
Misha V. Belkindas	*ODW Consulting, Washington D.C., USA*	Colm A. O'Muircheartaigh	*University of Chicago, Chicago, USA*
Sanjay Chaudhuri	*National University of Singapore, Singapore*	Ralf Münnich	*University of Trier, Trier, Germany*
Eugeniusz Gatnar	*National Bank of Poland, Warsaw, Poland*	Oleksandr H. Osaulenko	*National Academy of Statistics, Accounting and Audit, Kiev, Ukraine*
Krzysztof Jajuga	*Wrocław University of Economics, Wrocław, Poland*	Viera Pacáková	*University of Pardubice, Pardubice, Czech Republic*
Alina Jędrzejczak	*University of Łódź, Poland*	Tomasz Panek	*Warsaw School of Economics, Warsaw, Poland*
Marianna Kotzeva	*EC, Eurostat, Luxembourg*	Mirosław Pawlak	*University of Manitoba, Winnipeg, Canada*
Marcin Kozak	*University of Information Technology and Management in Rzeszów, Rzeszów, Poland*	Mirosław Szreder	*University of Gdańsk, Gdańsk, Poland*
Danute Krapavickaite	*Institute of Mathematics and Informatics, Vilnius, Lithuania*	Imbi Traat	*University of Tartu, Tartu, Estonia*
Martins Liberts	*Central Statistical Bureau of Latvia, Riga, Latvia*	Vijay Verma	*Siena University, Siena, Italy*
Risto Lehtonen	*University of Helsinki, Helsinki, Finland*	Gabriella Vukovich	*Hungarian Central Statistical Office, Budapest, Hungary*
Achille Lemmi	*Siena University, Siena, Italy*	Zhanjun Xing	*Shandong University, Shandong, China*

FOUNDER/FORMER EDITOR

Jan Kordos — *Warsaw School of Economics, Warsaw, Poland*

EDITORIAL OFFICE **ISSN 1234-7655**

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

© Copyright by Polish Statistical Association, Statistics Poland and the authors, some rights reserved. CC BY-SA 4.0 licence

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

CONTENTS

From the Editor III

Submission information for authors IX

Invited papers

Sen A., Lahiri P., Estimation of mask effectiveness perception for small domains using multiple data sources 1

Research articles

Yabaci A., Sigirli D., Comparison of tree-based methods used in survival data 21

Chaudhuri A., Samaddar S., Estimating the population mean using a complex sampling design dependent on an auxiliary variable 39

Goldoust M., Mohammadpour A., Generalized extended Marshall-Olkin family of lifetime distributions 55

Kowalczyk B., Wieczorkowski R., New improved Poisson and negative binomial item count techniques for eliciting truthful answers to sensitive questions 75

Oluyede B., Moako T., Chipepa F., The odd power generalized Weibull-G power series class of distributions: properties and applications 89

Abu-Shawiesh M. O. A., Sinsomboonthong J., Golam Kibria B. M., A modified robust confidence interval for the population mean of distribution based on deciles 109

Chaturvedi A., Kumar S., Estimation procedures for reliability functions of Kumaraswamy-G Distributions based on Type II Censoring and the sampling scheme of Bartholomew 129

Janus J., Long-term sovereign interest rates in Czechia, Hungary and Poland: a comparative assessment with an affine term structure model 153

Yasmeen U., Noor-ul-Amin M., Hanif M., Variance estimation in stratified adaptive cluster sampling 173

Akeem Ajibola Adepoju A. A., Abdulkadir S. S., Danjuma J., Chiroma H., Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart 185

Research Communicates and Letters

Islamiyati A., Raupong , Anisa K., Sari U., Estimation the confidence interval of the regression coefficient of the blood sugar model through multivariable linear spline with known variance 201

Conference announcement

The 3rd Congress of Polish Statistics will take place from 26 to 28 April 2022 in Cracow, Poland 213

The IAOS 2022 Conference will take place from 26 to 28 April 2022 at the Convention Center in Cracow, Poland 214

Special issue announcement

A New Role for Statistics: The Joint Special Issue of "Statistics in Transition new series" and "Statystyka Ukraïny", June 2022 215

About the Authors 217

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. III–VIII

From the Editor

It is with great pleasure that we can announce the upgrading of Statistics in Transition new series (SiTns) in terms of points allocated by the Ministry of Education and Science to scientific journals, and by the same token, to the articles published in SiTns – from 40 to 70 points.

Also, the formal and legal relations with authors publishing in our journal will change as well – namely papers submitted from now on will be made available in the framework of Creative Commons Attribution-ShareAlike 4.0 (CC BY-SA 4.0) free licences. The practical implication of this is that the authors will retain all their copyrights, and the readers will be able to use the work according to the provisions of the licence. Authors will be granting the licence to the publisher of SiTns in a statement submitted along with the paper (the new statement form can be downloaded from https://sit.stat.gov.pl/ForAuthors).

The presented March issue contains 12 articles by authors from 13 countries: Turkey, India, Iran, Poland, Botswana, Jordan, Thailand, USA, Pakistan, Canada, Nigeria, Saudi Arabia, and Indonesia. We are convinced that such a geographic diversity adds to the value of also the thematically diversified problems discussed in the articles presented in our journal.

Invited paper

The issue starts with the Invited paper entitled ***Estimation of mask effectiveness perception for small domains using multiple data sources*** by **Aditi Sen** and **Partha Lahiri**. The paper discusses the impacts of pandemics on public health and related societal issues. Due to the fact that mask wearing is one of the few precautions against COVID-19, the authors develop a synthetic estimation method to estimate proportions of perceived mask effectiveness for small area using a logistic model that combines information from multiple data sources. The authors select the working model using an extensive data analysis facilitated by a new variable selection criterion for survey data and benchmarking ratios and propose a jackknife method to estimate variance of their proposed estimator. From the data analysis, it is evident that the proposed synthetic method outperforms direct survey-weighted estimator with respect to commonly used evaluation measures. To quantify people's perception of mask effectiveness and to prevent the spread of COVID-19 for small areas, the authors use Understanding America Study's (UAS) survey data on COVID-19 as the primary data source. The issue

of mask effectiveness perception is a critical one and helps in understanding the future impacts or spread of the disease at the state level. Wearing masks is undoubtedly one of the few and most effective precautionary measures.

Research articles

Aysegul Yabaci and **Deniz Sigirli** in their article ***Comparison of tree-based methods used in survival data*** present survival trees and forests as the popular non-parametric alternatives to parametric and semi-parametric survival models. The Conditional inference trees (Ctree), Conditional inference forests (Cforest) and Random survival forests (RSF) methods are discussed in detail and the performances of the survival forest methods, namely the Cforest and RSF have been compared with a simulation study. The results of the simulation demonstrate that the RSF method with a log-rank score distinction criteria outperforms the Cforest and the RSF with log-rank distinction criteria. As a result, it has been shown that the RSF method performs better than the Cforest. For both methods, it can be said that the Aalen estimator performs better than the other estimators. The performance of both methods was better when the proportional hazard assumption was not provided. In addition, the RSF method shows that the logrank distinction criteria, which is one of two different separation criteria, performs better than the logrank score distinction criteria.

The paper entitled ***Estimating the population mean using a complex sampling design dependent on an auxiliary variable*** by **Arijit Chaudhuri** and **Sonakhya Samaddar** starts with a view that the simplest strategy to estimate a population total without bias is to employ Simple Random Sampling (SRS) with replacement (SRSWR) and the expansion estimator based on it. Anything other than that including SRS Without Replacement (SRSWOR) and usage of the expansion estimator is a complex strategy. In the paper, the authors examine if from a complex sample at hand a gain in efficiency may be unbiasedly estimated comparing the "rival population total-estimators" for the competing strategies and how suitable model-expected variances of rival estimators compete in magnitude as examined numerically through simulations.

Mehdi Goldoust and **Adel Mohammadpour** discuss ***Generalized extended Marshall-Olkin family of lifetime distributions***. The authors introduce a new generalized class of lifetime distributions, called the LPS2 family of distributions, by compounding a lifetime and twice power series distributions in a serial and parallel structure. The new models extend several distributions widely used in the lifetime literature such as the exponential power series, Weibull power series, and complementary of exponential power series distributions. A mathematical treatment of the new distributions is provided, including ordinary and incomplete moments, quantile, moment generating and mean residual functions. The maximum likelihood

estimation technique is used to estimate the model parameters and a simulation study is conducted to investigate the performance of the maximum likelihood estimates. The authors perform a Monte Carlo simulation study to assess the finite sample behaviour of the maximum likelihood estimators. Some members of the LPS2 family are fitted to two real data sets to illustrate the usefulness of the new distributions. They provide better fits than other competing models consistently.

Barbara Kowalczyk's and **Robert Wieczorkowski's** article ***New improved Poisson and negative binomial item count techniques for eliciting truthful answers to sensitive questions*** is devoted to demonstrating how Item Count Techniques (ICTs) – pioneered by Miller – are working in the context of indirect survey questioning methods designed to deal with sensitive features. These techniques have gained the support of many applied researchers and encountered further theoretical development. The two new item count methods called Poisson and negative binomial ICTs were also proposed. However, if the population parameters of the control variable are not given from the outside source, the methods are not very efficient. In the paper the authors analyse best linear unbiased and maximum likelihood estimators of the population proportion of the sensitive attribute in the new introduced models. Theoretical results presented in the manuscript are supported by a comprehensive simulation study. The improved procedure allowed increasing efficiency of the estimation as compared to the original Poisson and negative binomial ICTs. In the article three new models are proposed: Poisson-Poisson neutral questions ICT, Poisson-negative binomial neutral questions ICT, and negative binomial-negative binomial neutral questions ICT. Newly proposed methods maintain privacy of respondents at the same level regarding the sensitive question. At the same time the three newly proposed techniques increase efficiency of the estimation, which is very important in indirect methods of questioning.

Broderick Oluyede, Thatayaone Moakofi, and **Fastel Chipepa** discuss a new class of distributions in the paper entitled ***The odd power generalized Weibull-G power series class of distributions: properties and applications***. The authors develop a new class of distributions, namely the odd power generalized Weibull-G power series (OPGW-GPS) class of distributions and present some special classes of the proposed distribution. Structural properties have also been derived. The authors conducted a simulation study to evaluate the consistency of the maximum likelihood estimates. Moreover, two real data examples on selected data sets were provided to illustrate the usefulness of the new class of distributions. The proposed model outperforms several non-nested models on selected data sets. Furthermore, from the results shown in the manuscript, the authors conclude that the OPGW-WP distribution is indeed a better model compared to several selected models since it is associated with the lowest values for all the the goodness-of-fit statistics (and the largest p-value for the K-S statistic).

In the next paper, ***A modified robust confidence interval for the population mean of distribution based on deciles,*** a new approach to estimating the population mean of a skewed distribution is considered by **Moustafa Omar Ahmed Abu-Shawiesh, Juthaphorn Sinsomboonthong**, and **B. M. Golam Kibria**. Acknowledging that the confidence interval is an important statistical estimator used to estimate the population location and dispersion parameters, the authors look for a robust modified confidence interval building upon an adjustment of the Student's t confidence interval based on the decile mean and decile standard deviation for estimating the population mean of a skewed distribution. The efficiency of the proposed interval estimator is evaluated using an extensive Monte Carlo simulation study. The coverage ratio and average width of the proposed confidence interval are compared with some existing, widely used confidence intervals. The simulation results show that, in general, the proposed interval estimator performs significantly well. For illustration purposes, three real-life data sets are analyzed, which support the findings obtained from the simulation study to some extent. Consequently, the authors recommend practitioners using the robust modified confidence interval for estimating the population mean when the data is generated by a normal or skewed distribution.

The paper entitled ***Estimation procedures for reliability functions of Kumaraswamy-G Distributions based on Type II Censoring and the sampling scheme of Bartholomew*** by **Aditi Chaturvedi** and **Surinder Kumar** discusses Kumaraswamy-G distributions and derive a Uniformly Minimum Variance Unbiased Estimator (UMVUE) and a Maximum Likelihood Estimator (MLE) of the two measures of reliability, namely $R(t) = P(X > t)$ and $P = P(X > Y)$ under Type II censoring scheme and sampling scheme of Bartholomew (1963). Authors also develop interval estimates of the reliability measures. A comparative study of the different methods of point estimation has been conducted on the basis of simulation studies. An analysis of a real data set has been presented for illustration purposes. The paper focuses on developing classical estimators for different parameters and reliability functions of Kumaraswamy-G distributions under various sampling schemes and investigating their properties. However, an interesting alternative to MLE and UMVU estimators can be provided by the empirical Bayes approach or ML-II estimators based on the robust Bayesian approach of Shrivastava et al..

The paper by **Jakub Janus** entitled ***Long-term sovereign interest rates in Czechia, Hungary and Poland: a comparative assessment with an affine term structure model*** provides a comparative evaluation of the behaviour of long-term sovereign yields in Czechia, Hungary and Poland from 2001 to 2019. An affine term structure model developed by Adrian, Crump and Moench (2013) is used as an empirical framework for the decomposition of the bond yields into term premium and risk-neutral

components. The paper aimed to examine the sovereign 10-year bond yields in three Central European economies: Czechia, Hungary, and Poland, from 2001 to 2019. The ACM term structure to extract time-varying risk-neutral and term premium components were developed. The evolution of these components, along with their relative role in driving the actual interest rates were discussed. The international comovements of 10-year yields between the CE economies and Germany was studied, and as an extension the baseline term structure model was corrected.

Uzma Yasmeen, Muhammad Noor-ul-Amin, and **Muhammad Hanif** in their article focus on ***Variance estimation in stratified adaptive cluster sampling***. In many sampling surveys, the use of auxiliary information at either the design or estimation stage, or at both these stages is usual practice. Auxiliary information is commonly used to obtain improved designs and to achieve a high level of precision in the estimation of population density. Adaptive cluster sampling (ACS) was proposed to observe rare units with the purpose of obtaining highly precise estimations of rare and specially clustered populations in terms of least variances of the estimators. This sampling design proved to be more precise than its more conventional counterparts, including simple random sampling (SRS), stratified sampling, etc. In this paper, a generalised estimator is anticipated for a finite population variance with the use of information of an auxiliary variable under stratified adaptive cluster sampling (SACS). The bias and mean square error expressions of the recommended estimators are derived up to the first degree of approximation. A simulation study showed that the proposed estimators have the least estimated mean square error under the SACS technique in comparison with variance estimators in stratified sampling.

The last article prepared by **Akeem Ajibola Adepoju, Sauta S. Abdulkadir, Danjuma Jibasen**, and **Haruna Chiroma** propose ***Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart***. The paper aims to develop an Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart (IT2FEWMA) under the fuzzy type-2 condition. This development will facilitate monitoring small and moderate shifts in the production process in conditions of uncertainty. The manuscript extends the control limits of the classical control chart of the exponentially weighted moving average (EWMA). The IT2FEWMA is advantageous over the classical EWMA due to its flexibility over the control limits, but it is not capable of detecting a big shift in the process due to the fact that classical EWMA does not have such capacity too. This article is a new addition to the existing Statistical Process Control Tools. It is useful when the process engineer needs to monitor a process whose measurement is obtained in fuzzy environment and a small shift needs to be detected.

Research Communicates and Letters

The Research Communicates & Letters section presents a paper by **Anna Islamiyati, Raupong, Anisa Kalondeng**, and **Ummi Sari** entitled ***Estimating the confidence interval of the regression coefficient of the blood sugar model through a multivariable linear spline with known variance***. The estimates from confidence intervals are more powerful than point estimates, because there are intervals for parameter values used to estimate populations. In relation to global conditions, involving issues such as type 2 diabetes mellitus, it is very difficult to make estimations limited to one point only. Therefore, in this article, the authors estimate confidence intervals in a truncated spline model for type 2 diabetes data. They use a non-parametric regression model through a multi-variable spline linear estimator. The use of the model results from the irregularity of the data, so it does not form a parametric pattern. Subsequently, the authors obtained the interval from beta parameter values for each predictor. Body mass index, HDL cholesterol, LDL cholesterol and triglycerides all have two regression coefficients at different intervals as the number of the found optimal knot points is one. This value is the interval for multivariable spline regression coefficients that can occur in a population of type 2 diabetes patients.

Włodzimierz Okrasa
Editor

© Włodzimierz Okrasa. Article available under the CC BY-SA 4.0 licence

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, 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, March 2022
Vol. 23, No. 1, 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, March 2022
Vol. 23, No. 1, pp. 1–20, DOI 10.21307/stattrans-2022-001
Received – 01.01.2022; accepted – 24.02.2022

Estimation of mask effectiveness perception for small domains using multiple data sources

Aditi Sen[1], **Partha Lahiri**[2]

ABSTRACT

Understanding the impacts of pandemics on public health and related societal issues at granular levels is of great interest. COVID-19 is affecting everyone in the globe and mask wearing is one of the few precautions against it. To quantify people's perception of mask effectiveness and to prevent the spread of COVID-19 for small areas, we use Understanding America Study's (UAS) survey data on COVID-19 as our primary data source. Our data analysis shows that direct survey-weighted estimates for small areas could be highly unreliable. In this paper, we develop a synthetic estimation method to estimate proportions of perceived mask effectiveness for small areas using a logistic model that combines information from multiple data sources. We select our working model using an extensive data analysis facilitated by a new variable selection criterion for survey data and benchmarking ratios. We suggest a jackknife method to estimate the variance of our estimator. From our data analysis, it is evident that our proposed synthetic method outperforms the direct survey-weighted estimator with respect to commonly used evaluation measures.

Key words: cross-validation, jackknife, survey data, synthetic estimation.

1. Introduction

Mask effectiveness perception is a topic of great relevance during the COVID-19 pandemic with emergence of new variants, multiple waves and fluctuating infection r ates. In the United States, national estimates of mask effectiveness perception can be derived by weighted means or proportions from respondent level data from a national survey like the Understanding America Study. However, to draw conclusions for small areas (e.g., states) for which sample sizes are small, direct estimates are inappropriate and misleading with very low or high estimates and highly variable standard errors.

In this paper, we explore a synthetic estimation of the perception on mask effectiveness, i.e., proportion of people considering mask to be highly effective at the state level. The is an indirect method of borrowing strength from similar areas. A synthetic estimator is not area specific in the study variable of interest and can be applied to any probability and non-probability sample design. Such methods are often employed in practice for their simplicity and ability to produce estimates for areas with no sample from the sample survey. Moreover, when the survey does not provide any sample for many areas, a synthetic method may be appealing to public policy makers as the same estimation method is applied to all areas, irrespective of whether an area has sample or not. There is a widespread use of synthetic

[1]PhD Student, Applied Mathematics & Statistics, and Scientific Computation. E-mail: asen123@umd.edu
[2]Director and Professor, The Joint Program in Survey Methodology & Professor, Department of Mathematics, University of Maryland, College Park, MD 20742, USA. E-mail: plahiri@umd.edu

© Aditi Sen, Partha Lahiri. Article available under the CC BY-SA 4.0 licence

estimation in different small area applications; e.g. Ghosh (2020), Marker (1995) and others. A synthetic method uses explicit or implicit models to link several disparate databases in producing efficient estimates for small areas. Hansen et al. (1953) presented an early example of a regression method to produce synthetic estimates of the median number of radio stations heard during the day for over 500 counties of the United States. Stasny et al. (1991) developed a regression-synthetic method for estimating county acreage of wheat using a non-probability sample of farms along with auxiliary data on planted acreage and district indicators. Marker (1999) and Rao and Molina (2015) presented more examples of synthetic small area estimators based on regression models. For our problem, we combine UAS data with the census data and Covid Tracking Report data to develop synthetic estimates of mask effectiveness perception for the states.

In Section 2, we describe primary and supplementary data used in this paper. In Section 3, we evaluate performances of the state level direct survey-weighted estimates. The performance of the direct method is poor, which motivates synthetic estimation, described in Section 4. In this Section, we introduce a JACKKNIFE method to estimate variance of the synthetic estimator. We report main results from our data analysis in Section 5. In this Section, we introduce a new model selection criterion for complex survey data. Finally, we evaluate synthetic estimates by comparative analogy of plotting with direct estimates for a handful of states, some small like District of Columbia, Rhode Island, North Dakota and large states like New York, California, Florida. We conclude the paper by summarizing the utility of the methods described in the paper and discussing how they can be extended to any other binary, categorical or continuous variable from this survey or any other survey with little adjustments or modifications.

2. Data used

For this study, we will use the UAS as the primary data containing study variable on the perception of mask effectiveness and supplementary data containing information for building small domain modelling and estimation procedures.

2.1. The Primary Data: Understanding America Study (UAS)

The Understanding America Study (UAS), conducted by the University of Southern California (USC), is an internet panel of households representing the entire United States. A household is broadly defined as anyone living together with the person who signed up for participating in the UAS. Using members of the population-representative UAS panel, USC's Center for Economic and Social Research (CESR) launched the Understanding Coronavirus in America tracking survey on March 10, 2020. The survey provides useful information on attitudes, behaviours, including health care avoidance behaviour, mental health, personal finances around the novel coronavirus pandemic in the United States.

Initial requests were sent out to the UAS panel members in order to determine their willingness to participate in an ongoing Coronavirus of UAS surveys. Among 9,063 UAS panel members who responded to the initial request, 8,547 were found eligible to participate in the survey. On average until November, 2020 (wave 16) about six thousand respondents

Table 1: Understanding of America Survey (UAS) wave details

Wave Number	Wave Name	Time period	Sample size
1	UAS 230	March 10, 2020 - March 31, 2020	6,932
2	UAS 235	April 1, 2020 - April 28, 2020	5,478
3	UAS 240	April 15, 2020 - May 12, 2020	6,287
4	UAS 242	April 29, 2020 - May 26, 2020	6,403
5	UAS 244	May 13, 2020 - June 9, 2020	6,407
6	UAS 246	May 27, 2020 - June 23, 2020	6,408
7	UAS 248	June 10, 2020 - July 8, 2020	6,346
8	UAS 250	June 24, 2020 - July 22, 2020	6,077
9	UAS 252	July 8, 2020 - Aug 5, 2020	6,289
10	UAS 254	July 22, 2020 - Aug 19, 2020	6,371
11	UAS 256	Aug 5, 2020 - Sep 2, 2020	6,238
12	UAS 258	Aug 19, 2020 - Sep 16, 2020	6,284
13	UAS 260	Sep 2, 2020 - Sep 30, 2020	6,284
14	UAS 262	Sep 16, 2020 - Oct 14, 2020	6,129
15	UAS 264	Sep 30, 2020 - Oct 27, 2020	6,181
16	UAS 266	Oct 14, 2020 - Nov 11, 2020	6,181

participate in the surveys, as seen from sample size in Table 1. Beginning in March 2020, the first round was UAS 230, which fielded from March 10 to March 31, 2020, with most responses happening during the period of March 10-14, 2020. UAS 230 is the first round of the survey that includes questions specifically tailored to COVID-19. These questions were repeated in subsequent longitudinal waves. The survey is being conducted in multiple waves. As of November 11, 2020,there are 16 waves, as described in Table 1 with their time periods.

For each wave, eligible panel members are randomly assigned to respond on a specific day so that a full sample is invited to participate over a 14-day period. Respondents have 14 days to complete the survey but receive an extra monetary incentive for completing the survey on the day they are invited to participate. Thus, except for the first wave, the data collection period for each wave is four weeks with a two-week overlap between any two consecutive waves. Each wave data consists of, on an average, six thousand observations.

The UAS is sampled in batches, through address-based sampling. The batches are allocated for national estimation and also for special population estimation (Native Americans, California, and Los Angeles county). Essentially UAS is a multiple-frame survey with four frames: Nationally Representative Sample, Native Americans, Los Angeles (LA) County, and California. Table 2 shows the relationship between the batches and frames, but each batch draws from only one frame.

As of November 2020, there are 21 batches, the latest being added in August, 2020. Most batches use a two-stage probability sample design in which zip codes are drawn first and then households are drawn at random from the sampled zip codes (except for two small sub-groups that are simple random samples from lists). The National batches draw zip

Table 2: Relationship between Batches and Frames in the Understanding America Survey

Batch	Frame
1	U.S.
2,3	Native American
4	Los Angeles County young mothers
5 to 12	U.S.
13,14,18,19	Los Angeles County
15,16	California
17,20,21	U.S.

codes without replacement, but the Los Angeles County batches draw with replacement and do sometimes contain the same zip code in different batches.

The base weights account for the differential probability of sampling a zip-code and an address within it. The base weights are then adjusted for nonresponse. Finally, at the national level, the distribution of nonresponse adjusted weights is calibrated to that of the 2018 Current Population Survey (CPS) weights with respect to selected demographic variables. Weights are provided for all batches, except for batch 4, which comprises Los Angeles County young mothers, and non-Native American households in batches 2 and 3. Angrisani et al. (2019) describe the sampling and weighting for UAS in great detail.

The survey includes a national bi-weekly long-form questionnaire and a weekly Los Angeles County short-form questionnaire administered in each bi-weekly wave. The survey data contains information on different demographic variables such as age, race, sex, and Hispanic origin, education, marital status, work status, identifiers for the states and zip-codes, and various outcome variables affecting human lives (e.g., mental stress, personal finances, COVID-19 like symptoms, testing results, etc.) The data also contains base and final weights so survey-weighted direct estimates for different outcome variables of interest can be produced.

2.2. Supplementary Data

The COVID Tracking Project: Both national and state level data can be downloaded from https://covidtracking.com. We use the data as a source of state specific auxiliary variables in our models. The COVID Tracking Project collects and publishes testing data daily for the United States as a whole and also for states and territories. From this data we understand that for 50 states and the District of Columbia (DC) combined the total test count has been increasing fast with more than 1 million in April 2020 to close to total 200 million by end of November 2020. The daily test count also increased from around 180K in April 2020 to 1.5 million in November 2020. There are various state specific auxiliary variables that could be potentially predictive of the perception on mask effectiveness. They include COVID-19 daily total testing, total test results (positive/negative), death, recovery count (as obtained from Johns Hopkins data on coronavirus), hospitalization, ventilation, etc. With

the increase in tests (due to better supply of testing kits, increase in awareness, etc.) or increase in positive cases (due to mask mandate being relaxed, advent of a new variant, etc.), one may argue that people's perceptions of mask effectiveness may change. Thus for this study, we use the following auxiliary variables that could be potentially useful in explaining our outcome variable on perception of mask effectiveness:

(i) totalTestResults: total number of tests with positive or negative results,

(ii) positive: total number of positive tests.

To make the above two auxiliary variables comparable across 50 states and the District of Columbia, we have used appropriate scaling factors to create the following two auxiliary variables, which we have used in our modelling:

(i) Testing rate: (Total tests with positive or negative results)/(Total population of state),

(ii) Positivity rate: (Total positive tests)/ (Total tests with positive or negative results).

Population density data: We use population density estimates in our modelling. Population density estimates for US states in 2010 are obtained from the U.S. Census Bureau (2020). For this study, we have created a categorical variable from it with three levels as follows:

- low - when population density of a state is less than or equal to 101 people per square mile (1st quartile from 2010 census data), e.g. North Dakota, Wyoming, Alaska etc.,
- medium - when population density is greater than 101 but less than or equal to 231 people per square mile (median), e.g. Georgia, Michigan, Virginia, etc.,
- high - when population density is greater than 231 people per square mile, e.g. New York, California, District of Columbia, etc.

Democratic party affiliation: For a given state, we have created a binary variable, which takes on the value 1 if the Governor of the state is a Democrat and 0 otherwise. The information is prior to the 2020 election and obtained from Wikipedia (2020).

Region membership of the states: Since 1950, the United States Census Bureau defines four statistical regions, with nine divisions. Using information obtained from the Census Bureau (2010) we have marked each state as one of the four regions - Northeast, Midwest, South and West.

Census Bureau's Population Estimates Program (PEP): For our synthetic estimation method, we need population counts for different demographic groups in the 50 states and the District of Columbia. The Census Bureau releases various tables of population estimates. On June 2020, the Population Division of the U.S. Census Bureau released annual state

resident population estimates by age, sex, race, and Hispanic origin for the period April 1, 2010 to July 1, 2019. The Census Bureau essentially obtains these estimates using the 2010 decennial census as the base and updates by births, deaths, migration etc. available from the administrative records and others obtained from the American Community Survey (ACS) survey. We have used two data sources as follows:

1. SCPRC-EST2019-18+POP-RES: Estimates of the Resident Population Age 18 Years and Older for the US states from July 1, 2019 (released on December 2019), which can be directly used.

2. SC-EST2019-ALLDATA5: Estimates of population by Age, Sex, Race, and Hispanic Origin – 5 race groups (5 race alone or in combination groups). This data needs to be adjusted by filtering out 18+ population (with "AGE") for the above-mentioned domains (using variables "RACE" for white and rest as other race and "ORIGIN" for Hispanic or Non-Hispanic). Sex is not used, although present in the data and hence set to value 0 for all. The domain wise populations are then adjusted with a factor (i.e., multiplying with domain wise population/total state population) so that the sum of all the domains is equal to the total state level estimate mentioned before.

3. Direct estimation

For the mask effectiveness perception problem, we focus on the following question from survey questionnaire: *How effective is wearing a face mask such as the one shown here for keeping you safe from coronavirus?* This is a categorical variable with five possible answers: *(i) Extremely Ineffective, (ii) Somewhat Ineffective, (iii) Somewhat Effective, (iv) Extremely Effective, and (v) Unsure.* The answer choices of respondents have been used to create a binary variable that takes on the value 1 is taken if mask is considered to be Extremely Effective by respondent and 0 otherwise. Using this binary variable the direct estimate works really well at overall national level with low standard error.

The survey data contains respondents residing in 50 states and DC, but naturally they are not evenly distributed. For larger states like California or Florida, there is a sizable volume in the sample of even as high as 2000 respondents and for smaller states like Delaware or Wyoming, there is very little representation of even 3 or 4 respondents. In such scenarios, direct survey-weighted estimates are highly misleading. For example, we see for the first three waves 0% of people in Wyoming think mask is extremely effective, which happens because all the respondents in the sample take the value 0 for binary response variable perceived mask effectiveness. Hence this is not a good method to draw a conclusion for the whole population of the states.

We observe extremely variable standard error (SE) or margin of error (ME). Estimated SE, or equivalently, estimated ME for a state depends on the sample size and the value of estimated proportion. For a state with small sample size, say less than 12, SE is either 0 or very high. From computations of direct estimates, from multiple waves we see that for Rhode Island, a state whose contribution in the wave is small with 2 or 3 respondents, estimated SE in the first few waves (1 and 2) is 0%. We obtain 0 SE when all the observations are the same. In the case of Rhode Island the cause is latter. But as soon as we have a mix

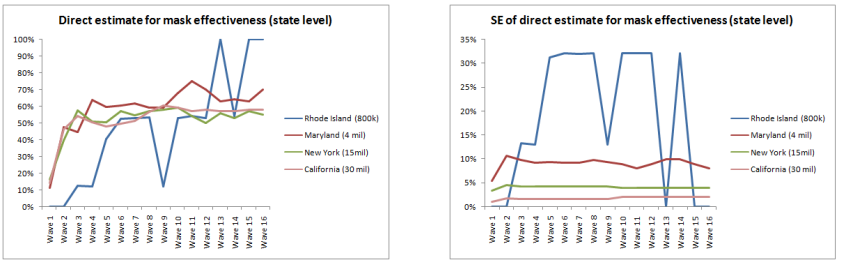


Figure 1: Direct estimates of perceived mask effectiveness and associated standard errors for four selected states.

of 0s and 1s, SE becomes very high, as high as even 30% from wave 5 onwards to wave 9 for Rhode Island.

Figure 1 displays erratic behaviour of direct estimates and standard errors for four states with varying population sizes (as estimated from the Census Bureau's PEP data)- one with high population (California - estimated adult population of 30 million from PEP), one with medium population (New York - estimated adult population of 15 million from PEP) – one with small population (Maryland - estimated adult population of 4 million from PEP) and one with very small population (Rhode Island - estimated adult population of 800k from PEP). The curves for Rhode Island are very unstable whereas, those for New York and California are quite stable. These SE estimates are thus surely very unstable or unreliable and typically, in public opinion polls margin of errors (2SE) is targeted at a low level such as 3% or 4%. Figure 1 for Rhode Island demonstrates high variability in state estimates for smaller states.

Along with high variability a demonstration of high bias in the direct state estimates can also be observed. Since we do not know the truth for perceived mask effectiveness, we cannot demonstrate bias properties for perceived mask effectiveness. But we can say if we consider another outcome variable for which "truth" is known from the PEP data, we can at least partially justify our claim. Using Figure 2 we show that UAS estimates of proportions of people falling in the four demographic groups or domains we considered do not match up with PEP data for states, but they more or less match at the national level. For large states like California, the difference between PEP estimate of the percentage of adult population and UAS direct wave estimate is negligible. This is similar for medium sized states like Maryland and New York, but for small states like Rhode Island and North Dakota, referring to Figure 2, we see the percentages vary significantly with even 0% or no contribution in some domains.

4. Synthetic method

For developing the synthetic estimation of perceived mask effectiveness for small areas, i.e., at state level, we first define the following notations and then derive a formula for the estimator from a logistic regression model. Let y_k denote the value of outcome (or

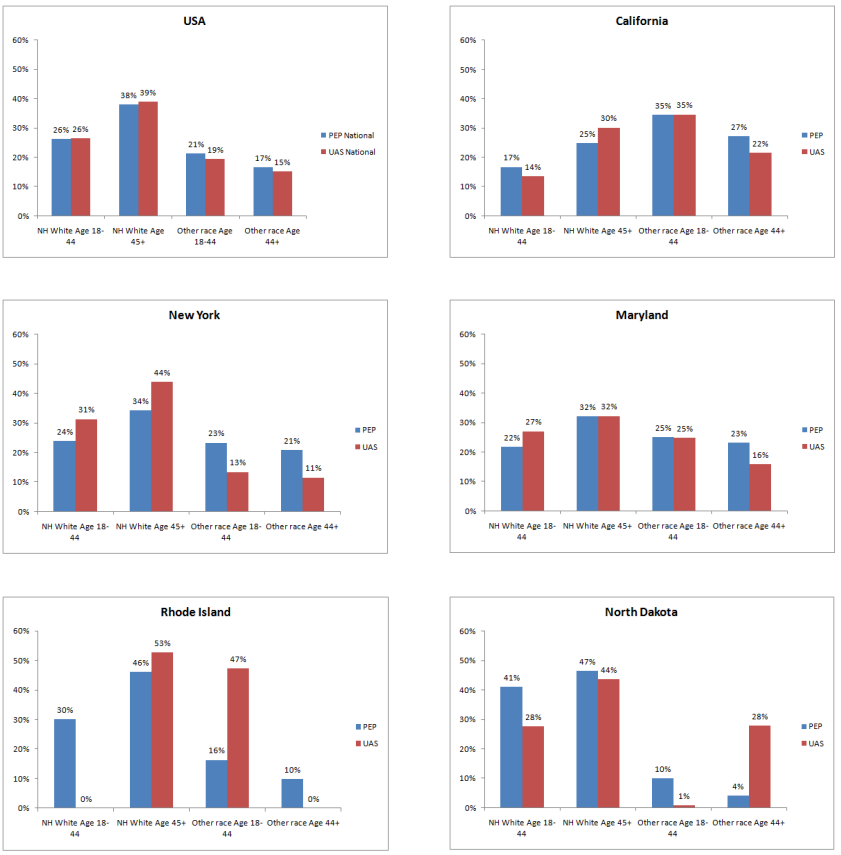


Figure 2: PEP vs. UAS estimates of 4 domains

dependent) variable for the kth respondent ($k = 1, \cdots, n$), where n denotes the number of respondents in a given wave (say, wave 2 covering April 1-April 28, 2020) of the UAS survey. The outcome variable is binary as defined by $y_k = 1$ if respondent k considers mask wearing to be extremely effective. Let $x_k = (x_{k1}, \cdots, x_{kp})'$ denote the value of a vector of auxiliary variables (same as independent variables or predictor variables or covariates) for respondent k. We have focused on the following two criteria for selecting the auxiliary variables for the unit level logistic regression model: (i) auxiliary variables should have good explanatory power in explaining the outcome variable of interest; (ii) total or mean of these auxiliary variables for the population should be available from a big data such as a large survey, administrative records or decennial census. Let N_i and N_{gi} be the population size of the adult (18+) and the gth group in state i, respectively. As discussed previously in the data section N_{gi} and N_i values are obtained from the US Census Bureau. Let y_{gik} be the value of the outcome variable for kth respondent in state i for the gth group ($g = 1, \cdots, G$; $i = 1, \cdots, m$; $k = 1, \cdots, N_{ig}$). Here we have $m = 51$ (50 states and DC) small areas. Let z_i be a vector of state specific auxiliary variables. For the estimation of mask-effectiveness variable for the 50 states and the District of Columbia, we write the population model as:

$$\text{Level 1: } y_{gik}|\theta_{gi} \overset{ind}{\sim} f(.;\theta_{gi}), \quad \text{Level 2: } h(\theta_{gi}) = x_g'\beta + z_i'\gamma, \tag{1}$$

for $k = 1, \cdots, N_{gi}$, $g = 1, \cdots, G; i = 1, \cdots, m$, where $f(.;\theta_{gi})$ is a suitable distribution with parameter θ_{gi} (here for binary variable this is a Bernoulli distribution with success probability θ_{gi}); $h(\theta_{gi})$ is a suitable known link function (for this application, we take logit link); β and γ are unknown parameters to be estimated using UAS micro data, i.e., at the respondent or unit level using survey weights.

We estimate population mean for state i by: $\hat{\bar{Y}}_i^{syn} = \sum_{g=1}^{G}(N_{gi}/N_i)\hat{\theta}_{gi} = \sum_{g=1}^{G}(N_{gi}/N_i)$ $h^{-1}(x_g'\hat{\beta} + z_i'\hat{\gamma})$, where h^{-1} is the inverse function of h; $\hat{\beta}$ and $\hat{\gamma}$ are survey-weighted estimators of β and γ, respectively. If $h(\cdot)$ is a logit function, we have $\hat{\bar{Y}}_i^{syn} = \sum_{g=1}^{G}(N_{gi}/N_i)\hat{\theta}_{gi} = \sum_{g=1}^{G}(N_{gi}/N_i)\exp(x_g'\hat{\beta} + z_i'\hat{\gamma})\left[1 + \exp(x_g'\hat{\beta} + z_i'\hat{\gamma})\right]^{-1}$.

We propose a jackknife method to estimate the variance of the proposed synthetic estimator. We obtain jth jackknife resample by deleting all survey observations in batch j. Thus we have $m = 20$ jackknife resamples from wave 14 onwards because there are 20 batches in total, whereas earlier for waves 1 to 13 there were in total 19 batches in each wave data, the latest addition being "21 MSG 2020/08 Nat. Rep. Batch 11" in August 2020 and LA County Young mothers is not present in any of the waves. For each jackknife resample, we recompute replicate synthetic estimate using (1). We will get m such replicate estimates, say, $\hat{\bar{Y}}_{i(-j)}^{syn}$ ($j = 1, \cdots, m$). We can then estimate the variance of $\hat{\bar{Y}}_i^{syn}$ by

$$v(\hat{\bar{Y}}_i^{syn}) = \frac{m-1}{m}\sum_{j=1}^{m}\left(\hat{\bar{Y}}_{i(-j)}^{syn} - \frac{1}{m}\sum_{j=1}^{m}\hat{\bar{Y}}_{i(-j)}^{syn}\right)^2. \tag{2}$$

Table 3: Direct national estimates of perceived mask effectiveness (associated standard errors) for selected demographic groups and first five waves.

Direct Estimate	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5
Overall National	14% (0.6%)	41% (0.9%)	47% (0.9%)	46% (0.9%)	44% (0.9%)
NH White Age(18-44)	12% (1.1%)	33% (1.7%)	39% (1.6%)	37% (1.6%)	33% (1.6%)
NH White Age(45+)	11% (0.7%)	40% (1.2%)	46% (1.1%)	46% (1.1%)	44% (1.1%)
Other race Age(18-44)	20% (1.7%)	48% (2.6%)	54% (2.3%)	50% (2.3%)	49% (2.3%)
Other race Age(45+)	17% (1.7%)	47% (2.7%)	58% (2.5%)	58% (2.5%)	58% (2.4%)

5. Data analysis

At national level in order to understand the broader question on the identification of demographic factors influencing effectiveness perceptions certain domains or groups are created based on race-ethnicity x age. These four groups are Non Hispanic White Age 18-44, Non Hispanic White Age 45+, Other race Age 18-44 and Other race Age 45+. Considerable variation among these groups is observed across multiple waves with all standard errors (SE) from direct estimates around 2%, after which it is chosen for further estimation study. The direct survey-weighted estimates at the national level as well as domain level from waves 1 to 5 are provided in Table 3 along with the standard errors in parenthesis; see also Figure 3. We observe that the overall national estimate and the domain NH White Age(45+) behave similarly (e.g, 46% and 44% for waves 4 and 5, respectively). The Other Race Age (45+) domain has the highest perception of mask effectiveness (e.g., 58% for waves 4 and 5), whereas the domain NH White Age (18-44) has the least value of such estimate (e.g., 37% and 33% at wave 4 and 5, respectively). Thus this breakdown of the population into domains can be used further for modelling. We have used R **survey** package to compute such estimates with the weights of respondents as provided in the wave data. We refer to the papers by Lumley (2004, 2010, 2020) for understanding the R **survey** package.

From the aforementioned observations, it is clear that direct estimates are not stable even at the state level. The synthetic estimators essentially would borrow strength from other states through implicit or explicit models and combine information from the sample survey, various administrative/census records, or previous surveys. Synthetic estimators are highly effective and appealing in small area estimation. Referring to synthetic estimation methods explained in Lahiri and Pramanik (2019), we employ a unit level logistic model with respondent level characteristics like the age x race/ethnicity along with state level auxiliary variables such as regional identifier (e.g., Northeast, Midwest, South or West), party affiliation of state governor or DC mayor (Democratic or Republic) and even the state level COVID-19 testing or positivity rate. Thus we have combined the data in UAS survey with

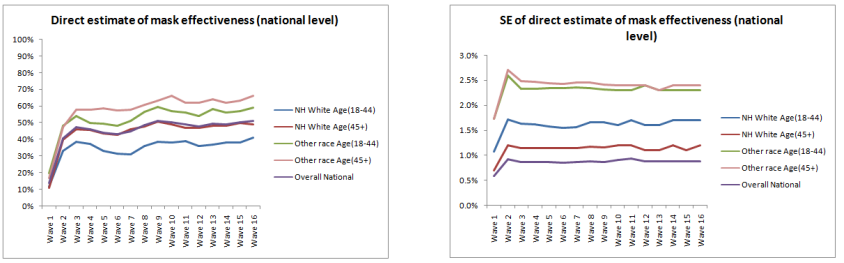


Figure 3: National direct wave estimates of perceived mask effectiveness and associated standard error direct estimates; overall estimates as well as estimates for four groups are provided.

the US Census Bureau data and Covid Tracking Project data to derive state level synthetic estimates of population means and totals for the variable of interest.

5.1. Variable Selection

For all the 16 waves, we first fit the full model, i.e., the model with all auxiliary variables listed earlier. Table 4 displays significant auxiliary variables in all the waves. We then concentrate our focus on the models given in Table 5. These are logistic regression models for the indicator response variable perceived mask effectiveness with different combinations of auxiliary variables. In every case, we use R survey package to run weighted logistic regression with quasi-Bernoulli family, where weights are the final post-stratification weights as provided by UAS and design is defined with such weights and no strata or cluster.

The full model, i.e., M1, is our starting point. True values of some of the coefficients of M1 may be zero; if the sample size is large, those coefficients will be estimated at near zero. But, if we keep too many covariates in a model, the estimates may be subject to high variability (and thereby we may lose some predictive power if we select a model with a lot of covariates.)

We now explore the possibility of reducing the number of auxiliary variables from M1. There are a large number of possible models so we proceed systematically. To this end, we fit M1 for all the 16 waves. Table 4 reports significant auxiliary variables for each of the 16 waves. In all the models, we include intercept (whether or not it is significant). Using information in Table 4, we create Table 5, which lists a number of competing models with less number of model parameters. We now explain why we want to consider models M1-M7 for further comparison.

All the auxiliary variables except for the democratic party affiliation appear in at least one wave. Thus, a natural question is what happens if we drop the democratic party affiliation from M1, which motivates keeping M2 for further investigation. Positivity rate is significant only in wave 5. This suggests inclusion of model M3 for further investigation. The factors NH Whites (18-44), NH Whites (45+), Other Race (18-44), population density are all significant for 5 waves: 6,7, 12, 14, and 16. So the model M7 seems to be a natural choice. We then consider models M3-6. Note that, in addition to NH Whites (18-44), NH

Table 4: Significant covariates in Model 1 for different waves; from R package output of significance code and p-value pairs to be interpreted as '***' for [0, 0.001], '**' for (0.001, 0.01], '*' for (0.01, 0.05], '•' for (0.05, 0.1], ' ' for (0.1, 1]

Wave	intercept	NH White Age(18-44) (indicator)	NH White Age(44+) (indicator)	Other race Age(18-44) (indicator)	testing rate	positivity rate	population density (categorical)	region Northeast (indicator)	region Midwest (indicator)	region South (indicator)	Democratic party (indicator)
1	***	*	***						•	**	
2		***	*		•						
3		***	***				***				
4		***	***	*			***			**	
5	*	***	***	*		**	**	•			
6		***	***	*			***				
7		***	***	*			**				
8		***	***				***		*		
9		***	***		•		***				
10		***	***	*	*		**				
11		***	***	•			***		•	•	
12		***	***	*			***				
13		***	***	•	*		**			**	
14		***	***	•			***				
15		***	***	•					•		
16		***	***	*			***				

Table 5: A list of competing models

Model	intercept	NH White Age(18-44) (indicator)	NH White Age(44+) (indicator)	Other race Age(18-44) (indicator)	testing rate	positivity rate	population density (categorical)	region Northeast (indicator)	region Midwest (indicator)	region South (indicator)	Democratic party (indicator)
M1	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
M2	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
M3	✓	✓	✓	✓	✓		✓	✓	✓	✓	
M4	✓	✓	✓	✓	✓		✓			✓	
M5	✓	✓	✓	✓			✓			✓	
M6	✓	✓	✓	✓	✓		✓				
M7	✓	✓	✓	✓			✓				

Whites (45+), Other Race (18-44), population density, each of these three models includes an additional auxiliary variable significant in at least one wave. For example, M4 includes an additional auxiliary variable testing rate because all M4 coefficients are significant in wave 10.

To select one out of the seven models listed in Table 5, we apply a cross-validation leave-one-state-out method. We now describe the method. We leave out the entire UAS survey data on the outcome variable y_i (e.g., perceived mask effectiveness) for state i and predict the vector of outcome variables for all sampled units of the leave out state using x_g for the sampled unit and z_{-i} for the leave out state. Let $f(y_i|y_{-i})$ denote the joint density of y_i, all the observations in state i, conditional on the data from the rest of the states, say y_{-i}. For the Bernoulli distribution of y_i for state i, using independence, we have for known model parameters β and γ:

$$\begin{aligned}
\log f(y_i|y_{-i};\beta,\gamma) &= \sum_{g=1}^{G}\sum_{k=1}^{n_{gi}}\left[y_{gik}\log\theta_{gi}+(n_{gi}-y_{gik})\log(1-\theta_{gi})\right] \\
&= \sum_{g=1}^{G}\sum_{k=1}^{n_{gi}}\left[y_{gik}\log\left(\frac{\theta_{gi}}{1-\theta_{gi}}\right)+n_{gi}\log(1-\theta_{gi})\right] \\
&= \sum_{g=1}^{G}\sum_{k=1}^{n_{gi}}\left[y_{gik}(x_g'\beta+z_i'\gamma)-n_{gi}\log\left(1+\exp(x_g'\beta+z_i'\gamma)\right)\right].
\end{aligned}$$

Using data from the rest of states, i.e., $y_{(-i)}$ we get survey-weighted estimates β and γ and plug in the above expression. Let these estimates be $\hat{\beta}_{w,(-i)}$ and $\hat{\gamma}_{w;(-i)}$. We then define our model selection criterion as:

$$C=\sum_{i=1}^{m}\sum_{g=1}^{G}\sum_{k=1}^{n_{gi}}w_{gik}\left[y_{gik}(x_g'\hat{\beta}_{w,(-i)}+z_i'\hat{\gamma}_{w;(-i)})-n_{gi}\log\left(1+\exp(x_g'\hat{\beta}_{w,(-i)}+z_i'\hat{\gamma}_{w;(-i)})\right)\right].$$

For each of the models in Table 5, we compute C model selection measures for all the waves (wave 1-16). In Table 6, we report the quantiles (minimum, first quartile, median, third quartile, maximum) and mean of C values (over the 16 waves) for each model in Table 5. We divide C value from each model by the sample size of the wave to scale down the numbers for ease of comparison. The C values are all negative, as these are logarithm of fractions. For every state, iteratively regressions are run and regression estimates are obtained, which are used in the formula. Using Table 6, we select M2 as the best performing model because this model produces maximum value of all descriptive statistics reported in Table 6.

5.2. Synthetic estimation of the perception of mask effectiveness for the states

In this section, we consider benchmarked synthetic estimates, which are obtained from the synthetic estimates after a ratio adjustment. These benchmarked synthetic estimates, when appropriately aggregated over the 50 states and the District of Columbia, yield the national direct estimate. We compare both synthetic and benchmarked synthetic estimates

Table 6: Cross validation leave one state out statistic for all models

Model	0%	25%	50%	75%	100%	Mean
M1	-89	-83	-75	-68	-16	-71
M2	-79	-55	-10	-8	-7	-28
M3	-92	-86	-83	-78	-20	-78
M4	-92	-86	-82	-77	-21	-78
M5	-91	-84	-80	-76	-19	-76
M6	-68	-59	-55	-51	-17	-54
M7	-90	-82	-76	-69	-15	-72

with the corresponding direct sample survey estimates (i.e., weighted proportion of people from UAS who believe mask is extremely effective) for the 50 states and the District of Columbia. This gives us an idea about the magnitude of biases in the synthetic and benchmarked synthetic estimates because direct estimates, though unreliable, are unbiased or approximately so. In Figure 4, we have 6 plots corresponding to 6 states (3 with small population - District of Columbia, North Dakota, Rhode Island, and 3 with large population - California, New York, Florida) of point estimates (direct and benchmarked synthetic) vs waves, which display time series trends from wave 1 to wave 16. The direct estimates of perceived mask effectiveness for small states could be unreasonable. For example, for the District of Columbia, direct estimates are 0% for both waves 1 and 2. On the other hand, benchmarked synthetic estimates 18% and 39% are more reasonable for these two waves – they are more in line with the national estimates for such waves. Similarly, for Rhode Island unreasonable 100% perceived mask effectiveness direct estimates for waves 13 and 15 have also been modified to more reasonable benchmarked synthetic estimates.

Figure 5 displays standard errors of direct and benchmarked synthetic estimates. The proposed jackknife method is used to compute standard errors for benchmarked synthetic estimates. We denote standard errors of direct and benchmarked synthetic estimates by SE and STD, respectively. If we focus on the error graphs, the values from direct estimates get as high as 32% for small states (i.e. one with low contribution to overall sample size). Using benchmarked synthetic estimates at the state level, the error has reduced to almost 6 times with as low as 6% standard error from the jackknife method. For larger states like New York and Florida, the errors reduce using benchmarked synthetic estimate, although not to a great extent. For the state contributing most to the sample size, California, the standard errors are more or less similar.

For the chosen model M2, we create a state level comparative diagram of benchmarked synthetic estimates with direct estimates in Figure 6 using data from wave 16. As at the state level, the synthetic estimates and the corresponding benchmarked synthetic estimates are really close, we have not plotted synthetic estimates for ease of viewing. We observe that our synthetic estimates are much more stable than the corresponding direct estimates. The states arranged in increasing order of population sizes show that the issue of highly variable state level direct estimates for the smaller states has been mitigated by the synthetic method.

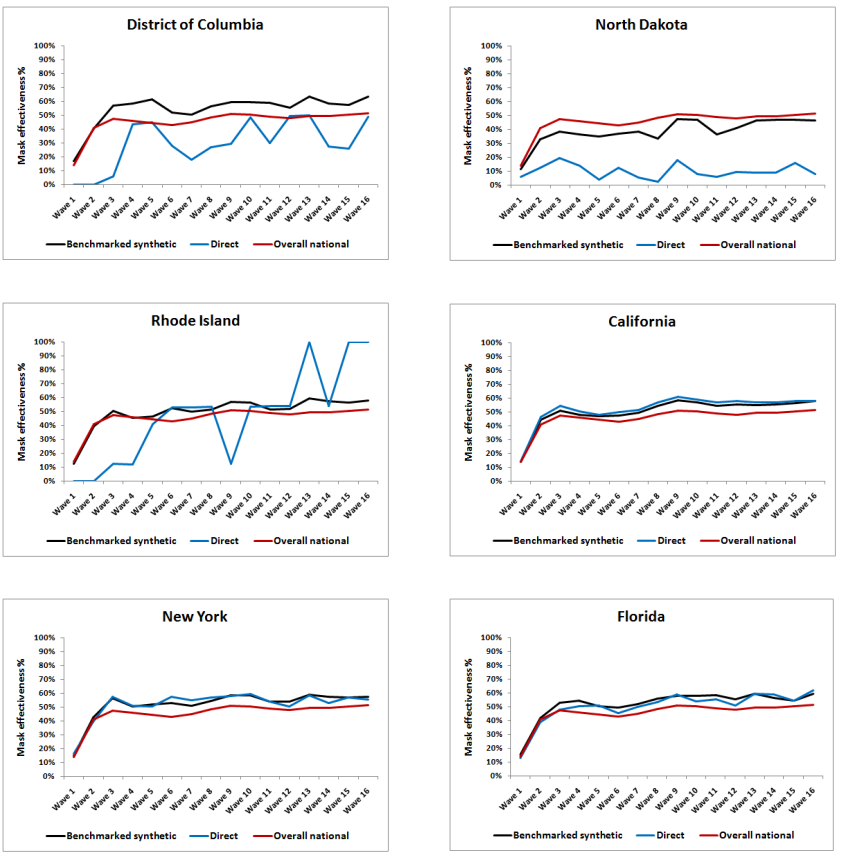


Figure 4: Time series trend of direct and benchmarked synthetic estimate for 6 sample states (3 small, 3 large)

Figure 5: Time series trend of SE of direct and benchmarked synthetic estimate for 6 sample states (3 small, 3 large)

For largely populated states as well as for small ones, the benchmarked synthetic estimates are doing a good job of estimating the proportion of the response variable. We next check the robustness of the synthetic estimator in terms of variance through the jackknife method.

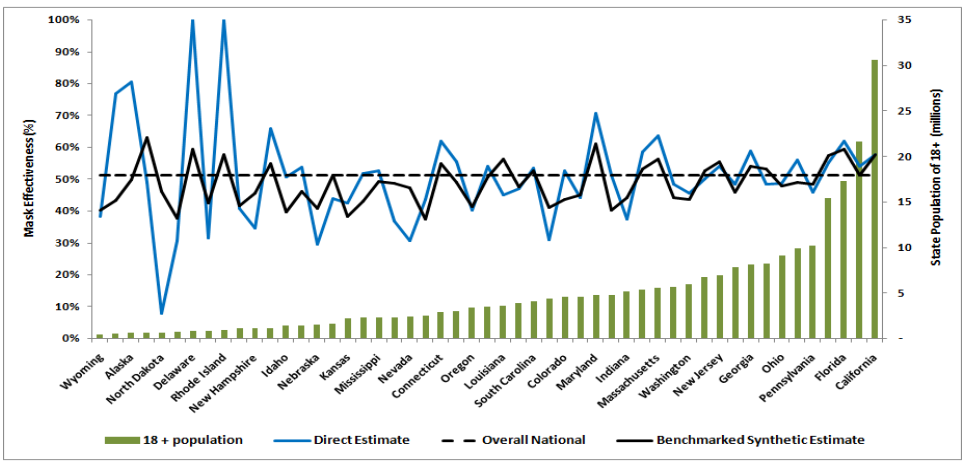


Figure 6: State level comparison of direct and synthetic estimates from M1 for wave 16

We fitted M2 for wave 16 data and obtained jackknife estimates of variances and hence standard errors at the state level. We provide a comparative view with the SE from direct estimates at the state level. The two graphs in Figure 7 are based on the wave 16 data. In the x-axis, states are arranged in increasing order of sample sizes. In the first graph, the y-axis is the ratio of direct estimate (survey-weighted) and synthetic estimate. In the second graph, the y-axis is the ratio of STD and SE, where SE is the standard error of direct estimate coming right from UAS (treating states as domains) and STD is the jackknife standard error of benchmarked synthetic estimate. For states with small sample sizes (e.g. Rhode Island, Wyoming), we see a lot of differences between the survey-weighted direct estimates and the synthetic estimates. For states with large sample sizes (e.g., California), the ratio is approaching to 1 (as plotted by the straight line) as the auxiliary variables used to construct the synthetic estimator are reasonable. We observe that all the jackknife estimates are much smaller than direct estimates and we conclude that the model is a fair one at estimating the perceived mask effectiveness at state level.

We define Benchmark Ratio (BR) as the ratio of the overall direct national estimate to the synthetic estimate (aggregated at the national level). The synthetic estimates, which are obtained at the state level, are aggregated by multiplying by the ratio of the adult state population to the overall US adult population estimate and then adding up. The closer the value of BR is to 1 the better is the model. We see from Table 7 that BR is close to 1 for all waves, using which we compute the Benchmarked or BR synthetic estimate.

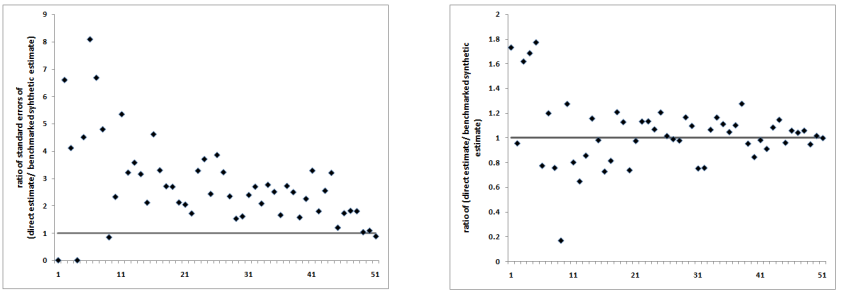


Figure 7: Comparison of direct with benchmarked synthetic estimates through the ratio SE/STD and ratio of estimates from M1 for wave 16; domains arranged in increasing order of sample size.

Table 7: Benchmarking ratios and national synthetic and benchmarked synthetic estimates for last five waves; synthetic estimates are based on Model 1.

Model	0%	25%	50%	75%	100%	Mean
Benchmarking Ratio	0.98	0.98	0.99	0.99	1.00	0.99
Synthetic	14.08%	45.22%	48.63%	50.43%	51.94%	46.00%
Benchmarked Synthetic	13.86%	44.62%	48.02%	49.66%	51.22%	45.38%

6. Conclusion

The method of estimating population means or totals for the states of USA explained in the paper provides sensible and numerically sound estimates and the model selection with all standard error of estimates within 2%. We noticed high variability of synthetic estimates at the state level estimation. We further note that while direct UAS estimates are designed to produce approximately unbiased estimates at the national level, they are subject to biases for the state level estimation. Biases in the direct proportion estimates at the state level may arise from the fact that they are essentially ratio estimates since the state sample sizes are random and expected sample sizes are small for most states. Moreover, the UAS weights are not calibrated at the state level.

From our investigation, we found that synthetic estimates improve on UAS direct estimates in terms of variance reduction, especially for the small states. But since synthetic estimates are derived using a working model, they are subject to biases when working model is not reasonable. However, we observe that the benchmarking ratios for all waves are consistently around 1 showing lack of evidence for bias. Our benchmarked synthetic estimates are close to the synthetic estimates because the benchmarking ratios are close to 1. None-theless by benchmarking synthetic estimates we achieve data consistency and it is reasonable to expect to reduce biases as well. We add that it is possible to reduce biases at the state level by benchmarking the synthetic estimates to the UAS direct estimates for a group of states

(e.g., benchmarking with a division). This may be needed for other synthetic estimation problems.

Infection rates has declined declining in most parts of the USA. However, with differential vaccination hesitancy rates across the US states and emergence of new COVID-19 variants, identification of granular level mask effectiveness perception rates may remain an important problem in the US. While we wait to reach herd immunity through aggressive vaccination program, good control of the spread of COVID-19 and its different variants in different parts of the world is essential. Thus, it will be of interest to understand mask effectiveness perception rates in communities throughout the world, especially where infection rates are high. Not only COVID-19, but for other infectious diseases, mask effectiveness perception is likely to stay relevant. While we illustrate the proposed synthetic methodology for state level estimation of perceived mask effectiveness, the methodology is general and can be applied to granular sub-state levels with no sample from the primary survey data. Moreover, similar synthetic methodology can be developed in the future to estimate granular level proportions related to personal finance, mental health, etc.

7. Acknowledgements

The project described in this paper relies on data from survey(s) administered by the Understanding America Study, which is maintained by the Center for Economic and Social Research (CESR) at the University of Southern California. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of USC or UAS. The collection of the UAS COVID-19 tracking data is supported in part by the Bill & Melinda Gates Foundation and by grant U01AG054580 from the National Institute on Aging. The second author's research was partially supported by the U.S. National Science Foundation Grant SES-1758808.

References

Angrisani, M., A. Kapteyn, E. Meijer, and H.-W. Saw, (2019). Sampling and weighting the Understanding America Study, Working Paper, No. 2019-004, Los Angeles, CA: University of Southern California, Center for Economic and Social Research.

Census Bureau, (2010). Census Regions and Divisions of the United States.

Ghosh, M., (2020). Small area estimation: Its evolution in five decades. Statistics in Transition New Series, Special Issue on Statistical Data Integration, pp. 1–22.

Hansen, M., W. Hurwitz, and W. Madow, (1953). Sample Survey Methods and Theory, Vol. 1. Wiley-Interscience.

Lahiri, P. and S. Pramanik, (2019). Estimation of average design-based mean squared error of synthetic small area estimators, Austrian Journal of Statistics, 48, pp. 43–57.

Lumley, T., (2004). Analysis of complex survey samples, Journal of Statistical Software, 9(1), pp. 1–19, R package verson 2.2.

Lumley, T., (2010). Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R. John Wiley and Sons.

Lumley, T., (2020). Survey: analysis of complex survey samples, R package version 4.0.

Marker, D., (1995). Small Area Estimation: A Bayesian Perspective. Phd thesis, University of Michigan, Ann Arbor, MI.

Marker, D., (1999). Organization of small area estimators using a generalized linear regression framework, Journal of Offcial Statistics, 15, pp. 1–24.

Rao, J. N. K., I. Molina, (2015). Small Area Estimation, 2nd Edition, Wiley.

Stasny, E., P. Goel, and D. Rumsey, (1991). County estimates of wheat production, Survey Methodology, 17 (2), pp. 211–225.

U.S. Census Bureau, (2020). Census Bureau Population Estimates. Wikipedia, (2020). Political party strength in U.S. states.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23 No. 1, pp. 21–38, DOI 10.21307/stattrans-2022-002
Received – 03.05.2021; accepted – 19.07.2021

Comparison of tree-based methods used in survival data

Aysegul Yabaci[1], **Deniz Sigirli**[2]

ABSTRACT

Survival trees and forests are popular non-parametric alternatives to parametric and semi-parametric survival models. Conditional inference trees (Ctree) form a non-parametric class of regression trees embedding tree-structured regression models into a well-defined theory of conditional inference procedures. The Ctree is applicable in a varietyof regression-related issues, involving nominal, ordinal, numeric, censored, as well as multivariate response variables and arbitrary measurement scales of covariates. Conditional inference forests (Cforest) consitute a survival forest method which combines a large number of Ctrees. The Cforest provides a unified and flexible framework for ensemble learning in the presence of censoring. The random survival forests (RSF) methodology extends the random forests method enabling the approximation of rich classes of functions while maintaining generalisation errors low. In the present study, the Ctree, Cforest and RSF methods are discussed in detail and the performances of the survival forest methods, namely the Cforest and RSF have been compared with a simulation study. The results of the simulation demonstrate that the RSF method with a log-rank score distinction criteria outperforms the Cforest and the RSF with log-rank distinction criteria.

Key words: tree-based methods, conditional inference trees, conditional inference forests, random survival forests.

1. Introduction

Tree-based methods constitute classification and regression models in the form of a tree structure according to data sets. Understanding the decision rules used in the creation of tree structures makes the use of the method common. Decision trees perform decision making with a multi-stage and sequential approach in solving the classification and regression problem (Safavian et al. 1991).

The Classification and Regression Trees (C&RT) provide a visual representation of the effect of independent variables on dependent variables and the interaction between

[1] Department of Biostatistics and Medical Informatics, Bezmialem Vakif University, Faculty of Medicine, Istanbul, Turkey. E-mail: aysegulyabaci@gmail.com. ORCID: https://orcid.org/0000-0002-5813-3397.

[2] Department of Biostatistics, Faculty of Medicine, Uludag University, Bursa, Turkey.
E-mail: denizsigirli@hotmail.com. ORCID: https://orcid.org/ 0000-0002-4006-3263.

them, which is used to estimate the class membership of a discrete or continuous dependent variable without pre-requisite presentation of the independent variable. In general, if the dependent variable is categorical, the name of the method is the classification tree, and if it is continuous, the method is called the regression tree (Breiman et al. 1984).

Survival trees and forests are popular non-parametric alternatives to parametric and semi-parametric survival models. A single tree set can be classified according to survival characteristics by taking into consideration independent variables, while a very powerful estimating tool can be obtained through tree sets created by the combination of trees.

The aim of this study is to evaluate the performances of random survival forests (RSF) and conditional inference forest (Cforest) methods as tree-based methods used in survival data analysis, for different conditional censored survival function estimators, for different sample sizes and for cases where the proportional hazard assumption is provided and not provided.

2. Methods of comparison

2.1. Conditional inference trees – Ctree

Let $\tilde{T}$ show the actual time of death and C be the time of censoring, $T = min(\tilde{T}, C)$ is the dependent variable and $\Delta = I(\tilde{T} \leq C)$ is the state variable. Let $\boldsymbol{X} = (X_1, \dots, X_p)$ be the vector of p dimensional covariate from $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_p$ sample space. The situation in which covariates are measured on any scale is discussed. Given the covariate of $\boldsymbol{X}$, T is the conditional distribution of the dependent variable, presented in the form of $\mathcal{F}_{\boldsymbol{T}|\boldsymbol{X}}$ to be a function of the common variables of $\mathcal{F}_{\boldsymbol{T}|\boldsymbol{X}}$ in the Eq. (1) that is bound to suppose.

$$\mathcal{F}_{T|\boldsymbol{X}} = \mathcal{F}(T \mid f(X_1, \dots X_p)). \tag{1}$$

Let $\mathcal{L}$ be given as in Eq. (2), where some X_{ji} (j=1,..,p ; i=1,..,n) covariate values are missing, and independent and identically distributed observation values are n units of the random sample 'learning sample'.

$$\mathcal{L} = \{(\boldsymbol{T}_i, \Delta_i, \boldsymbol{X}_i) \quad i = 1, \dots, n\}. \tag{2}$$

For each node in the tree, there is a unit weights vector. Let the unit weight vector be shown as $\boldsymbol{w} = [w_1, \dots w_n]$. If the observation values of the relevant variable are located on this node, the corresponding value in the weight vector is 1, and if not 0 (Hothorn et al. 2006b).

The following steps are taken to create conditional inference trees:

Step 1: For $\boldsymbol{w}$ unit weights, the general null hypothesis that there is independence between any of the p covariates and the dependent variable is tested. If this hypothesis is not rejected, then it is stopped. In other cases, $\boldsymbol{X}_{j^*}$ is chosen as the j^*nth covariate, which has the strongest relationship with T.

Step 2: The X_{j^*} variable, which divides the $A^* \subset X_{j^*}$ set into two discrete sets A^* and X_{j^*}/A^*, is selected.

Step 3: Step-1 and step-2, $\boldsymbol{w}_{left}$ and $\boldsymbol{w}_{right}$ unit weights are modified and repeated.

In Step 1, the absence hypothesis is as follows:

$$H_0 = \bigcap_{j=1}^{p} H_0^j$$

Here, p partial hypotheses are defined as follows:

$$H_0^j : \mathcal{F}_{T|\boldsymbol{X}_j} = \mathcal{F}_T \;\; ; \; \mathrm{j} = 1, \dots, \mathrm{p}$$

When the H_0 hypothesis cannot be rejected at the specified α level of significance, the division stops. The relationship between the T and each X_j , $j = 1, \dots, p$ covariate is tested by the H_0^j hypotheses, which are partial hypotheses. For this hypothesis, the test statistics or p values are used to select the covariate that is the most associated with T. Weights of w_i can be set to 0 or 1.

The symmetric group of all permutations of elements corresponding to the unit weight $w_i = 1$ is shown with $S(\mathcal{L}, w)$. In this case, the relationship between T and X_j , $(j = 1, \dots, p)$ is measured by the linear test statistic given below (Hothorn et al. 2006b).

$$\boldsymbol{T}_j(\mathcal{L}, w) = vec\left(\sum_{i=1}^{n} w_i g_j(X_{ji}) h((T_i, (T_1, \dots . T_n))')\right) \in \mathbb{R}^{p_j q} \tag{3}$$

Where $g_j : \mathcal{X}_j \to \mathbb{R}^{p_j}$ is the non-random transformation of the covariate X_j. For continuous covariate, $g_{ji}(x) = x$ unit transformation can also be applied. Also, it is possible to rank or nonlinear transformations. The effect function $h : \mathcal{T} \times \mathcal{T}^n \to \mathbb{R}^q$ is based on response variables in symmetric permutation and is obtained as in Eq. (4). In survival data, h can be selected as log-rank score.

$$h(T_i, (T_1, \dots . T_n)) = \sum_{k=1}^{n} w_k \, I(T_k \le T_i) \quad i = 1, .. n \tag{4}$$

To divide the covariate selected in Step-1 into two, the permutation test is used in Step-2. The test statistic, which is a special case of $\boldsymbol{T}_j(\mathcal{L}, w)$ the test statistic, is calculated as in Eq.(5).

$$\boldsymbol{T}_{j^*}^{A}(\mathcal{L}, w) = vec\left(\sum_{i=1}^{n} w_i \, I\left(X_{j^* i} \in A\right) h(T_i, (T_1, \dots . T_n)')\right) \in \mathbb{R}^q \tag{5}$$

This linear statistic gives two sampling test statistics that measure the discordance between samples $\{T_i|w_i > 0 \; ve \; X_{ji} \in A; i = 1,..,n\}$ and $\{T_i|w_i > 0 \; ve \; X_{ji} \notin A; i = 1,..n\}$. Conditional expected value $\mu_{j^*}^A$ and covariance $\Sigma_{j^*}^A$ are calculated as in Eq. (6) and Eq. (7) respectively.

$$\mu_j = \mathbb{E}\left(\boldsymbol{T}_j(\mathcal{L},w)\middle|S(\mathcal{L},w)\right) = vec((\textstyle\sum_{i=1}^n w_i g_j(X_{ji}))\; \mathbb{E}\,(h \mid S(\mathcal{L},w))') \tag{6}$$

$$\textstyle\sum_j = \mathbb{V}(\boldsymbol{T}_j(\mathcal{L},w) \mid S(\mathcal{L},w)\,). \tag{7}$$

Using this expected value and covariance, $\boldsymbol{T}_{j^*}^A(\mathcal{L},w)$'s standardized test statistic is obtained from $c(t_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A)$. The distinction that corresponds to the maximum of this test statistic is indicated by "A^* ". The test statistic that is maximized over all possible subsets of A is as in Eq. (8)

$$A^* = argmax_A \; c(t_{j^*}^A, \mu_{j^*}^A, \Sigma_{j^*}^A). \tag{8}$$

Then, as stated in Step 2 of the algorithm, "w_{left}" and "w_{right}" unit weights are determined by the functions $\boldsymbol{w}_{left,i} = w_i \; \mathrm{I}(\boldsymbol{X}_{j^*i} \in A^*)$ and $\boldsymbol{w}_{right,i} = w_i \; \mathrm{I}(\boldsymbol{X}_{j^*i} \notin A^*)$ and the weights are modified and repeat Step-1 and Step-2.

2.2. Conditional inference forest method – Cforest

Assume that the conditional distribution function of T the response variable is dependent on random variable X with the function $f: \mathcal{X} \to \mathbb{R}$. In this case, $\mathcal{F}_{T|X} = \mathcal{F}_{T|f(X)}$. The conditional censoring survival function is given in the form of $G(T \mid \mathbf{X}) \approx \mathbb{P}(C > t \mid \mathbf{X} = \mathrm{x})$. Let ψ be the function space of all candidate estimators $\psi: \mathcal{X} \to \mathbb{R}$. Estimation of the regression function *f*, as defined by full data loss function *L* is found by minimizing the expected value of the risk function. However, the full data function cannot be calculated because all data cannot be reached in the presence of censored observation. Therefore, instead of the full data loss function, the observed data loss function $L = (T, \psi(\mathbf{X}) \mid \eta)$ is used. In this case, the expected value of the observed data loss function is obtained as given in Eq. (9). Here, the expected value of the full loss data function is intended to be minimized according to the candidate estimators $\psi \in \Psi$ (Hothorn et al. 2006a).

$$\mathbb{E}_{T,\boldsymbol{X}} L_{full}\big(T, \psi(\mathbf{X})\big) = \int L(T, \psi(\mathbf{X}) \mid \eta)\, d\mathcal{F}_{T,\Delta,\mathbf{X}} = \mathbb{E}_{T,\Delta,\mathbf{X}}\, L(T, \psi(\mathbf{X}) \mid \eta). \tag{9}$$

In Eq. (9), η is the nuisance parameter and can be defined as a conditional censored survival function. The observed loss data function can be defined as Eq. (10) by using $G(T \mid \boldsymbol{X})^{-1}$.

$$L(T, \psi(\boldsymbol{X}) \mid G) = L\big(T, \psi(\boldsymbol{X})\big) \frac{\Delta}{G(T|\boldsymbol{X})}. \tag{10}$$

The full data loss function is weighted by the inverse of the probability of censored after T time. In this case, the expected value of the observed data loss function is obtained as in Eq. (11).

$$\widehat{\mathbb{E}}_{T,\Delta,\boldsymbol{X}} L(T,\psi(\boldsymbol{X}) \mid G) = n^{-1}\sum_{i=1}^{n} L\big(T_i,\psi(\boldsymbol{X_i}) \mid \hat{G}\big) = n^{-1}\sum_{i=1}^{n} L\big(T_i,\psi(\boldsymbol{X_i}) \mid \hat{G}\big)\frac{\Delta_i}{\hat{G}(T_i \mid \boldsymbol{X_i})}. \tag{11}$$

The regression function predictor $\hat{f}$ is obtained by minimizing this equation according to the candidate predictors $\psi \in \Psi$. Here the G conditional censored survival function is unknown and its estimator is used instead. As a $\hat{G}$ estimator, the nonparametric estimator, Cox estimator or the cumulative Aalen estimator can be used. In the case of $w_i = \Delta_i\, \hat{G}(T_i \mid \boldsymbol{X_i})^{-1}$, $\mathbf{w} = (w_1, w_2, .., w_n)$ is called IPC (the inverse probability of censored weights).

The conditional inference forest (cforest) algorithm has been proposed by *Hothorn et al* to find the values of ψ that minimize the expected value of the observed data loss function. $\mathbf{w}$ weight vector is calculated by using the observed learning sample $\mathcal{L} = \{(T_i, \Delta_i, \boldsymbol{X_i}); i = 1, \dots, n\}$ and $w_i = \Delta_i\, \hat{G}(T_i \mid \boldsymbol{X_i})^{-1}$. If the learning sample contains a censored observation value, it is $w_i = 0$ because it is $\Delta_i = 0$. The steps of the algorithm are as follows (Hothorn et al. 2006a):

Step 1: Set $m = 1$ and $M > 1$.

Step 2: From the multinomial distribution with parameter n and $(\sum_{i=1}^{n} w_i)^{-1}\boldsymbol{w}$, a random vector of the unit numbers $\mathbf{v}_m = (v_{m1}, \dots, v_{mn})$ is drawn.

Step 3: With a regression tree, the sample space $\mathcal{X}$ is divided into $K(m)$ cells and created $\pi_m = \big(R_{m1}, \dots, R_{mK(m)}\big)$ pieces are created. This regression tree is created using the learning sample $\mathcal{L}$ with case counts $\mathbf{v}_m$. In the permutations of the $\mathcal{L}$ learning sample, *i th* observation takes place once.

Step 4: Increase m by one, repeat Step 2 and step 3 until $m = M$.

In Step 3, using the learning sample obtained in Step 2, a survival tree is obtained with a conditional inference trees algorithm. Let $\mathcal{T}_m$ denote m th survival tree and $\mathcal{T}_m(\boldsymbol{x})$ denote terminal node with $\boldsymbol{x}$ covariate value in the m th tree. Each $\boldsymbol{x}$ value will take place on a single terminal node.

$$\widetilde{N}_i(s) = I(T_i \le s, \Delta_i = 1) \text{ and } \tilde{Z}_i(s) = I(T_i > s)$$

$$\widetilde{N}_m^*(s,x) = \sum_{i=1}^{n} v_{im} I(X_i \in \mathcal{T}_m(x))\widetilde{N}_i(s) \tag{12}$$

$$\tilde{Z}_m^*(s,x) = \sum_{i=1}^{n} v_{im} I\big(X_i \in \mathcal{T}_m(x)\big)\tilde{Z}_i(s) \tag{13}$$

Where $\tilde{N}_m^*(s,x)$ and $\tilde{Z}_m^*(s,x)$ are respectively the number of uncensored events in the terminal node up to the time of s corresponding to the $\boldsymbol{x}$ covariate value, and the number of units at risk at s time. In this case, when $\boldsymbol{x}$ is given a covariate, the ensemble survival function for t time is equal to that of Eq. (14).

$$\hat{S}^{cforest}(t \mid x) = \prod_{s \le t} \left(1 - \frac{\sum_{m=1}^{M} \tilde{N}_m^*(s,x)}{\sum_{m=1}^{M} \tilde{Z}_m^*(s,x)}\right). \quad (14)$$

2.2. Random survival forest method – RSF

The algorithm steps of the RSF method are as follows:

Step 1: Extract M bootstrap sample from the original data. Each bootstrap sample should exclude average 37% of the original data. The data that is excluded is called out-of-bag data (OOB).

Step 2: Create a survival tree for each bootstrap samples. On each node of the tree, randomly $\sqrt{p}$ candidate variable is selected. The node is separated by using candidate variables that maximise the survival difference between child nodes.

Step 3: continue the split until at least one observed case remains on each terminal node.

Step 4: Cumulative hazard function (CHF) is calculated for each tree. Average to obtain the ensemble CHF.

Step 5: Using OOB data, estimation error is calculated for the ensemble cumulative hazard function (Ishwaran et al. 2008a).

Logrank test is being used to compare two groups survival, by putting equal weights to each individual (Mantel N. 1966; Karadeniz et al. 2018). Two methods can be used as separation criteria in the algorithm. The first is the log-rank distinction and the second is the log-rank score distinction (Segal 1988; Ciampi et al.1986; Hothorn and Lausanne 2003).

i. Log-rank distinction criteria

Let T_i ; $i = 1,..,n$ denote the survival time of i th unit and X_j covariate for the distinction on a node, $X_j \le c$ and $X_j > c$ according to the cut point of c. Let $s_1 < s_2 < \cdots < s_z$ denote discrete time of death on a node for $z = 1, \ldots, N$. For the m *th* tree, $\tilde{N}_{md}^*(s_z, x)$ show the number of people dying in s_z time on child nodes d=1,2.

$\tilde{N}_m^*(s_z, x) = \tilde{N}_{m1}^*(s_z, x) + \tilde{N}_{m2}^*(s_z, x)$ is in format. For the m *th* tree, , $\tilde{Z}_{md}^*(s_z, x)$ indicates the number of units at risk at s_z time on child nodes d=1,2. In this case, $\tilde{Z}_m^*(s_z, x) = \tilde{Z}_{m1}^*(s_z, x) + \tilde{Z}_{m2}^*(s_z, x)$ and $\tilde{Z}_{m1}^*(s_z, x) = \#\{T_i \ge s_z\,, x_i \le c\}$, $\tilde{Z}_{m2}^*(s_z, x) = \#\{T_i \ge s_z\,, x_i > c\}$. Where x_i, is the value that the X_j covariate takes for unit i *th.* n_d is the total number of units observed in the d *th* child node. Thus, $n_1 = \#\{i: x_i \le c\}$ and $n_2 = \#\{i: x_i > c\}$ are equal to $n = n_1 + n_2$.

The log-rank test statistic for the c cut-off value of the X_j covariate is as in Eq. (15).

$$LogRank(X_j, c) = \frac{\sum_{z=1}^{N}\left(\tilde{N}^*_{m1}(s_z,X_j) - \tilde{Z}^*_{m1}(s_z,X_j)\frac{\tilde{N}^*_m(s_z,X_j)}{\tilde{Z}^*_m(s_z,X_j)}\right)}{\sqrt{\sum_{z=1}^{N}\frac{\tilde{Z}^*_{m1}(s_z,X_j)}{\tilde{Z}^*_m(s_z,X_j)}\left((1-\frac{\tilde{Z}^*_{m1}(s_z,X_j)}{\tilde{Z}^*_m(s_z,X_j)})(\frac{\tilde{Z}^*_m(s_z,X_j)-\tilde{N}^*_m(s_z,X_j)}{\tilde{Z}^*_m(s_z,X_j)-1})\tilde{N}^*_m(s_z,X_j)\right)}}. \tag{15}$$

$\left|LogRank(X_j, c)\right|$ provides a measure for node distinction. The distinction occurs between the two terminal nodes that has the highest $\left|LogRank(X_j, c)\right|$ value. The best distinction value of $\left|LogRank(X_j^*, c^*)\right| \geq \left|LogRank(X_j, c)\right|$ is determined by the value of the X_j covariate and c cut-off value (Segal 1988; Hothorn and Lausen 2003).

ii. Log-rank score distinction criteria

Another distinction rule is the log-rank score distinction rule proposed by Hothorn and Lusen (2003). Assume that the values of the X_j covariate are sorted as $x_1 \leq x_2 \leq \cdots \leq x_n$. For each T_i survival time, ranks are obtained as in Eq. (16).

$$\alpha_i = \Delta_i - \sum_{k=1}^{\Gamma_i}\frac{\Delta_k}{n-\Gamma_k+1}. \tag{16}$$

Where, $\Gamma_k = \#\{s: T_s \leq T_k\}$. In this case, the log-rank score statistic is obtained as in Eq. (17).

$$LogRankskor(X_j, c) = \frac{\sum_{x_i \leq c}\alpha_i - n_1\overline{\alpha}}{\sqrt{n_1(1-\frac{n_1}{n})s_\alpha^2}}. \tag{17}$$

In Eq. (17), $\overline{\alpha}$ and s_α^2 is defined as the sample mean and sample variance of ranks, respectively. $LogRankscore(X_j, c)$ provides log-rank score for node distriction.

2.4. Estimators used in estimating G conditional censored survival function

***i.* Nonparametric estimator**

Let $G(T \mid X) \approx P(C > t \mid X = x)$ and $K(t)$ denote respectively conditional survival function of the censoring time and any kernel function. The nonparametric estimator used by Graf et al. is given in Eq. (18). (Gerds and Schumacher 2007).

$$\hat{G}_{NonPar} = \left\{G: sup_t \frac{|G(T|X) - G(T|X')|}{|X - X'|^\alpha} \leq K(t) > 0\right\}. \tag{18}$$

***ii.* Cox estimator**

Let α and $H_0(t)$ denote respectively regression coefficient and initial cumulative hazard function. Cox regression estimator is given in Eq. (19) (Gerds and Schumacher 2007).

$$\hat{G}_{Cox} = \left\{G_{\alpha,H_0(t)}: G(T \mid X) = exp\{-\exp(\alpha' X)H_0(t)\}; \alpha \in \mathbb{R}^d\right\}. \tag{19}$$

***iii*. Aalen estimator**

Let $\alpha(t)$ denote time-dependent regression coefficient. Cumulative Aalen regression estimator are given as in Eq. (20) (Gerds and Schumacher 2007).

$$\hat{G}_{Aalen} = \left\{ G_\alpha : G(T \mid X) = exp\left\{ - \int_{s=0}^{t} X' \alpha(s).ds \right\} \right\}. \tag{20}$$

2.5. Criteria used to evaluate model performance

2.5.1. Brier Score – BS

The prediction error defined as the time dependent expected Brier score is one of the measures for assessing the predictive performances of rival survival modeling strategies. If the score is close to zero, the class estimates are accepted to be reliable. Let $\Delta_i = I\left(\tilde{T}_i \leq t\right)$ be state of *i th* unit for *t* time. When *X* is given, the probability of survival predicted at *t* time for the *i th* unit is shown as $\hat{S}(t \mid X_i)$. In this case, the Brier score is the same as the Eq. (21)

$$BS\left(t, \hat{S}\right) = E\left[I(\tilde{T}_i > t) - \hat{S}(t \mid X_i)^2 \right]. \tag{21}$$

The expected value is calculated based on the data of the *i th* unit which is not included in the learning set. The first critical value for the Brier score is 33%. This corresponds to the risk predicted by the random number drawn from the U[0,1] distribution. The second critical value is 25% and corresponds to 50% risk estimation for each unit. Another criterion is the Brier score value obtained from the model from which all independent variables are extracted (Ishwaran et al. 2008a). Residual squares are weighted using the inverse probabilities of the censored weights given in Eq. (22).

$$\widehat{W}_i(t) = \frac{I(\tilde{T}_i \leq t)\Delta_i}{\hat{G}(\tilde{T}_i - | X_i)} + \frac{I(\tilde{T}_i > t)}{\hat{G}(t | X_i)}. \tag{22}$$

Here $\hat{G}(t \mid x) \approx P(C_i > t \mid X_i = x)$ is the estimate of the conditional survival function for the *i th* unit of censoring time. If an independent set of data D_n is available, the expected Brier score is the same as in Eq. (23).

$$\widehat{BS}\left(t, \hat{S}\right) = \frac{1}{n} \sum_{i \in D_n} \widehat{W}_i(t) \{ I(\tilde{T}_i > t) - \hat{S}(t \mid X_i) \}^2. \tag{23}$$

Where *n* is the number of units in D_n (*i*=1,.,*n*) and calculated from the learning data $\hat{S}$.

2.5.2. Integrated Brier Score – IBS

Prediction errors can be summed up with IBS as follows:

$$IBS(TH, \tau) = \frac{1}{\tau} \int_{t=0}^{\tau} TH(t, \hat{S}) dt \tag{24}$$

Where *TH* is the prediction error obtained using methods such as Apperr (apparent prediction), BootCvErr (Boostrap Cross Validation prediction), NoInfErr (ignorance prediction error), boot632pluserr (0.632+ prediction). τ is the time of maximum observation ($\tau > 0$).

2.5.3. Concordance Index – C-Index

Concordance Index is the probability of concordance between the predicted and the observed survival. Model performance increases as the C Index value approaches to 1. C-Index is not based on a fixed point of time, unlike other indexes that measure the performance of survival (Ishwaran, 2008). C-Index is calculated with Steps 1-3:

Step 1: Create all possible pairs of units on the data set.

Step 2: If pairs is censored which the unit corresponding to shorter survival time, the pair is neglected. If both pairs are alive and $T_i = T_j$, i and j pairs are neglected. "Allowed" can be expressed as the total number of pairs that are not neglected.

Step 3: When $T_i \neq T_j$, if it has worse prediction results with shorter survival time, it gets a value of 1, if the prediction results are equal it gets a value of 0.5 for each allowed pair. For each allowable pair, if $T_i = T_j$ and both are dead, the result is worse than that which is dead, then it gets a value of 1, otherwise it gets a value of 0.5. "Concordance" represents the sum of the values received by all allowed pairs.

C-Index is defined below:

$$C = \frac{Concordance}{Allowed}.$$

3. Material and method

Simulation studies were carried out under different scenarios in order to compare the performance of RSF and Cforest methods from tree-based methods used in survival data. In addition, Aalen, Cox and nonparametric estimators were evaluated for the performance of the RSF method in the case of using different separation criteria and conditional survival function of the censoring time (Gerds and Schumacher 2007). For this purpose, data derivation were made with two different scenario. The first scenario examine the situation in which the proportional hazard assumption is provided and the second senario examine the situation in which the proportional hazard assumption is not provided (Ishwaran and *et* al. 2010; Zhu and Kosorok 2012). In both scenarios, the criterion for the number of independent variables randomly chosen in each division was taken as the square root of the number of variables *p*. Sample size was determined as 100, 200 and 300. The number of trees created is *M*=100, bootstrap number is *B*=100, the test set (out of bag data) uses 37% of the total sample size and the training set (in bag data) uses 63% of the total sample size. The number of units on each terminal node is limited to 6. Simulation was carried out with 1000 repetitions.

Scenario 1: The number of independent variables was taken as p = 25. Let $X = (X_1, \dots, X_{25})$, $\Sigma_{ij} = \rho^{|i-j|}$ (ρ=0.9) and diagonal elements 1. Covariates were derived from the multivariate normal distribution with $\Sigma_{p\times p}$ variance-covariance matrix and $[0]_{p\times 1}$ mean vector. Let b_0=0,1, survival times were derived independently from exponential distribution with $\mu = b_0 \times \sum_{i=11}^{20} X_i$. Censored times were derived independently from exponential distribution with $\mu/2$. The state variable was obtained as $\Delta = I(\tilde{T} \leq C)$. For this scenario, censored rate was approximately 30%.

Scenario 2: The number of independent variables was taken as p = 25. Let $X = (X_1, \dots, X_{25})$, $\Sigma_{ij} = \rho^{|i-j|}$ (ρ=0.75) and and diagonal elements 1. Covariates were derived from the multivariate normal distribution with $\Sigma_{p\times p}$ variance-covariance matrix and $[0]_{p\times 1}$ mean vector. Survival times were derived independently from log-normal distribution with $\mu = 0.1 \times \left|\sum_{i=1}^{5} X_i\right| + 0.1 \times \left|\sum_{i=21}^{25} X_i\right|$. Censored time was derived from the log normal distribution with $\mu + 0.5$ mean. The state variable was obtained as $\Delta = I(\tilde{T} \leq C)$. For this scenario, censored rate was approximately 30%. Model performances were evaluated with IBS and C Index.

Pec, party, randomForestSRC packets were used in R 3.4.1 program in simulation study(Hothorn and et al. 2005; Mogensen and et al. 2012a; Ishwaran and et al. 2008b).

4. Evaluation

The results of the simulation study were presented by taking into consideration scenario 1 and Scenario 2 with Table 1-12. The mean and standard values of RSF and Cforest method with two separate criteria for Aalen, Cox and nonparametric estimators, three estimators used in the calculation of IPC weights and three sample sizes were presented in the table.

Table 1. The mean and standard error values according to the C-Index criteria of RSF and Cforest method for different survival times in cases where n=100 and the proportional hazard assumption is provided

C-Index (scenario 1, n=100)	Survival Time					
	0.5		2		3.5	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.9111	0.0007	0.8780	0.0010	0.8639	0.0016
RSF(logrank-Cox)	0.9133	0.0006	0.8808	0.0009	0.8652	0.0009
RSF(logrank-Aalen)	0.9202	0.0004	0.8887	0.0008	0.8689	0.0008
RSF(logrankscore- nonparametric)	0.8849	0.0009	0.8477	0.0010	0.8165	0.0012
RSF(logrankscore-Cox)	0.8868	0.0008	0.8481	0.0009	0.8268	0.0011
RSF(logrankscore-Aalen)	0.8927	0.0007	0.8575	0.0008	0.8309	0.0010
Cforest(non parametric)	0.8333	0.0010	0.8205	0.0020	0.7931	0.0024
Cforest(Cox)	0.8319	0.0009	0.8249	0.0010	0.8126	0.0011
Cforest(Aalen)	0.8319	0.0009	0.8249	0.0010	0.8126	0.0011

Table 2. The mean and standard error values according to the C-Index criteria of RSF and Cforest method for different survival times in cases where n=200 and the proportional hazard assumption is provided

C-Index (scenario 1, n=200)	Survival Time					
	0.5		2		3.5	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.9675	0.0008	0.8786	0.0009	0.8639	0.0016
RSF (logrank-Cox)	0.9678	0.0007	0.8908	0.0008	0.8652	0.0009
RSF (logrank-Aalen)	0.9680	0.0005	0.8987	0.0007	0.8689	0.0008
RSF (logrankscore- nonparametric)	0.8850	0.0009	0.8475	0.0010	0.8170	0.0017
RSF(logrankscore-Cox)	0.8870	0.0008	0.8485	0.0009	0.8281	0.0011
RSF (logrankscore-Aalen)	0.8935	0.0007	0.8590	0.0008	0.8309	0.0010
Cforest(nonparametric)	0.8689	0.0010	0.8249	0.0020	0.7931	0.0024
Cforest (Cox)	0.8689	0.0010	0.8596	0.0010	0.8126	0.0011
Cforest (Aalen)	0.8769	0.0009	0.8596	0.0010	0.8126	0.0011

Table 3. The mean and standard error values according to the C-Index criteria of RSF and Cforest method for different survival times in cases where n=300 and the proportional hazard assumption is provided

C-Index (scenario 1, n=300)	Survival Time					
	0.5		2		3.5	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.9838	0.0009	0.8886	0.0010	0.8739	0.0012
RSF(logrank-Cox)	0.9857	0.0006	0.8908	0.0009	0.8752	0.0006
RSF (logrank-Aalen)	0.9869	0.0004	0.8990	0.0007	0.8789	0.0003
RSF (logrankscore- nonparametric)	0.8950	0.0009	0.8475	0.0012	0.8276	0.0014
RSF (logrankscore-Cox)	0.8965	0.0006	0.8485	0.0010	0.8381	0.0011
RSF (logrankscore-Aalen)	0.8970	0.0005	0.8590	0.0008	0.8509	0.0010
Cforest(nonparametric)	0.8879	0.0010	0.8249	0.0020	0.7931	0.0024
Cforest(Cox)	0.8889	0.0010	0.8596	0.0010	0.8125	0.0013
Cforest (Aalen)	0.8969	0.0009	0.8596	0.0010	0.8126	0.0011

Table 4. The mean and standard error values according to the C-Index criteria of RSF and Cforest method for different survival times in cases where n=100 and the proportional hazard assumption is not provided

C-Index (scenario 2, n=100)	Survival Time					
	0.5		2		3.5	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.9802	0.0002	0.9434	0.0004	0.9076	0.0006
RSF (logrank-Cox)	0.9801	0.0002	0.9435	0.0004	0.9079	0.0006
RSF (logrank-Aalen)	0.9834	0.0001	0.9552	0.0002	0.9282	0.0002
RSF (logrankscore- nonparametric)	0.9801	0.0002	0.8576	0.0015	0.8265	0.0009
RSF (logrankscore-Cox)	0.9799	0.0002	0.8581	0.0009	0.8368	0.0006
RSF (logrankscore-Aalen)	0.9825	0.0002	0.8775	0.0008	0.8409	0.0002
Cforest(nonparametric)	0.9323	0.0012	0.8602	0.0010	0.8342	0.0011
Cforest (Cox)	0.9324	0.0012	0.8603	0.0010	0.8361	0.0009
Cforest (Aalen)	0.9336	0.0010	0.8605	0.0009	0.8398	0.0008

Table 5. The mean and standard error values according to the C-Index criteria of RSF and Cforest method for different survival times in cases where n=200 and the proportional hazard assumption is not provided

C-Index (scenario 2, n=200)	Survival Time					
	0.5		2		3.5	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.9765	0.0009	0.9385	0.0005	0.9056	0.0010
RSF(logrank-Cox)	0.9789	0.0003	0.9436	0.0003	0.9083	0.0004
RSF (logrank-Aalen)	0.9795	0.0001	0.9462	0.0002	0.9152	0.0003
RSF(logrankscore- nonparametric)	0.8950	0.0017	0.8675	0.0015	0.8275	0.0018
RSF(logrankscore-Cox)	0.8970	0.0005	0.8785	0.0007	0.8381	0.0010
RSF (logrankscore-Aalen)	0.8995	0.0003	0.8990	0.0005	0.8409	0.0007
Cforest(nonparametric)	0.9015	0.0015	0.8194	0.0016	0.8002	0.0012
Cforest (Cox)	0.9028	0.0010	0.8291	0.0009	0.8013	0.0008
Cforest (Aalen)	0.9041	0.0002	0.8291	0.0009	0.8035	0.0005

Table 6. The mean and standard error values according to the C-Index criteria of RSF and Cforest method for different survival times in cases where n=300 and the proportional hazard assumption is not provided

C-Index (scenario 2, n=300)	Survival Time					
	0.5		2		3.5	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.9948	0.0008	0.9100	0.0012	0.8839	0.0011
RSF (logrank-Cox)	0.9957	0.0005	0.9108	0.0010	0.8852	0.0006
RSF (logrank-Aalen)	0.9969	0.0003	0.9120	0.0004	0.8889	0.0002
RSF (logrankscore- nonparametric)	0.8970	0.0009	0.8475	0.0010	0.8576	0.0012
RSF (logrankscore-Cox)	0.8985	0.0006	0.8585	0.0008	0.8581	0.0011
RSF (logrankscore-Aalen)	0.8990	0.0005	0.8690	0.0007	0.8609	0.0010
Cforest(nonparametric)	0.8979	0.0013	0.8749	0.0020	0.7998	0.0022
Cforest (Cox)	0.8989	0.0010	0.8896	0.0010	0.8125	0.0009
Cforest (Aalen)	0.8989	0.0009	0.8896	0.0010	0.8126	0.0005

Table 7. The mean and standard error values according to the IBS criteria of RSF and Cforest method for different survival times in cases where n=100 and the proportional hazard assumption is provided

IBS (scenario 1, n=100)	AppErr		BootCvErr		NoInfErr		Boot632plusErr	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.0298	0.0056	0.1660	0.0360	0.2850	0.0060	0.1389	0.0279
RSF (logrank-Cox)	0.0292	0.0042	0.1646	0.0253	0.2814	0.0050	0.1384	0.0258
RSF (logrank-Aalen)	0.0286	0.0021	0.1630	0.0156	0.2787	0.0038	0.1372	0.0168
RSF (logrankscore-nonparametric)	0.0300	0.0070	0.1745	0.0380	0.3050	0.0120	0.1439	0.0289
RSF (logrankscore-Cox)	0.0295	0.0068	0.1736	0.0293	0.3014	0.0090	0.1424	0.0280
RSF (logrankscore-Aalen)	0.0287	0.0035	0.1720	0.0166	0.2987	0.0058	0.1412	0.0267
Cforest(nonparametric)	0.1010	0.0200	0.1534	0.0210	0.2576	0.0104	0.1386	0.0225
Cforest (Cox)	0.1007	0.0191	0.1512	0.0198	0.2399	0.0098	0.1384	0.0210
Cforest (Aalen)	0.0974	0.0120	0.1489	0.0127	0.2342	0.0090	0.1363	0.0133

*AppErr: apparent prediction, BootCvErr: Boostrap Cross-Validation prediction, noinferr: ignorance prediction error, Boot632plusErr: 0.632+ prediction

Table 8. The mean and standard error values according to the IBS criteria of RSF and Cforest method for different survival times in cases where n=200 and the proportional hazard assumption is provided

IBS (scenario 1, n=200)	AppErr		BootCvErr		NoInfErr		Boot632plusErr	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.0288	0.0036	0.1640	0.0335	0.2845	0.0050	0.1379	0.0259
RSF (logrank-Cox)	0.0282	0.0022	0.1626	0.0233	0.2794	0.0045	0.1373	0.0238
RSF (logrank-Aalen)	0.0276	0.0019	0.1615	0.0145	0.2767	0.0028	0.1362	0.0148
RSF(logrankscore-nonparametric)	0.0290	0.0070	0.1645	0.0367	0.3030	0.0110	0.1418	0.0260
RSF (logrankscore-Cox)	0.0285	0.0068	0.1636	0.0291	0.2914	0.0085	0.1412	0.0270
RSF (logrankscore-Aalen)	0.0277	0.0035	0.1620	0.0164	0.2867	0.0050	0.1409	0.0167
Cforest(nonparametric)	0.1008	0.0197	0.1514	0.0187	0.2456	0.0094	0.1376	0.0215
Cforest (Cox)	0.0987	0.0172	0.1508	0.0166	0.2297	0.0066	0.1368	0.0200
Cforest (Aalen)	0.0961	0.0110	0.1469	0.0115	0.2210	0.0050	0.1338	0.0113

*AppErr: apparent prediction, BootCvErr: Boostrap Cross-Validation prediction, noinferr: ignorance prediction error, Boot632plusErr: 0.632+ prediction

Table 9. The mean and standard error values according to the IBS criteria of RSF and Cforest method for different survival times in cases where n=300 and the proportional hazard assumption is provided

IBS (scenario 1, n=300)	AppErr		BootCvErr		NoInfErr		Boot632plusErr	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.0275	0.0032	0.1628	0.0325	0.2275	0.0045	0.1365	0.0247
RSF (logrank-Cox)	0.0271	0.0020	0.1616	0.0228	0.2283	0.0037	0.1355	0.0222
RSF (logrank-Aalen)	0.0265	0.0012	0.1609	0.0136	0.2247	0.0018	0.1320	0.0136
RSF (logrankscore-nonparametric)	0.0285	0.0063	0.1635	0.0357	0.3020	0.0100	0.1417	0.0249
RSF (logrankscore-Cox)	0.0283	0.0064	0.1626	0.0271	0.2904	0.0075	0.1410	0.0260
RSF (logrankscore-Aalen)	0.0276	0.0027	0.1617	0.0154	0.2357	0.0040	0.1401	0.0147
Cforest(nonparametric)	0.1006	0.0177	0.1513	0.0167	0.2454	0.0094	0.1366	0.0215
Cforest (Cox)	0.0977	0.0162	0.1504	0.0146	0.2294	0.0066	0.1358	0.0200
Cforest (Aalen)	0.0951	0.0106	0.1459	0.0105	0.2308	0.0050	0.1328	0.0113

*AppErr: apparent prediction, BootCvErr: Boostrap Cross-Validation prediction, noinferr: ignorance prediction error, Boot632plusErr: 0.632+ prediction

Table 10. The mean and standard error values according to the IBS criteria of RSF and Cforest method for different survival times in cases where n=100 and the proportional hazard assumption is not provided

IBS (scenario 2, n=100)	AppErr		BootCvErr		NoInfErr		Boot632plusErr	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.0296	0.0056	0.1658	0.0350	0.2840	0.0054	0.1379	0.0269
RSF (logrank-Cox)	0.0290	0.0042	0.1642	0.0243	0.2804	0.0044	0.1374	0.0248
RSF (logrank-Aalen)	0.0285	0.0021	0.1627	0.0146	0.2777	0.0032	0.1362	0.0158
RSF (logrankscore-nonparametric)	0.0297	0.0070	0.1743	0.0370	0.3040	0.0116	0.1429	0.0279
RSF (logrankscore-Cox)	0.0292	0.0068	0.1732	0.0283	0.3004	0.0087	0.1414	.0270
RSF(logrankscore-Aalen)	0.0295	0.0035	0.1718	0.0156	0.2977	0.0054	0.1402	0.0257
Cforest(nonparametric)	0.1007	0.0200	0.1531	0.0200	0.2566	0.0102	0.1376	0.0215
Cforest (Cox)	0.1005	0.0191	0.1508	0.0188	0.2389	0.0097	0.1374	0.0200
Cforest (Aalen)	0.0964	0.0120	0.1479	0.0117	0.2332	0.0087	0.1353	0.0123

*AppErr: apparent prediction, BootCvErr: Boostrap Cross-Validation prediction, noinferr: ignorance prediction error, Boot632plusErr: 0.632+ prediction

Table 11. The mean and standard error values according to the IBS criteria of RSF and Cforest method for different survival times in cases where n=200 and the proportional hazard assumption is not provided

IBS (scenario 2, n=200)	AppErr		BootCvErr		NoInfErr		Boot632plusErr	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.0278	0.0026	0.1630	0.0325	0.2835	0.0040	0.1369	0.0249
RSF (logrank-Cox)	0.0272	0.0012	0.1616	0.0223	0.2784	0.0035	0.1363	0.0228
RSF (logrank-Aalen)	0.0266	0.0009	0.1605	0.0135	0.2757	0.0018	0.1352	0.0138
RSF (logrankscore-nonparametric)	0.0280	0.0060	0.1635	0.0357	0.3020	0.0100	0.1408	0.0249
RSF (logrankscore-Cox)	0.0275	0.0058	0.1626	0.0281	0.2904	0.0075	0.1405	0.0260
RSF (logrankscore-Aalen)	0.0267	0.0025	0.1610	0.0154	0.2857	0.0040	0.1402	0.0157
Cforest(nonparametric)	0.1007	0.0187	0.1504	0.0177	0.2446	0.0084	0.1366	0.0205
Cforest (Cox)	0.0977	0.0162	0.1506	0.0156	0.2287	0.0056	0.1358	0.0195
Cforest (Aalen)	0.0951	0.0100	0.1459	0.0105	0.2300	0.0043	0.1328	0.0108

*AppErr: apparent prediction, BootCvErr: Boostrap Cross-Validation prediction, noinferr: ignorance prediction error, Boot632plusErr: 0.632+ prediction

Table 12. The mean and standard error values according to the IBS criteria of RSF and Cforest method for different survival times in cases where n=300 and the proportional hazard assumption is not provided

IBS (scenario 2, n=300)	AppErr		BootCvErr		NoInfErr		Boot632plusErr	
	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$	$\bar{x}$	$s_{\bar{x}}$
RSF(logrank-nonparametric)	0.0265	0.0022	0.1618	0.0315	0.2817	0.0035	0.1358	0.0237
RSF (logrank-Cox)	0.0261	0.0010	0.1606	0.0218	0.2773	0.0027	0.1352	0.0222
RSF (logrank-Aalen)	0.0255	0.0002	0.1608	0.0126	0.2737	0.0008	0.1347	0.0126
RSF (logrankscore-nonparametric)	0.0275	0.0053	0.1625	0.0347	0.3010	0.0098	0.1407	0.0239
RSF (logrankscore-Cox)	0.0273	0.0054	0.1616	0.0261	0.2902	0.0065	0.1400	0.0250
RSF (logrankscore-Aalen)	0.0266	0.0017	0.1607	0.0144	0.2847	0.0030	0.1399	0.0137
Cforest(nonparametric)	0.1005	0.0167	0.1503	0.0157	0.2444	0.0084	0.1356	0.0205
Cforest (Cox)	0.0967	0.0152	0.1501	0.0136	0.2305	0.0056	0.1348	0.0197
Cforest (Aalen)	0.0941	0.0104	0.1449	0.0102	0.2284	0.0040	0.1318	0.0103

*AppErr: apparent prediction, BootCvErr: Boostrap Cross-Validation prediction, noinferr: ignorance prediction error, Boot632plusErr: 0.632+ prediction

5. Results and discussion

In this study, Cforest method (Hothorn and et al. 2006a), which aims to minimize the proposed empirical risk function for right-censored data and a community with a low correlation structure by creating different trees, and RSF method (Ishwaran and et al. 2008a), which is an extension of Brieman's random forest method for right-censored data, are compared according to C-Index and IBS criteria.

According to the C-Index criterion; in all cases, RSF method has higher mean C -Index values and lower standard error values than Cforest method. When we examined the sample size, it was observed that the mean C-Index values for both scenarios and both methods were increased and standard error values were decreased with the increase in sample size. It is observed gave the best results for the RSF method and the non parametric estimator has lower mean C-Index values than the Aalen estimator and Cox estimator. In the Cforest method, it was observed that the nonparametric estimator had lower C-Index mean values and similar results were obtained in Cox and Aalen estimator. When the RSF method was examined in terms of two different separation criteria, it was determined that the logrank distinction had higher mean C-Index values and lower standard error values. Compared to the situation in which the proportional hazard assumption is provided and not provided, it has been observed that both methods perform better in the absence of the proportional hazard assumption.

However, when the proportional hazard assumption provided, there has been a further decrease in the mean C-Index values for the RSF method compared to the Cforest method.

According to the IBS criterion; for all cases, in both scenarios, and for all $\hat{G}$ estimation methods (Cox, Aalen and Nonparametric), the RSF method has lower mean and standard error values than the Cforest method. With the increase in sample size, model performance was observed to increase in all cases according to IBS criteria. For all methods and for both scenarios, Aalen estimator has a lower error value than nonparametric estimator and Cox estimator. When examined according to RSF separation criteria, it was determined that logrank distinction criteria had lower IBS mean values and standard error values. In this study, it was observed that all methods performed better in the case that the proportional hazard assumption is not provided, compared to the case that the proportional hazard assumption is provided.

Mogensen and et al. (2012) examined the performance of the RSF, Cforest and Cox regression models using the "cost" data set included in the PEC package. As a result, while some cross-validation methods found the performance of the methods to be similar, some cross-validation methods found the performance of the RSF method to be higher.

Gerds and Schumacher (2007) used marginal Kaplan-Meier, Cox, Aalen and nonparametric estimators for calculating IBS values. However, if the censored mechanism of the Kaplan-Meier estimator is dependent on the common variables, it gives error. For this reason, they recommended the use of three other predictors for the case where the censored survival function is dependent on the common variables. In their simulation study, they stated that the Aalen estimator was better than the Cox estimator. The results of our simulation study showed that the Aalen estimator has better performance in both methods. Ciampi (1986) proposed the use of logrank test statistic to compare two child nodes in decision trees. Ishwaran and et al.(2008a) stated that the model obtained by using logrank criteria is higher than C-Index value when they apply the RSF method on 11 sets of data according to different separation rules. According to the results of our simulation study, it was determined that the logrank distinction criteria showed higher performance than the logrank score distinction criteria in the case where the proportional hazard assumption is provided and not provided in the RSF method.

As a result, it has been shown that the RSF method performs better than the Cforest. For both methods, it can be said that the Aalen estimator performs better than the other estimators. The performance of both methods was better if the proportional hazard assumption was not provided. In addition, the RSF method shows that the logrank distinction criteria, which is one of two different separation criteria, performs better than the logrank score distinction criteria.

References

Breiman, L., Friedman, J., Olshen, R. and Stone, C., (1984). Classification and regression trees. Wadsworth Int. *Group*, 37(15), pp. 237–251.

Ciampi, A., Thiffault, J., Nakache, J. P. and Asselain, B., (1986). Stratification by stepwise regression, correspondence analysis and recursive partition: a comparison of three methods of analysis for survival data with covariates. *Computational statistics & data analysis*, 4(3), pp. 185–204.

Gerds, T. A., Schumacher, M., (2007). Consistent estimation of the expected Brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6), pp. 1029–1040.

Hothorn, T., Bühlmann, P., Dudoit, S., Molinaro, A. and Van Der Laan, M. J., (2006a). Survival ensembles. *Biostatistics*, 7(3), pp. 355–373.

Hothorn, T., Hornik, K. and Zeileis, A., (2006b). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3), pp. 651–674.

Hothorn, T., Hornik, K. and Zeileis, A., (2005). party: A laboratory for recursive part (y) itioning. *R package version 0.3-2*.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H. and Lauer, M. S., (2008). Random survival forests. *The annals of applied statistics*, 2(3), pp. 841–860.

Ishwaran, H., Kogalur, U. B., (2019). randomForestSRC: Fast unified random forests for survival, regression, and classification (RF-SRC). *R Package Version*, 2(1).

Mogensen, U. B., Ishwaran, H. and Gerds, T. A., (2012). Evaluating random forests for survival analysis using prediction error curves. *Journal of statistical software*, 50(11), 1.

Safavian, S. R., Landgrebe, D., (1991). A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3), pp. 660–674.

Segal, M. R., (1988). Regression trees for censored data. *Biometrics*, pp. 35–47.

Team, R. C., (2013). R: A language and environment for statistical computing.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 39–54, DOI 10.21307/stattrans-2022-003
Received – 03.12.2019; accepted – 31.03.2021

Estimating the population mean using a complex sampling design dependent on an auxiliary variable

Arijit Chaudhuri[1], **Sonakhya Samaddar**[2]

ABSTRACT

In surveying finite populations, the simplest strategy to estimate a population total without bias is to employ Simple Random Sampling (SRS) with replacement (SRSWR) and the expansion estimator based on it. Anything other than that including SRS Without Replacement (SRSWOR) and usage of the expansion estimator is a complex strategy. We examine here (1) if from a complex sample at hand a gain in efficiency may be unbiasedly estimated comparing the "rival population total-estimators" for the competing strategies and (2) how suitable model-expected variances of rival estimators compete in magnitude as examined numerically through simulations.

Key words: Des Raj and symmetrized Des Raj estimator and associated variance, Hansen-Hurwitz estimation and variance, Hartley-Ross, Horvitz-Thompson, Lahiri-Midzuno-Sen, Murthy, Rao-Hartley-Cochran procedures vis-a-vis SRSWOR and SRSWR.
AMS Subject classification: 62 DO5.

1. Introduction

Stratifed SRSWOR is supposed to outperform unstratified SRSWOR because the conventional unbiased estimator of the population mean in the former has a variance as a function of the 'Within Sum of Squares' contrasted with the latter involving the 'Total Sum of Squares' if the strata are well constructed and maybe, effectively controlled Between strata variability. Using the survey data from a stratified SRSWOR it is well known vide Cochran (1977) and JNK Rao (1961) how the gain in stratification may duly be estimated vis-a-vis unstratified SRSWOR.

It is our interest to extend this approach covering a few competitive pairs of strategies in each of which it is difficult to work out plausible variance formulae in closed form illustrated in Section 2 below.

Covering pairs of sampling strategies for estimating population totals when variance formulae are available for unbiased estimators, we intend to examine how more complicated complex strategies may be justified from the efficiency gaining point of view vis-a-vis SRSWR and SRSWOR as the basic procedures by postulating simplified regression models thereby working out their model-based expected values of the variances of rival unbiased estimators for the population total.

Details are given in the Section 3 below.

[1]Indian Statistical Institute, Kolkata, India. The corresponding author.
E-mail: arijitchaudhuri1@rediffmail.com.ORCID: https://orcid.org/0000-0002-4305-7686.
[2]Indian Statistical Institute, Kolkata, India. The corresponding author. E-mail: sonakhya003@gmail.com.
ORCID: https://orcid.org/0000-0002-9462-0520.

A comparative study by simulations is presented in Section 4. Comments are also stated there.

To our knowledge the literature covers no follow-up of JNK Rao's (1961) approach treating any other strategies. Our Section 2 below is a novel exercise removing this deficiency taking account of several worthy alternatives. Secondly, considering a simple special case of Fairfield Smith's (1938) popular super-population model and bringing several useful and popular sampling strategies under this umbrella we, as a novelty, study, by simulation, how the numerical model-expected design variances of unbiased estimators of finite population totals (or means) for complex and simple sampling strategies fare among each other.

2. Estimating Gain in Efficiency

2.1. (PPSWOR, Des Raj Estimator) strategy versus (SRSWOR, Expansion Estimator)

Suppose y is a variable of interest taking values y_i for the respective units i of a finite population $U = (1,\ldots,i,\ldots,N)$, with a total $Y = \sum_{i=1}^{N} y_i$.

Let positive values x_i of another positively correlated variable x be all known for the units i of U, with a total $X = \sum_{i=1}^{N} x_i$ and $p_i = \frac{x_i}{X}$ be the unit-wise normed size measures. $\overline{X}$, $\overline{Y}$ denote the population means of x and y.

Probability proportional to size measures x_i (PPS) without replacement (PPSWOR) sample selection method is implemented by selecting a number, say, n($\geq$2) units from U ordered as the $1^{st}, 2^{nd}, \ldots, n^{th}$, namely $i_1, i_2, \ldots, i_j, \ldots, i_n$ with respective probabilities

$$p_{i1}, \frac{p_{i2}}{1-p_{i1}}, \ldots, \frac{p_{ij}}{1-p_{i1}-\ldots-p_{ij-1}}$$

$j = 1,2,\ldots,n.$

Then, Des Raj's unbiased estimator for Y is

$$t_D = \frac{1}{n}(t_1 + t_2 + \ldots + t_n)$$

with

$$\begin{aligned} t_1 &= \frac{y_{i1}}{p_{i1}}, \\ t_2 &= y_{i1} + \frac{y_{i2}}{p_{i2}}(1-p_{i1}), \ldots, \\ t_j &= y_{i1} + y_{i2} + \ldots + \frac{y_{ij}}{p_{ij}}(1-p_{i1}-p_{i2}-\ldots-p_{ij-1}), j = 1,2,\ldots,n. \end{aligned}$$

The formula for the exact variance of t_D is given by Roychoudhury (1957). But its closed form expression is pretty complicated. Nevertheless, an unbiased estimator for $V(t_D)$

is given by Des Raj(1956) as

$$v(t_D) = \frac{1}{2n^2(n-1)} \sum_{j=1}^{n} \sum_{k=1,k\neq j}^{n} (t_j - t_k)^2$$

which is pretty simple in form.

Suppose a PPSWOR sample chosen as above is at hand as $s = (i_1, i_2, \ldots, i_n)$ along with the values $y_{i1}, y_{i2}, \ldots, y_{in}$.

Suppose we consider a comparable strategy composed of an SRSWOR sample s_{WOR} of size n and the expansion estimator based on it as

$$N\overline{y} = \frac{N}{n} \sum_{i \in s_{WOR}} y_i$$

with variance

$$V_{SWOR}(N\overline{y}) = \frac{(N-n)N^2}{Nn(N-1)} \sum_{i=1}^{N} (y_i - \overline{Y})^2$$

where $\overline{y}$ denotes the sample mean.

Then, an unbiased estimator for this is derived as follows: We have

$$V(t_D) = E(t_D^2) - Y^2$$

So an unbiased estimator for Y^2 is

$$\hat{Y^2} = t_D^2 - v(t_D) \qquad \ldots(2.1)$$

Also an unbiased estimator for $\sum_1^N y_i^2$ is $t_D(y^2)$, which is t_D as above with every y in t_D replaced by corresponding y^2. So, an unbiased estimator for $V(N\overline{y})$ is

$$v_1 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2}{N-1} \left[t_D(y^2) - \frac{\hat{Y^2}}{N} \right] \qquad \ldots(2.2)$$

with $\hat{Y^2}$ as given in (2.1).

Then $G_1 = v_1 - v(t_D)$ unbiasedly estimates gain in efficiency of (PPSWOR,t_D) over (SRSWOR,$N\overline{y}$).

2.2. (PPSWOR, Symmetrized Des Raj Estimator) versus (SRSWOR, Expansion Estimator)

Given the ordered sample as in section (2.1) as $s = (i_1, i_2, \ldots, i_n)$ and Des Raj's estimator $t_D = t_D(s)$ based on this ordered s, let s^* be the set of all samples obtained by permuting the n units in s in all possible $n!$ ways and

$$p(s^*) = \sum_{s \to s^*} p(s),$$

writing $\sum_{s\to s^*}$ to denote the sum over all possible samples in the set s^*. Then

$$t^*_{SD} = t^*_{SD}(s') = \frac{\sum_{s\to s^*} p(s)t_D(s)}{\sum_{s\to s^*} p(s)} = t^*_{SD}(s),$$

say, for any member s' in set s^* is the 'Symmetrized Des Raj' estimator for Y. Also, it is well known, vide (Chaudhuri(2010), p19) that

$$V(t^*_{SD}(s)) = V(t_D(s)) - E(t_D - t^*_{SD})^2.$$

Hence, an unbiased estimator for $V(t^*_{SD})$ is

$$v(t^*_D) = v(t_D) - (t_D - t^*_{SD})^2.$$

If the survey data $(s', y_i | i \in s')$ are at hand, an unbiased estimate for $V_{SWOR}(N\bar{y})$ follows as

$$v_2 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2}{(N-1)} \left[t^*_{SD}(y^2) - \frac{(\hat{Y}^2)'}{N}\right] \qquad \ldots(2.3)$$

Writing $t^*_{SD}(y^2)$ as $t^*_{SD}(s')$ with each y_i in $t^*_{SD}(s')$ replaced by y_i^2 and

$$(\hat{Y}^2)' = (t^*_{SD}(s'))^2 - v_2(t^*_{SD}) \qquad \ldots(2.4)$$

2.3. (Lahiri-Midzuno-Sen sampling with Ratio Estimator) versus (SRSWOR, Expansion Estimator)

Lahiri-Midzuno-Sen's (1951, 1952, 1953) or LMS sample is selected by choosing on the first draw from U a unit i with selection probability p_i followed by an SRSWOR in (n-1) draws from the remaining (N-1) units excluding the first chosen unit i from U.

Then, $t_R = X\frac{\bar{y}}{\bar{x}}$ is the exact unbiased ratio estimator for Y based on such a sample, with $\bar{x}$ denoting the sample mean of x.

Vide Chaudhuri(2010) an exactly unbiased estimator of variance of t_R is

$$v(t_R) = \sum^{N} \sum_{i<j=1}^{N} a_{ij} \frac{I_{sij}}{\sum_{i\in s} p_i} \left(\frac{N-1}{n-1} - \frac{1}{\sum_{i\in s} p_i}\right);$$

here s is the LMS sample of size n, and

$$I_{sij} = \begin{cases} 1 & i,j \in s \\ 0 & otherwise \end{cases}$$

and

$$a_{ij} = p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2;$$

also $\pi_i =$ the inclusion probability of i in an LMS sample is given by

$$\pi_i = \frac{N-n}{N-1} p_i + \frac{n-1}{N-1}$$

and

$$\pi_{ij} = \frac{N-n}{(N-1)(N-2)}(p_i + p_j) + \frac{(n-2)(n-1)}{(N-1(N-2))}$$

is the inclusion probability of i and j in an LMS sample of size n. So, an unbiased estimator of $V(N\overline{y})$ from the LMS sample is

$$\hat{V}_{SWOR}(N\overline{y}) = v_3 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{N^2}{(N-1)} \left[\sum_{i\in s} \frac{y_i^2}{\pi_i} - \frac{1}{N}(t_R^2 - v(t_R))\right]$$

because

$$V(t_R) = E(t_R^2) - Y^2$$

and Y^2 is unbiasedly estimated by $t_R^2 - v(t_R)$. So, $v_3 - v(t_R)$ unbiasedly estimates the gain in efficiency of (LMS,t_R) over (SRSWOR,$N\overline{y}$)

2.4. (SRSWOR,Hartley-Ross estimator) versus (SRSWOR, Expansion estimator)

Based on an SRSWOR s of size n an unbiased estimator for Y given by Hartley and Ross (1954) is

$$\hat{Y}_{HR} = N\left[\overline{r} + \left(\frac{N-1}{N}\right)\left(\frac{n}{n-1}\right)\frac{1}{\overline{\overline{X}}}(\overline{y} - \overline{rx})\right] = N\left[\overline{r} + c(\overline{y} - \overline{rx})\right],$$

say, writing $\overline{r} = \frac{1}{n}\sum_{i\in s}\frac{y_i}{xi}$ and $\overline{x}$, $\overline{y}$ are sample means of x and y and $\overline{\overline{X}}$ is the population mean of x.

An unbiased estimator for $V(\hat{Y}_{HR})$ is given by

$$v(\hat{Y}_{HR}) = (\hat{Y}_{HR})^2 - \left[\frac{N}{n}\sum_{i\in s} y_i^2 + \frac{N(N-1)}{n(n-1)}\sum\sum_{i\neq j\in s} y_i y_j\right]$$

because for SRSWOR $\pi_i = \frac{n}{N}\ \forall i$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}\ \forall i \neq j$

An unbiased estimator for $V(N\overline{y})$ from an SRSWOR s of size n is

$$v_4 = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{N^2}{(n-1)}\sum_{i\in s}(y_i - \overline{y})^2.$$

So $v_4 - v(\hat{Y}_{HR})$ tells us how much we may gain in efficiency on using $\hat{Y}_{HR}$ rather than $N\overline{y}$.

In Section 3 below we consider situations when for complex surveys variances of unbiased estimators for Y have manageably elegant forms.

3. How under a simple model expected variances fare relative to each other

Model

We assume that

$$y_i = \beta x_i + \varepsilon_i,\ i \in U = (1,2,3,\ldots,N)$$

Here β is an arbitrary unknown constant which determines y's dependence on x's. $x_i's$ are auxilliary variables which are known for all population units. $\varepsilon_i's$ are error terms with zero mean and some common variance τ^2, which is also not fixed.
Every expectation that we have taken in the upcoming Sections 3.1 to 3.9 are based on the above mentioned model.

This model is a simple special case of the well-known popular Fairfield Smith's (1938) super-population model under which the model-variance of ε_i is $\tau^2 x_i^\gamma$ for i=1,2,...,N. In the literature most strategies are treated utilizing this model and the literature on comparison among model expected variances of design- unbiased estimators of finite population totals (or means) is rather vast. But in this paper we may draw attention to the following few, namely the text by Sarndal, Swensson and Wretman (1992) and a few papers in peer-reviewed journals namely by JNK Rao and Bayless, D.L. (1969), JNK Rao and Bayless, D.L. (1970), TJ Rao (1967) and Chaudhuri and Arnab (1979). The last-mentioned paper, Chaudhuri and Arnab (1979), is worthy of attention because in it, expressing Model- (Fairfield Smith's)- expected variances of ratio estimator based on LMS scheme by E_1, that of Rao-Hartley-Cochran estimator by E_2 and that of Horvitz-Thompson estimator based on an IPPS sample by E_3, it is shown that
(i) $E_1 < E_2 < E_3$ if $\gamma < 1$,
(ii) $E_1 > E_2 > E_3$ if $\gamma > 1$ and
(iii) $E_1 = E_2 = E_3$ if $\gamma = 1$.

3.1. Strategy 1: (SRSWR, Expansion Estimator)

For SRSWR in n draws from population of size N the expansion estimator $N\bar{y}$ is unbiased for $Y = \sum_{i=1}^{N} y_i$ with variance

$$V(N\bar{y}) = \frac{N^2}{n}\sigma^2 = \frac{N}{n}\sum_{i}^{N}(y_i - \overline{Y})^2; \quad \sigma^2 = \frac{1}{N}\sum_{1}^{N}(y_i - \overline{Y})^2.$$

Under a model its expected value is

$$\mathscr{E}(V(N\bar{y})) = \frac{N}{n}\mathscr{E}\left[\sum_{i=1}^{N}(y_i - \overline{Y})^2\right]$$

$\mathscr{E}$ denotes generically a model-based expectation operator.

Then

$$\mathscr{E}(V(N\bar{y})) = \frac{N(N-1)}{n}\left[\tau^2 + \beta^2 S_{xx}\right] = (srswr)$$

where

$$S_{xx} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2.$$

3.2. Strategy 2: (SRSWOR,$N\overline{y}$)

We have in this case

$$V(N\overline{y}) = \frac{N^2(N-n)}{Nn(N-1)}\sum_{i=1}^{N}(y_i - \overline{Y})^2.$$

And thus,

$$\mathscr{E}(V(N\overline{y})) = \frac{N(N-n)}{n}(\tau^2 + \beta^2 S_{xx}) = (srswor).$$

3.3. Strategy 3: (PPSWR, Hansen-Hurwitz Estimator t_{HH})

The Hansen Hurwitz estimator (1943) is given by

$$t_{HH} = \frac{1}{n}\sum_{r=1}^{n}\frac{y_r}{p_r},$$

with
y_r=y-value for the unit chosen on r-th draw
p_r=probability of the unit being chosen on r-th draw

$$V(t_{HH}) = \frac{1}{n}\left[\sum_{1}^{N}\frac{y_i^2}{p_i} - Y^2\right]$$

with

$$\mathscr{E}(V(t_{HH})) = \frac{\tau^2}{n}\left(N\overline{X}\sum_{1}^{N}\frac{1}{x_i} - N\right) = (ppswr).$$

3.4. Strategy 4: (PPSWR, Horvitz-Thompson Estimator

For PPSWR sampling in n draws the inclusion-probabilities are

$$\pi_i = 1 - (1-p_i)^n$$

$$\pi_{ij} = 1 - (1-p_i)^n - (1-p_j)^n + (1-p_i-p_j)^n.$$

Following Chaudhuri and Pal (2003) the Horvitz & Thompson's (1952) estimator (HTE), $t_{HT} = \sum_{i\in s}\frac{y_i}{\pi_i}$ based on a PPSWR sample s in n draws has the variance

$$V(t_{HT})_{PPS} = \sum_{i<j}^{N}\sum^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_{1}^{N}\frac{y_i^2}{\pi_i}\alpha_i$$

with $\alpha_i = 1 + \frac{1}{\pi_i}\sum_{j\neq i}\pi_{ij} - \sum_1^N \pi_i$. Then

$$\mathscr{E}(V(t_{HT})_{PPS}) = \beta^2\left[\sum_1^N \frac{x_i^2}{\pi_i} + \sum\sum_{i\neq j} x_i x_j \frac{\pi_{ij}}{\pi_i\pi_j} - X^2\right] + \tau^2\left(\sum_1^N \frac{1}{\pi_i} - N\right) + \beta^2\sum_1^N \alpha_i\frac{x_i^2}{\pi_i} + \tau^2\sum_1^N \frac{\alpha_i}{\pi_i} = (ppswrht).$$

3.5. Strategy 5: (SRSWR,HTE)

For SRSWR in n draws the inclusion-probabilities are

$$\pi_i = 1 - (\frac{N-1}{N})^n$$

$$\pi_{ij} = 1 - 2(\frac{N-1}{N})^n + (\frac{N-2}{N})^n.$$

For the HTE based on SRSWR in n draws the variance is

$$V(t_{HT})_{SRS} = \sum_1^N\sum_{i<j}^N(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_1^N \frac{y_i^2}{\pi_i}\alpha_i$$

with

$$\mathscr{E}(V(t_{HT})_{SRS}) = \beta^2\left[\sum_1^N \frac{x_i^2}{\pi_i} + \sum\sum_{i\neq j} x_i x_j \frac{\pi_{ij}}{\pi_i\pi_j} - X^2\right] + \tau^2\left(\sum_1^N \frac{1}{\pi_i} - N\right) + \beta^2\sum_1^N \alpha_i\frac{x_i^2}{\pi_i} + \tau^2\sum_1^N \frac{\alpha_i}{\pi_i} = (srswrht).$$

3.6. Strategy 6: (SRSWR, N times the mean of the sampled distinct units only

From Chaudhuri (2010, pp. 35-36)we know that the sample mean of the distinct units in a sample s chosen by SRSWR in n-draws is unbiased for the population mean $\overline{Y}$ and the expansion estimator given by N multiplied by this mean $\bar{y}_d$, say, $\hat{Y}_d = N\bar{y}_d$ has the variance

$$V(\hat{Y}_d) = N^2\left[\frac{1}{N}\sum_{i=1}^N\left(\frac{j}{N}\right)^{n-1} - \frac{1}{N}\right]S^2$$

writing $S^2 = \frac{1}{(N-1)}\sum_{i=1}^N(y_i - \overline{Y})^2$ and so

$$\mathscr{E}(V(\hat{Y}_d)) = N^2\left[\frac{1}{N}\sum_1^N\left(\frac{j}{N}\right)^{n-1} - \frac{1}{N}\right](\tau^2 + \beta^2 S_{xx}) = (srswrd)$$

writing $S_{xx} = \frac{1}{N-1}\sum_1^N (x_i - \overline{X})^2$.

3.7. Strategy 7: (Rao-Hartley-Cochran Sampling, Rao-Hartley-Cochran Estimator)

A sample of size n by Rao-Hartley-Cochran (RHC(1962)) scheme is taken by choosing from the population first a sample of N_1 units by SRSWOR, then a sample of size N_2 from the remaining $(N - N_1)$ units of the population and successively and similarly, finally an SRSWOR of size N_n keeping $N_1 + N_2 + \ldots + N_n = N$ and for the sake of efficiency taking $N_i = \left[\frac{N}{n}\right]$ for $i = 1, 2, \ldots, k$ and the last (n-k) of these N_iFLs as $\left[\frac{N}{n}\right] + 1$ with the restriction $N_1 + N_2 + \ldots + N_n = N$. Such a choice is uniquely possible. For the parts of the population so constructed, the values of p_i are noted and

$$Q_i = p_{i1} + \ldots + p_{iN_i}$$

for the i-th pair or group is noted.

Then, writing $\sum_n$ as the sum over these n pairs or groups and $\sum_n \sum_n$ as the sum over the distinct pairs of these groups follows the RHC's unbiased estimator for Y as

$$t_{RHC} = \sum_n \frac{y_{ij}}{p_{ij}} Q_i$$

on taking independently across these n groups just one unit say labelled ij from the i-th group denoting the associated y value as y_{ij}. Then, it follows that

$$V(t_{RHC}) = \frac{\sum_n N_i^2 - N}{N(N-1)} \sum_n \sum_n p_i p_j \left(\frac{y_i}{p_i} - \frac{y_j}{p_j}\right)^2.$$

Also,

$$\mathscr{E}(V(t_{RHC})) = \frac{\left(\sum_n N_i^2 - N\right)}{N-1} \tau^2 \left(\overline{X} \sum_1^N \frac{1}{x_i} - 1\right) = (rhc).$$

3.8. Strategy 8: (An Inclusion Probability Proportional to size (IPPS or πPS) sampling, Horvitz-Thompson Estimator

The Horvitz Thompson estimator based on a sample s of size (of distinct unit) n is $t_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i}$ and $\pi_i = np_i$, $i \in Population$

$$V(t_{HT}) = \sum \sum_{i<j} (\pi_i \pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

and

$$\mathscr{E}(V(t_{HT})) = \frac{N\tau^2}{n} \left(\overline{X} \sum_1^N \frac{1}{x_i} - n\right) = (ippsht).$$

3.9. Strategy 9: Lahiri-Midzuno-Sen (LMS) sampling Scheme, Horvitz-Thompson Estimator (HTE)

For the sample s of size n for the LMS scheme, the HT estimator is $t_{HT} = \sum_{i\in s} \frac{y_i}{\pi_i}$ with the variance

$$V(t_{HT}) = \sum\sum_{i<j} (\pi_i\pi_j - \pi_{ij}) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$$

with

$$\pi_i = \frac{(N-n)}{(N-1)} p_i + \frac{(n-1)}{(N-1)}$$

and

$$\pi_{ij} = \frac{(N-n)(n-1)}{(N-1)(N-2)}(p_i + p_j) + \frac{(n-1)(n-2)}{(N-1)(N-2).}$$

It follows that

$$\mathscr{E}(V(t_{HT})_{LMS}) \;=\; \beta^2 \left[\sum_1^N \frac{x_i^2}{\pi_i} + \sum\sum_{i\neq j} x_i x_j \frac{\pi_{ij}}{\pi_i\pi_j} - X^2\right] \;+\; \tau^2 \left(\sum_1^N \frac{1}{\pi_i} - N\right) \;=\; (lmsht).$$

4. A numerical study by Simulation

4.1. Simple Model yielding x,y values

Model: Let $y_i = \beta x_i + \varepsilon_i \qquad i \in U = (1,2\ldots,N)$ with β an arbitrarily chosen positive constant; x_iFLs are independently generated from distribution function

$$F(x) = 1 - e^{-\frac{1}{10}x}, \qquad x > 0.$$

The choice of x was from such a distribution mainly because we wanted to use positive values of explanatory variables keeping in mind the application of such a model in real life. The mean of 10 was taken to choose values with considerably moderate values.

ε_iFLs are independently randomly generated from the Normal distribution $N(0,1)$ for $i = 1,2,\ldots,N$.

Also, we take $\beta = 2.3$, 1.6, and 3.6 and $N = 23$. Using these different values of β we generated three sets of values which shall be treated as population. The generated values with $\beta = 2.3$ are reported in Table 1.

Table 1: Table of Population values

Sl.No	x	y	Sl.No	x	y
1	7.55	18.15	13	12.38	28.41
2	11.82	27.25	14	44.24	100.37
3	1.46	1.36	15	10.55	23.84
4	1.40	3.84	16	10.35	23.42
5	4.36	9.97	17	18.76	43.09
6	28.94	66.42	18	6.55	16.16
7	12.30	26.81	19	3.37	8.51
8	5.40	11.93	20	5.88	13.37
9	9.57	22.41	21	23.65	54.13
10	1.47	4.74	22	6.42	15.46
11	13.91	31.88	23	2.94	7.32
12	7.62	17.91			

Throughout this paper we shall use only these three sets of (x,y)-values whenever needed for illustrations as the finite population values of x and y. For the material presented in sections 2.1-2.4 we intend to use the values generated as given above to illustrate the realized magnitudes of estimated gains in efficiencies of pairs of competing strategies. For this, from the population of N=23 sets of (x,y)-values samples of size n=7 are chosen by appropriately defined procedures. The findings are presented in Section 4.2 below in specified tables.

In order to present numerical illustrations for the materials covered in Sections 3.1 through 3.9 we use the x-values of three generated populations mentioned above for the population of size N=23 but β values are differently taken and ε_i's are supposed to have a constant model variance τ^2 which are variably taken for illustration. On every occasion a sample of size n is illustrated with the value 7 but y-values are accordingly supposed to be generated yielding the specified model-expected variances illustrated in tables in Section 4.3 below.

4.2. Numerical study of material in Sections 2.1-2.4

For the purpose of presentation of materials covered in Sections 2.1-2.4, we have chosen 10 separate and independent samples each of size 7 from different populations generated as mentioned above. We worked on deriving the estimated variances and tabulated them below side by side.

Table 2: Estimated variances given in Sections 2.1-2.4 for samples of size n=7 from the (x,y)

$v(t_D)$	v_1	$v(t^*_{SD})$	v_2	$v(t_R)$	v_3	$v(\hat{Y}_{HR})$	v_4
i	ii	iii	iv	v	vi	vii	viii
1806.07	4908.88	1718.08	2418.30	325.73	15196.67	5219.82	25418.88
1.76	3291.07	1.31	395.99	31.21	7845.60	24429.38	18072.16
7.57	12883.59	7.54	1097.78	509.82	2496.31	10429.85	45723.08
3.58	19646.11	2.72	1854.53	35.15	3269.31	5805.09	6658,41
39.77	6404.84	39.74	2163.54	6.49	411.98	22844.20	4321.82
315.56	4776.37	315.49	4321.30	39.93	2237.48	11985.97	5815.61
9.85	6366.44	9.18	3045.95	18.25	1476.29	860.27	2640.17
324.63	3685.89	324.49	5052.97	117.28	4474.86	4470.89	7616.78
3152.35	1301.71	3147.20	8209.40	8.76	344.86	35946.81	6572.70
15.29	5249.95	11.14	2583.30	239.06	21144.43	2014.46	42327.04

Comments

From the values of the estimated variances we may say that (i) (PPSWOR, t_D) is substantially more gainful in efficiency over (ii) (SRSWOR, Expansion Estimator) both of which are much inferior to (iii) (PPSWOR, Symmetrized Des Raj Estimator). For the sample size n=7 we had to obtain $7! = 5040$ Des Raj estimates. But with powerful statistical software it did not cost us much time.

Compared to (vi) (SRSWOR,$N\bar{y}$) the strategy (v) (LMS, Ratio Estimator) is enormously more gainful as it should be because a size variable is employed. Compared to (viii) (SRSWOR,$N\bar{y}$), (vii) (SRSWOR, Hartley-Ross Estimator) is also more gainful, but presumably because an auxiliary size-measure is employed.

4.3. Numerical study of material in Sections 3.1-3.9

Using the values of x_i from three different populations and variously choosing β and τ^2 explained in Section 4.1, we present below in Table 3, 4 and 5 the values of the model expected variances of various unbiased estimators for a finite population total of a variable y of interest based on samples taken according to various schemes.

Table 3: Ten values for each of the model-expected variances for first population

	SRSWR	*SRSWOR*	*PPSWR*	*PPSWR*	*SRSWR*	SRSWR	*RHC*	*IPPS*	*LMS*
	$N\bar{y}$	$N\bar{y}$	t_{HH}	*HTE*	*HTE*	$N\bar{y}_d$	*RHC*	*HTE*	*HTE*
(β,τ^2)	$(srswr)$	$(srswor)$	$(ppswr)$	$(ppswrht)$	$(srswrht)$	$(srswrd)$	(rhc)	$(ippsht)$	$(lmsht)$
(0.1,5)	434.55	316.03	815.49	1051.10	448.74	388.17	609.19	716.92	298.45
(0.1,10)	795.98	578.89	1630.98	2073.27	805.91	711.03	1218.4	1433.83	564.18
(0.2,2)	437.05	317.85	326.20	524.57	509.11	390.41	243.68	286.77	237.13
(0.2,10)	1015.33	738.42	1630.98	2160.04	1080.6	906.97	1218.4	1433.83	662.31
(0.5,5)	2189.4	1592.29	815.49	1745.31	2646.19	1955.74	609.19	716.92	1083.45
(1.5,2)	16596.34	12070.07	326.16	6917.12	20744.01	14825.12	243.68	286.77	7465.73
(2.5,5)	46060.8	33498.76	815.49	19100.64	57582.57	41145.02	609.19	716.92	20708.61
(2.5,10)	46422.23	33671.62	1630.98	20122.81	57939.74	41467.87	1218.34	1433.83	20974.35
(2.5,25)	47506.51	34550.19	4077.44	23189.33	59011.27	42436.44	3046	3584.59	21771.56
(3,2)	65951.66	47964.85	326.19	26441.87	82547.43	58913.06	243.68	286.767	29544.03

Table 4: Ten values for each of the model-expected variances for second population

	SRSWR	*SRSWOR*	*PPSWR*	*PPSWR*	*SRSWR*	SRSWR	*RHC*	*IPPS*	*LMS*
	$N\bar{y}$	$N\bar{y}$	t_{HH}	*HTE*	*HTE*	$N\bar{y}_d$	*RHC*	*HTE*	*HTE*
(β,τ^2)	$(srswr)$	$(srswor)$	$(ppswr)$	$(ppswrht)$	$(srswrht)$	$(srswrd)$	(rhc)	$(ippsht)$	$(lmsht)$
(0.1,5)	421.76	306.74	827.45	1045.82	436.46	376.75	618.14	728.88	298.43
(0.1,10)	783.19	569.59	1654.91	2063.89	793.64	699.61	1236.28	1457.76	557.62
(0.2,2)	385.92	280.67	330.98	518.18	460.03	344.73	247.26	291.55	215.02
(0.2,10)	964.20	701.24	1654.91	2147.11	1031.51	861.30	1236.27	1457.76	639.33
(0.5,5)	1869.84	1359.89	827.45	1711.52	2339.40	1670.29	618.14	728.88	946.09
(1.5,2)	13720.30	9978.4	330.98	6648.23	17982.88	12256.02	247.26	291.55	6234.14
(2.5,5)	38071.79	27688.57	827.45	18354.19	49912.75	34008.62	618.14	728.88	17287.6
(2.5,10)	38433.21	27951.43	1654.91	19372.27	50269.93	34331.48	1236.27	1457.76	17552.79
(2.5,25)	39517.50	28740	4137.27	22426.51	51341.45	35300.05	3090.69	3644.41	18348.37
(3,2)	54447.49	39598.17	330.98	25371.23	71502.90	48636.65	247.26	291.55	24618.34

Table 5: Ten values for each of the model-expected variances for third population

	SRSWR	*SRSWOR*	*PPSWR*	*PPSWR*	*SRSWR*	SRSWR	*RHC*	*IPPS*	*LMS*
	$N\bar{y}$	$N\bar{y}$	t_{HH}	*HTE*	*HTE*	$N\bar{y}_d$	*RHC*	*HTE*	*HTE*
(β,τ^2)	$(srswr)$	$(srswor)$	$(ppswr)$	$(ppswrht)$	$(srswrht)$	$(srswrd)$	(rhc)	$(ippsht)$	$(lmsht)$
(0.1,5)	433.82	315.50	2637.29	3664.84	441.64	387.52	1970.15	2538.72	298.03
(0.1,10)	795.25	578.36	5274.58	7307.63	798.82	710.38	3940.30	5077.44	565.04
(0.2,2)	434.14	315.74	1054.91	1545.29	480.75	387.81	788.06	1015.49	230.89
(0.2,10)	1012.43	736.31	5274.59	7373.77	1052.24	904.38	3940.31	5077.44	658.11
(0.5,5)	2171.27	1579.11	2637.29	4193.92	2468.95	1939.54	1970.15	2538.72	1042.59
(1.5,2)	16433.17	11951.39	1054.92	6417.29	19148.91	14679.36	788.06	1015.49	7087.09
(2.5,5)	45607.53	33169.11	2637.29	17421.07	53151.72	40740.12	1970.15	2538.72	19656.68
(2.5,10)	45968.96	33431.97	5274.58	21063.87	53508.89	41062.98	3940.30	5077.44	19923.69
(2.5,25)	47053.24	34220.54	13186.47	31992.25	54580.42	42031.55	9850.76	12693.61	20724.72
(3,2)	65298.96	47490.15	1054.92	21297.84	76167.01	58330.01	788.06	1015.49	28027.93

Comments

Among the strategies (srswr) (SRSWR, $N\bar{y}$), (srswor) (SRSWOR, $N\bar{y}$) and (srswrd) (SRSWR,$N\bar{y}_d$), as anticipated (srswor) fares best and (srswrd) in between the other two. Moreover (srswrht) (SRSWR, t_{HT}) fares worse than all these three. Between the (ppswr) (PPSWR, t_{HH}) and the (rhc) RHC strategy, the latter performs better for every (β, τ^2) pair as it should. Interestingly, (ppswrht) (PPSWR, t_{HT}) fares worse than both.

(ippsht) (IPPS, t_{HT}) strategy fares competitively against (lmsht) (LMS, t_{HT}), the latter poorer as τ^2 is taken higher. Interestingly, they are found competitive against all four strategies with equal probability sampling making no use of size measures.

Most interestingly, the (rhc) RHC strategy fares by far the best among all the strategies under our competition here for almost all choices of our (β, τ^2)'s.

5. Discussions

The present work has clearly two distinct aspects. One of them is extending the well-known approach of comparing the classical stratified sampling strategy with the corresponding unstratified one in terms of the two variance estimates from the stratified sample at hand on identifying the Des Raj estimator combined with PPSWOR, the symmetrized Des Raj estimator with PPSWOR, the Hartley-Ross estimator based on SRSWOR and the mean of values of distinct sample units in SRSWR versus the over-all sample mean in SRSWR and the expansion estimator in SRSWOR as the situations when easy variance estimator formulae are easy to derive. The other being on observing that the first aspect can be impressively clarified through simulated illustrations, following it through an appeal to simulated illustrations also to compare model-based expectations of exact design variances of pairs of well-known unbiased estimators of population totals citing several complex and simple sampling strategies.

6. Conclusions

(i) In the first case in Section 5, as expected, the complex strategies numerically outperform the respective simpler ones. Thus, it is vindicated that to go for the complex alternatives is lucrative rather than to remain complacent about the simpler alternatives.
(ii) In the second case of Section 5, for various alternative pairs relative performances of model-expected variances are comparatively demonstrated in Section 4.3. The Rao-Hartley-Cochran (1962) strategy for our illustrated numerical situation is demonstrated to fare as the most effective strategy. But from this it cannot be claimed that one should always go for this in practice. Respective performances are well illustrated in Section 4.3 of course. We cannot make general conclusions beyond our illustrated example of course.

Acknowledgement

The authors gratefully acknowledge two referee's recommendations that led to this improved version over an earlier one.

References

Bayless, D. L. and Rao, J. N. K., (1970). *An empirical study of stabilities of estimators and variance estimators in unequal probability sampling (n=3or 4)*, Jour. Amer. Stat.Assoc. 65, pp. 1645–1667.

Chaudhuri, A., (2010). *Essentials of survey sampling*, Prentice Hall of India, New Delhi.

Chaudhuri, A. and Arnab, R., (1979). *On the relative efficiencies of sampling strategies under a super-population model*, Sankhya, Ser. C. 41, pp. 40–43.

Cochran, W., G., (1977). *Sampling Techniques.* John Wiley and Sons. New York.

Des Raj, (1956). *Some estimators in sampling with varying probabilities without replacement*, Jour. Amer. Stat. Assocn. 51, pp. 269–284.

Hansen, M. H. and Hurwitz, W. N., (1943). *On the theory of sampling from finite populations*, Ann. Math. Stat, 14, pp. 333–362.

Hartley, H. O. and Ross, A., (1954). *Unbiased ratio estimators*, Nature, 174, pp. 270–271.

Horvitz, D. G. and Thompson, D. J., (1952). *A generalization of sampling without replacement from a finite universe*, Jour. Amer. Stat. Assoc., 47, 663–685.

Lahiri, D. B., (1951). *A method of sample selection providing unbiased ratio estimates*, Bull. Int. Stat. Inst. 33(2), pp. 133–140.

Midzuno, H., (1952). *An Outline of the theory of sampling systems*, Annals. Inst. Stat. Math, 1, pp. 149–156.

Murthy, M. N., (1957). *Ordered and unordered estimators in sampling without replacement*, Sankhy$\bar{a}$, 18, pp. 379–390.

Rao, J. N. K., (1961). *On the estimate of variance in unequal probability sampling*, Annals. Inst. Stat. Math, 13, pp. 57–60.

Rao, J.N.K. and Bayless, D. L., (1969). *An empirical study of the stabilities of estimators and variance estimators in unequal probability of two units per stratum*, Jour. Amer. Stat. Assoc. 64, pp. 540–549.

Rao, J. N. K., Hartley, H. O., Cochran, N.G., (1962). *On a simple procedure of unequal probability sampling without replacement*, Jour. Roy. Stat. Soc. B. 24, pp. 482–491.

Rao, T. J., (1967). *On the choice of a strategy for a ratio method of estimation*, Jour. Roy.Stat.Soc. B. 29, pp. 392–397.

Roychoudhury, D. K., (1957). *Unbiased sampling design using information provided by linear function of auxiliary variate*, Chapter 5, thesis for Associateship of Indian Statistical Institute, Kolkata.

Sarndal. C. E., Swensson, B and Wretman, J., (1992). *Model Assisted Survey Sampling*, Springer Verlag, Heidelberg.

Sen, A. R., (1953). *On the estimator of the variance in sampling with varying probabilities*, J. Ind. Soc. Agri. Stat. 5(2), pp. 119–127.

Smith, H. F., (1938). *An empirical law describing heterogeneity in the yields of agricultural crops*, Jour. Agri. Sci. 28, pp. 1–23.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 55–74, DOI 10.21307/stattrans-2022-004
Received – 24.07.2020; accepted – 26.08.2021

Generalized extended Marshall-Olkin family of lifetime distributions

Mehdi Goldoust[1], **Adel Mohammadpour**[2]

ABSTRACT

We introduce a new generalized family of nonnegative continuous distributions by adding two extra parameters to a lifetime distribution, called the baseline distribution, by twice compounding a power series distribution. The new family, called the lifetime power series-power series family, has a serial arrangement of parallel structures, which extends the Marshall and Olkin structure. Four special models are discussed. A mathematical treatment of the new distributions is provided, including ordinary and incomplete moments, quantile, moment generating and mean residual functions. The maximum likelihood estimation technique is used to estimate the model parameters and a simulation study is conducted to investigate the performance of the maximum likelihood estimates. Its applicability is also illustrated by means of two real data sets.

Key words: compound distribution, hazard rate function, lifetime distribution, maximum likelihood estimation, power series distribution.

1. Introduction

Classical well-known distributions, such as Weibull, gamma and Lomax distributions, are widely used for modeling data in many disciplines, including engineering, statistics, medical sciences, economics, and insurance. However, in many practical situations, they cannot provide appropriate fits on real data sets. Throughout the two last decades, several generators have been proposed in the literature to extend well-known distributions by adding one or more parameters to the baseline distribution. Since 1997, when Marshall and Olkin proposed a way to add a parameter to the lifetime distribution, by compounding with the geometric distribution, several new families of distributions have been derived by compounding the power series distribution with many other nonnegative continuous distributions to provide more flexible distributions for modeling lifetime data.

Marshall and Olkin's (1997) method was based on the lifetime of a series or parallel system with an unknown amount of components. Their work was extended by Chahkandi and Ganjali (2009), which proposed the exponential power series (EPS) distribution. Furthermore, Morais and Barreto-Souza (2011) proposed the Weibull power series (WPS) distribution containing the EPS distribution as a particular case. On the other hand, Flores et al. (2013) introduced the complementary EPS distribution, complementary to the EPS

[1]Department of Mathematics, Behbahan Branch, Islamic Azad University, Behbahan, Iran.
E-mail: mehdigoldust@gmail.com. ORCID: https://orcid.org/0000-0002-5859-3350.
[2]Department of Statistics, Faculty of Mathematics & Computer Science, Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran. E-mail: adel@aut.ac.ir. ORCID: https://orcid.org/0000-0002-5079-7025.

distribution and Munteanu et al. (2014) presented complementary WPS distribution. Some more well-known generators based on Marshal-Olkin generated family (MO-G) are Kumaraswamy Marshall-Olkin by Alizadeh et al. (2015), Beta Marshall-Olkin by Alizadeh et al. (2015), exponentiated logarithmic Marshal-Olkin by Marinhoa and Cordeiro (2016), and Marshall-Olkin alpha power family by Nassar et al. (2019).

According to Ross (2010), any system can be represented both as a series arrangement of parallel structures or as a parallel arrangement of series structures. Using this key, the purpose of this paper is to introduce a new generator of lifetime distributions by compounding a lifetime distribution with twice power series distribution, obtaining what is referred to as the LPS^2 family of distributions. The proposed family is motivated by a system consisting of serial components with each component consisting of a parallel of components. Some researchers published real examples of systems made by the serial, and parallel components are resonant converters (Kazimierczuk et al., 1993), hybrid electric bus (Xiong et al., 2009), and hybrid envelope amplifiers (Hassan et al., 2012).

The paper is organized as follows. In Section 2, we introduce the new family of distributions. Four special cases of this family are defined in Section 3. Section 4 derives some of its mathematical properties. The explicit expressions for the moments, incomplete moments, generating function, and mean residual time are given in this section. The estimation of parameters using the maximum likelihood method is investigated in Section 5. In Section 6, a simulation study is performed to show the behaviour of asymptotic biases and mean square errors of maximum likelihood estimations (MLEs). Illustrative examples of two real data sets are given in Section 7. Finally, in Section 8, we present some concluding remarks.

2. The LPS^2 family of distributions

Let $X_{i,j}$s be a sequence of independent and identically (iid) random sample from a baseline lifetime distribution for $j = 1,2,\ldots,Z_i$, $i = 1,2,\ldots,U$, with probability density function (pdf) $\pi(x;\boldsymbol{\varsigma})$ and cumulative distribution function (cdf) $\Pi(x;\boldsymbol{\varsigma})$, where $\boldsymbol{\varsigma}$ denoted the parameter vector of baseline distribution. Suppose $Z_1, Z_2, \ldots, Z_U$ are iid zero truncated power series random variables with probability mass function (pmf)

$$P(z;\theta) = \frac{b_z \theta^z}{B(\theta)},$$

for $z = 1,2,\ldots$, where b_z depends only on z and $B(\theta) = \sum_{z=1}^{\infty} b_z \theta^z < \infty$. Furthermore, suppose U is a zero truncated power series random variable with pmf

$$P(u;\lambda) = \frac{a_u \lambda^u}{A(\lambda)},$$

for $u = 1,2,\ldots$, where a_u depends only on u and $A(\lambda) = \sum_{u=1}^{\infty} a_u \lambda^u < \infty$. Consider that $X_{i,j}$s, Z_is and U are independent, we define a system, which is made of U series components, that the ith component is made of Z_i components working in parallel. Figure 1 shows an illustration of this system. Then the lifetime of the system is

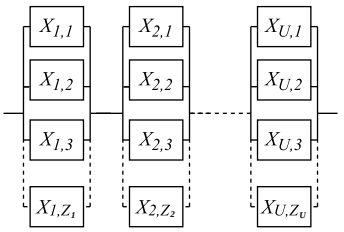


Figure 1: The system made up of series and parallel components.

$$X = \min\left\{\max\left\{X_{i,j}\right\}_{j=1}^{Z_i}\right\}_{i=1}^{U}.$$

Table 1 shows useful quantities of some members of the power series family (truncated at zero) such as the Poisson, geometric, logarithmic series, negative binomial and binomial distributions.

Table 1: Members of the power series family.

power series family	pmf	λ	extended λ after compounding	a_u	$A(\lambda)$	$I(\lambda)=\int_0^\lambda A'(v)\log\{A'(v)\}dv$
Poisson	$e^{-\lambda}\lambda^u/u!\left(1-e^{-\lambda}\right)$	$0<\lambda<\infty$	$\lambda\in(-\infty,0)\bigcup(0,+\infty)$	$1/u!$	$e^{\lambda}-1$	$(\lambda-1)e^{\lambda}+1$
Geometric	$(1-\lambda)\lambda^{u-1}$	$0<\lambda<1$	$\lambda\in(-\infty,0)\bigcup(0,1)$	1	$\lambda/(1-\lambda)$	$-2(\lambda+\log\{1-\lambda\})/(1-\lambda)$
Logarithmic series	$-\lambda^u/u\log\{1-\lambda\}$	$0<\lambda<1$	$\lambda\in(-\infty,0)\bigcup(0,1)$	$1/u$	$-\log\{1-\lambda\}$	$-\frac{1}{2}\log^2\{1-\lambda\}$
Negative Binomial	$\binom{m+u-1}{u}(1-\lambda)^m\lambda^u/1-(1-\lambda)^m$	$0<\lambda<1$	$\lambda\in(-\infty,0)\bigcup(0,1)$	$\binom{m+u-1}{u}$	$(1-\lambda)^{-m}-1$	$\log\{m\}A'(\lambda)-\frac{m+1}{m}\{(1-\lambda)^{-m}[m\log\{1-\lambda\}+1]-1\}$
Binomial	$\binom{m}{u}\lambda^u/((1+\lambda)^m-1)$	$0<\lambda<\infty$	$\lambda\in(-1,0)\bigcup(0,+\infty)$	$\binom{m}{u}$	$(1+\lambda)^m-1$	$(1+\lambda)^m\log\left\{m(1+\lambda)^{m-1}\right\}-\log\{m\}$

It could be shown that the marginal cdf of X is

$$F(x;\boldsymbol{\xi}) = 1-[A(\lambda)]^{-1}A\left(\lambda\left\{1-\frac{B(\theta\Pi(x;\boldsymbol{\varsigma}))}{B(\theta)}\right\}\right), \tag{1}$$

for $x>0$ and $\boldsymbol{\xi}=(\boldsymbol{\varsigma},\theta,\lambda)$. Hence, $S(x;\boldsymbol{\xi})=1-F(x;\boldsymbol{\xi})$ is the corresponding survival function and the pdf and hazard rate function (hrf) of LPS^2 family are defined as follows:

$$f(x;\boldsymbol{\xi}) = \frac{\lambda\theta\pi(x;\boldsymbol{\varsigma})B'(\theta\Pi(x,\boldsymbol{\varsigma}))}{A(\lambda)B(\theta)}A'\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x;\boldsymbol{\varsigma}))\right\}\right) \tag{2}$$

and

$$h(x;\boldsymbol{\xi}) = \frac{\lambda\theta\pi(x,\boldsymbol{\varsigma})B'(\theta\Pi(x;\boldsymbol{\varsigma}))A'\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x;\boldsymbol{\varsigma}))\right\}\right)}{B(\theta)A\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x;\boldsymbol{\varsigma}))\right\}\right)},$$

for $x>0$, respectively. Furthermore, $A'(.)$ and $B'(.)$ are the derivative of $A(.)$ and $B(.)$

functions, respectively. The hrf can be constant, decreasing, increasing, J-shaped, bathtub-shaped, and upside-down bathtub-shaped for a different type of the baseline and power series distributions (see Section 3). The LPS^2 family of distributions contains all compounded lifetime distributions, which were built by the Marshall and Olkin method. Here, some sub-models of the LPS^2 family are presented.

- When $Z_1, Z_2, \ldots = 1$ and the baseline is an exponential or a Weibull distribution, we obtain the EPS (Chahkandi and Ganjali, 2009) and WPS (Morais and Barreto-Souza, 2011) distributions respectively;
- when $U = 1$ and the baseline is an exponential or a Weibull distribution, we obtain the CEPS (Flores et al., 2013) and the max-Weibull power series (Munteanu et al., 2014) distributions respectively.

Furthermore, since

$$\lim_{\lambda \to 0} A(\lambda) = 0 \quad \text{and} \quad \lim_{\lambda \to 0} \frac{A(\lambda x)}{A(\lambda)} = x,$$

thereupon, we have

- The lifetime power series distributions with minimum structure is a special limiting case of the LPS^2 family of distributions when $\theta \to 0^+$. In general,

$$\begin{aligned} \lim_{\theta \to 0^+} F(x;\boldsymbol{\xi}) &= 1 - [A(\lambda)]^{-1} \lim_{\theta \to 0} A\left(\lambda \left\{1 - \frac{B(\theta \Pi(x;\boldsymbol{\varsigma}))}{B(\theta)}\right\}\right) \\ &= 1 - [A(\lambda)]^{-1} A(\lambda \{1 - \Pi(x;\boldsymbol{\varsigma})\}); \end{aligned}$$

- The complementary lifetime power series distributions with maximum structure is a special limiting case of the LPS^2 family of distributions when $\lambda \to 0^+$. In general,

$$\lim_{\lambda \to 0^+} F(x;\boldsymbol{\xi}) = 1 - \lim_{\lambda \to 0} \frac{A\left(\lambda \left\{1 - \frac{B(\theta \Pi(x;\boldsymbol{\varsigma}))}{B(\theta)}\right\}\right)}{[A(\lambda)]^{-1}} = [B(\theta)]^{-1} B(\theta \Pi(x;\boldsymbol{\varsigma}));$$

- The baseline distribution is a special limiting case of this new family when $\theta, \lambda \to 0^+$

$$\lim_{\theta \to 0^+} \lim_{\lambda \to 0^+} F(x;\boldsymbol{\xi}) = \Pi(x;\boldsymbol{\varsigma}).$$

3. Some special models

In this section, we consider some special cases of the LPS^2 distribution. These special models generalize some well-known distributions in the literature. We provide four special models of this family corresponding to the baseline exponential, Weibull, Lomax (Lx), and generalized half-normal (GHN) distributions. To illustrate the flexibility of the distributions, graphs of the pdf and hrf for some selected distributions are presented.

It should be noted that compounding of a lifetime geometric family (Marshall and Olkin, 1997) with a geometric distribution again, just expand the parameter space and its cdf doesn't change. Suppose $X_{i,j}$ is a sequence of the independent identically lifetime random variables with cdf $F(x;\boldsymbol{\varsigma})$. The cdf of a lifetime geometric-geometric family of distributions (compounding a lifetime and twice geometric distribution) with parameter $\boldsymbol{\xi} = (\boldsymbol{\varsigma}, \theta, \lambda)$ is

$$F(x;\boldsymbol{\xi}) = \frac{(1-\lambda)F(x;\boldsymbol{\varsigma})}{1-\theta+(\theta-\lambda)F(x;\boldsymbol{\varsigma})},$$

for $x > 0$. With a reparameterization $\gamma = \frac{\theta-\lambda}{1-\lambda}$, we can write

$$F(x;\boldsymbol{\varsigma},\gamma) = \frac{F(x;\boldsymbol{\varsigma})}{1-\gamma F(x;\boldsymbol{\varsigma})},$$

for $x > 0$ and $\gamma < 1$. The lifetime geometric-geometric family of distributions (LGG) is due to Marshall and Olkin (1997) with expanded geometric parameter space. On the other hand, the parameter space of truncated Poisson distribution in compound distributions could be extended to $(-\infty,0)\bigcup(0,+\infty)$, and the parameter space of truncated binomial distribution could be extended to $(-1,0)\bigcup(0,+\infty)$. A more similar extension of the parameter space may be done to power series parameters (see Table 1).

3.1. Exponential power series-power series distribution (EPS^{2}D)

The EPS^{2}D distribution is defined from (1) by taking $\Pi(x;\beta) = 1-e^{-\beta x}$. Then, its density function is given by

$$f(x) = \frac{\beta\theta\lambda e^{-\beta x}B'\left(\theta\left[1-e^{-\beta x}\right]\right)}{A(\lambda)B(\theta)}A'\left(\lambda\left\{1-[B(\theta)]^{-1}B\left(\theta\left[1-e^{-\beta x}\right]\right)\right\}\right),$$

for $x > 0$ and $\beta > 0$.

3.2. Weibull power series-power series distribution (WPS^{2}D)

The cdf and pdf of the Weibull distribution with scale parameter β and shape parameter α are given by $\Pi(x;\alpha,\beta) = 1-e^{-\beta x^{\alpha}}$ and $\pi(x;\alpha,\beta) = \alpha\beta x^{\alpha-1}e^{-\beta x^{\alpha}}$, respectively. The WPS^{2}D pdf follows by inserting these expressions in (2) as

$$f(x) = \frac{\alpha\beta\theta\lambda x^{\alpha-1}e^{-\beta x^{\alpha}}B'\left(\theta\left[1-e^{-\beta x^{\alpha}}\right]\right)}{A(\lambda)B(\theta)}A'\left(\lambda\left\{1-[B(\theta)]^{-1}B\left(\theta\left[1-e^{-\beta x^{\alpha}}\right]\right)\right\}\right),$$

for $x > 0$, $\alpha > 0$ and $\beta > 0$. Figures 2 and 3 display the pdf and hrf of the Weibull geometric-Poisson distribution (WGPD) for some selected parameter values.

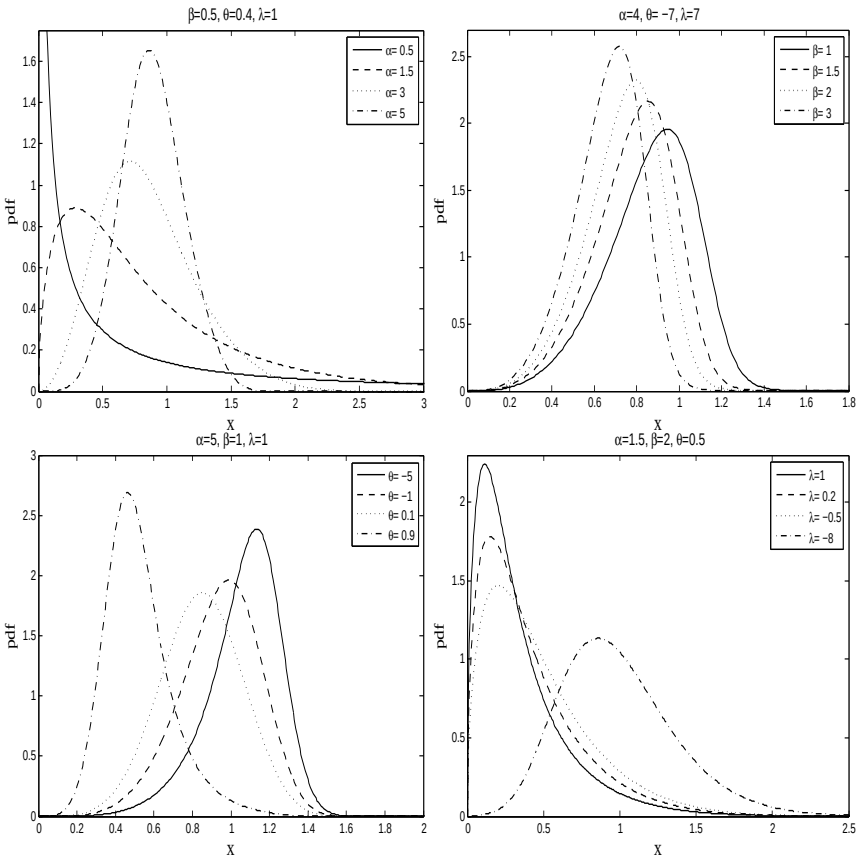


Figure 2: Graphs of the WGPD pdf for some selected values of the parameters.

3.3. Lomax power series-power series distribution (LxPS^{2}D)

The LxPS^{2}D distribution is defined from (2) by taking $\Pi(x;\beta) = 1-[1+\beta x]^{-\alpha}$ for the cdf of the Lomax distribution with parameters α and β. The LxPS^2 pdf is given by

$$f(x) = \frac{\alpha\beta\theta\lambda B'\left(\theta\left\{1-[1+\beta x]^{-\alpha}\right\}\right)}{A(\lambda)B(\theta)[1+\beta x]^{\alpha+1}} A'\left(\lambda\left\{1-[B(\theta)]^{-1}B\left(\theta\left\{1-[1+\beta x]^{-\alpha}\right\}\right)\right\}\right),$$

for $x>0$, $\alpha>0$ and $\beta>0$.

3.4. Generalized half-normal power series-power series distribution (GHNPS^{2}D)

Cooray and Ananda (2008) introduced generalized half-normal distribution with cdf and pdf

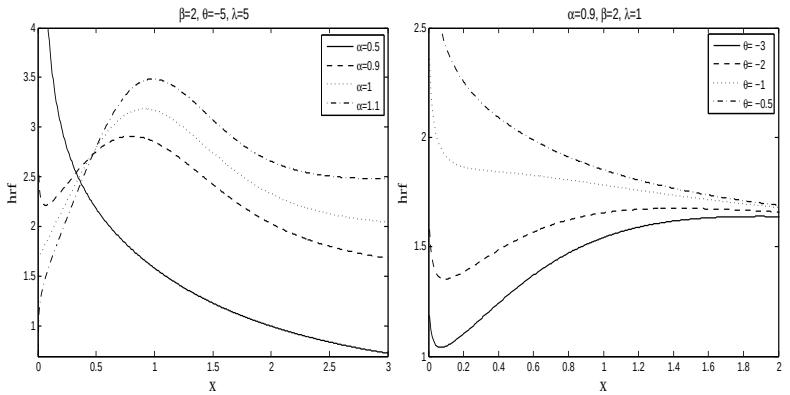


Figure 3: Graphs of the WGPD hrf for some selected values of the parameters.

$$\Pi(x) = 2\Phi(\beta x^{\alpha}) - 1 \text{ and } \pi(x) = \sqrt{\frac{2}{\pi}}\alpha\beta x^{\alpha-1} e^{-\frac{1}{2}(\beta x^{\alpha})^2},$$

respectively. The GHNPS^{2}D pdf follows by inserting these expressions in (2) as

$$f(x) = \frac{\sqrt{\frac{2}{\pi}}\alpha\beta\theta\lambda x^{\alpha-1} B'(\theta\,[2\Phi(\beta x^{\alpha}) - 1\,])}{A(\lambda)\,B(\theta)\,e^{\frac{1}{2}(\beta x^{\alpha})^2}} A'\left(\lambda\left\{1 - [B(\theta)]^{-1} B(\theta\,[2\Phi(\beta x^{\alpha}) - 1\,])\right\}\right),$$

for $x > 0$, $\alpha, \beta > 0$ and $\Phi(.)$ denotes the cdf of standard normal distribution.

4. Some useful properties

In this section, we derive some useful structural properties of the LPS^2 distributions. These include the two useful linear representations for (1) and (2) (Section 4.1), the r-th moment, moment generating function and mean residual lifetime (Section 4.2), the quantiles (Section 4.3).

4.1. Two useful linear representations

Let X be a LPS^2 random variable with parameters $\boldsymbol{\xi} = (\boldsymbol{\varsigma}, \theta, \lambda)$. Using the binomial expansion and $A'(\lambda) = \sum_{u=1}^{\infty} u a_u \lambda^{u-1}$, the cdf and pdf of X can be expanded as

$$F(x;\boldsymbol{\xi}) = \sum_{k=1}^{\infty}\sum_{j=0}^{\infty} \phi_{k,j}\Pi(x;\boldsymbol{\varsigma})^{k+j} \tag{3}$$

and

$$f(x;\boldsymbol{\xi}) = \pi(x,\boldsymbol{\varsigma}) \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \varphi_{k,j} \Pi(x;\boldsymbol{\varsigma})^{k+j-1}, \tag{4}$$

for $x > 0$, where $\phi_{k,j} = \phi_{k,j}(\theta,\lambda)$ and $\varphi_{k,j} = \varphi_{k,j}(\theta,\lambda) = (k+j)\phi_{k,j}$. For further details, see Appendix.

4.2. Moment properties

First, we derive the r-th moment for a random variable X. Therefore, the r-th moment of $X \sim \mathrm{LPS}^2(\boldsymbol{\varsigma},\theta,\lambda)$ is given by

$$\begin{aligned} \mu'_r &= E[X^r] = \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \varphi_{k,j} \int_0^{\infty} x^r \pi(x,\boldsymbol{\varsigma}) \Pi(x)^{k+j-1} dx \\ &= \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \varphi_{k,j} M(r,k+j-1), \end{aligned}$$

for $r > 0$, where $M(s,k+j-1)$ is the $(s,k+j-1)$th probability weighted moment (PWM) of baseline distribution defined by Greenwood et al. (1979) as follows:

$$M(i,j) = E\left[X^i \Pi(X)^j\right] = \int_0^{+\infty} x^i [\Pi(x)]^j d\Pi(x).$$

The moment generating function (mgf) of the LPS^2 family of distributions is given by

$$\begin{aligned} M_X(t) &= E\left[e^{tX}\right] = E\left[\sum_{s=0}^{\infty} \frac{(tX)^s}{s!}\right] = \sum_{s=0}^{\infty} \frac{t^s}{s!} E[X^s] \\ &= \sum_{s,j=0}^{\infty} \sum_{k=1}^{\infty} \frac{\varphi_{k,j}}{s!} M(s,k+j-1) t^s. \end{aligned}$$

Given the survival to time x_0, the residual life is the period from x_0 until the time of failure. From (4), the mean residual lifetime of the LPS^2 distribution is given by

$$\begin{aligned} m(x_0) &= E[X - x_0 | X > x_0] = [S(x_0;\boldsymbol{\xi})]^{-1} \int_{x_0}^{\infty} v f(v) dv - x_0 \\ &= [S(x_0;\boldsymbol{\xi})]^{-1} \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \varphi_{k,j} M_{x_0}(1,k+j-1) - x_0, \end{aligned}$$

where the upper incomplete probability weighted moment was defined as

$$M_{x_0}(i,j) = \int_{x_0}^{\infty} x^i [\Pi(x)]^j d\Pi(x).$$

4.3. Quantiles

If U is a uniform $[0,1]$ random variable then

$$X = \Pi^{-1}\left(\frac{1}{\theta}B^{-1}\left(B(\theta)\left[1-\frac{1}{\lambda}A^{-1}(UA(\lambda))\right]\right)\right)$$

is a LPS^2 random variable, where $\Pi^{-1}(.)$ is the inverse of baseline cdf. Furthermore, $A^{-1}(.)$ and $B^{-1}(.)$ are the inverse of $A(.)$ and $B(.)$ functions, respectively. It follows that the ωth quantile of the LPS^2 distributions is

$$x_\omega = \Pi^{-1}\left(\frac{1}{\theta}B^{-1}\left(B(\theta)\left[1-\frac{1}{\lambda}A^{-1}(\omega A(\lambda))\right]\right)\right).$$

The effects of the parameters on the skewness of random variable X can be shown based on quantiles. The Bowley skewness (Kenney and Keeping, 1962), also known as the quantile skewness coefficient, is defined by

$$B = \frac{x_{0.75}+x_{0.25}-2x_{0.5}}{x_{0.75}-x_{0.25}}.$$

Figure 4 graphs the Bowleie's measure for the WGPD distribution. The graph indicates the variability of this measures on the α, β, θ and λ parameters.

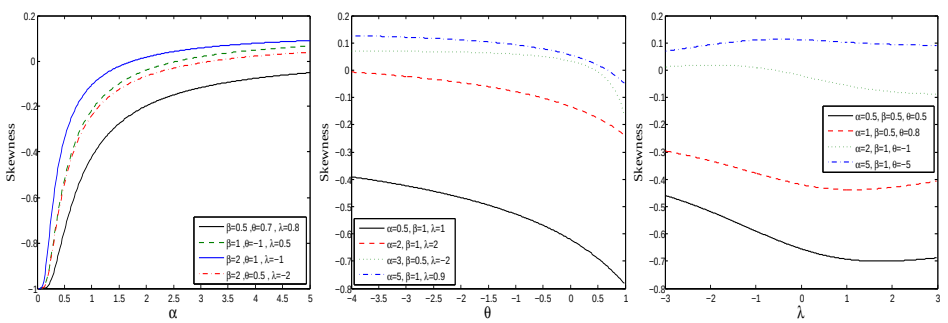


Figure 4: Graphs of skewness based on the quantiles of the WGPD distribution.

5. Estimation of the parameters

In this section, we determine the maximum likelihood estimates (MLEs) of the parameters of the LPS^2 family of distributions from complete samples only. Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample from the LPS^2 distribution with observed values $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and parameters $\boldsymbol{\xi} = (\boldsymbol{\varsigma}, \theta, \lambda)$. The log-likelihood function is given by

$$\ell(\boldsymbol{\xi} \mid \mathbf{x}) = n\log\theta + n\log\lambda - n\log[A(\lambda)] - n\log[B(\theta)] + \sum_{i=1}^{n}\log[\pi(x_i;\boldsymbol{\varsigma})] + \sum_{i=1}^{n}\log\left[A'\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x_i;\boldsymbol{\varsigma}))\right\}\right)\right]. \quad (5)$$

By differentiating (5) with respect to $\boldsymbol{\varsigma}$, θ and λ, and then equating these derivative to zero, we obtain the components of score vector $U_n(\boldsymbol{\xi}) = \left(\frac{\partial\ell}{\partial\boldsymbol{\varsigma}}, \frac{\partial\ell}{\partial\theta}, \frac{\partial\ell}{\partial\lambda}\right)$, where

$$\frac{\partial\ell}{\partial\boldsymbol{\varsigma}} = \sum_{i=1}^{n}\frac{\pi_{\boldsymbol{\varsigma}}(x_i;\boldsymbol{\varsigma})}{\pi(x_i;\boldsymbol{\varsigma})} - \frac{\lambda\theta}{B(\theta)}\sum_{i=1}^{n}\frac{\Pi_{\boldsymbol{\varsigma}}(x_i;\boldsymbol{\varsigma})B'\left(\theta\Pi_{\boldsymbol{\varsigma}}(x_i;\boldsymbol{\varsigma})\right)A''\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x_i;\boldsymbol{\varsigma}))\right\}\right)}{A'\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x_i;\boldsymbol{\varsigma}))\right\}\right)},$$

$$\frac{\partial\ell}{\partial\theta} = \frac{n}{\theta} - \frac{nB'(\theta)}{B(\theta)} - \lambda\sum_{i=1}^{n}\frac{B(\theta)\Pi(x_i;\boldsymbol{\varsigma})B'\left(\theta\bar{G}(x_i;\boldsymbol{\varsigma})\right) - B'(\theta)B(\theta\Pi(x_i;\boldsymbol{\varsigma}))}{[B(\theta)]^2} \times\frac{A''\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x_i;\boldsymbol{\varsigma}))\right\}\right)}{A'\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x_i;\boldsymbol{\varsigma}))\right\}\right)}$$

and

$$\frac{\partial\ell}{\partial\lambda} = \frac{n}{\lambda} - \frac{nA'(\lambda)}{A(\lambda)} + \sum_{i=1}^{n}\frac{\left\{1-[B(\theta)]^{-1}B\left(\theta\bar{G}(x_i;\boldsymbol{\varsigma})\right)\right\}A''\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x_i;\boldsymbol{\varsigma}))\right\}\right)}{A'\left(\lambda\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x_i;\boldsymbol{\varsigma}))\right\}\right)},$$

where

$$\Pi_{\boldsymbol{\varsigma}}(x_i;\boldsymbol{\varsigma}) = \frac{\partial\Pi(x_i;\boldsymbol{\varsigma})}{\partial\boldsymbol{\varsigma}} \text{ and } \pi_{\boldsymbol{\varsigma}}(x_i;\boldsymbol{\varsigma}) = \frac{\partial\pi(x_i;\boldsymbol{\varsigma})}{\partial\boldsymbol{\varsigma}}.$$

The maximum likelihood estimates, $\widehat{\boldsymbol{\xi}}$ of $\boldsymbol{\xi} = (\boldsymbol{\varsigma}, \theta, \lambda)$ are obtained by solving the nonlinear equations $U_n(\boldsymbol{\xi}) = \left(\frac{\partial\ell}{\partial\boldsymbol{\varsigma}}, \frac{\partial\ell}{\partial\theta}, \frac{\partial\ell}{\partial\lambda}\right) = \mathbf{0}$. These equations have no closed form and the values of the parameters $\boldsymbol{\varsigma}$, θ and λ must be found by using iterative methods. To solve these equations, it is usually more convenient to use nonlinear optimization methods such as the Newton-Raphson, quasi-Newton, or Nelder-Mead procedures. The Adequacy Model package version 1.0.8 available in the R programming language was used for numerical maximization in the data examples Section 7. For interval estimation of $(\boldsymbol{\varsigma}, \theta, \lambda)$ and hy-

pothesis tests on these parameters, we obtain the observed information matrix since the expected information matrix is very complicated and requires numerical integration. The $(p+2)\times(p+2)$ observed information matrix $J_n(\boldsymbol{\xi})$, where p is the dimension of the parameter vector $\boldsymbol{\varsigma}$, becomes

$$J_n(\boldsymbol{\xi}) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \boldsymbol{\varsigma}^2} & \frac{\partial^2 \ell}{\partial \boldsymbol{\varsigma} \partial \theta} & \frac{\partial^2 \ell}{\partial \boldsymbol{\varsigma} \partial \lambda} \\ \frac{\partial^2 \ell}{\partial \theta \partial \boldsymbol{\varsigma}} & \frac{\partial^2 \ell}{\partial \theta^2} & \frac{\partial^2 \ell}{\partial \theta \partial \lambda} \\ \frac{\partial^2 \ell}{\partial \lambda \partial \boldsymbol{\varsigma}} & \frac{\partial^2 \ell}{\partial \lambda \partial \theta} & \frac{\partial^2 \ell}{\partial \lambda^2} \end{pmatrix}.$$

Under the usual regularity conditions and that the parameters are in the interior of the parameter space, but not on the boundary and large n, the distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}\right)$ can be approximated by $N_{p+2}\left(\mathbf{0}, nJ_n^{-1}(\boldsymbol{\xi})\right)$. This approximation can be used to construct confidence intervals and tests of hypotheses.

6. A simulation study

In this section, we assess the performance of the MLEs of the WGPD distribution as the particular case of LPS^2 distribution with respect to sample size n. Samples of sizes 20, 50, 100, 200 and 500 are generated for different combinations of $\boldsymbol{\xi} = (\alpha, \beta, \theta, \lambda)$ from WGPD distribution by using (5). We repeated the simulation k =1000 times with parameter values I : $\alpha = 2$, $\beta = 1$, $\theta = 0.5$, $\lambda = 1$ and II : $\alpha = 1.5$, $\beta = 0.5$, $\theta = 0.7$, $\lambda = 0.8$, then the MLEs of the parameters are calculated. The standard deviation (SD) of the parameter estimates are computed by inverting the observed information matrices. The bias and mean squared errors (MSE) are given respectively by

$$\text{bias}_{\varepsilon}(n) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\varepsilon}_i - \varepsilon)$$

and

$$\text{MSE}_{\varepsilon}(n) = \frac{1}{1000}\sum_{i=1}^{1000}(\hat{\varepsilon}_i - \varepsilon)^2,$$

for $\varepsilon = \alpha, \beta, \theta, \lambda$ where $\hat{\varepsilon}_i$ is ith MLE of ε with standard error $s_{\hat{\varepsilon}_i}$. The empirical results are given in Table 2 indicate that the MLEs perform well for estimating the model parameters. According to the results, it can be concluded that as the sample size n increases, the MSEs decay toward zero. We also observe that for all the parameters, the biases decrease as the sample size n increases.

Table 2: The mean, bias, MSE, standard error of the MLE estimators.

		I					*II*				
n	Parameter	R.value	MLE	Bias	MSE	SD	R.value	MLE	Bias	MSE	SD
$n=20$	α	2	2.0912	0.0912	0.1129	0.4445	1.5	1.5412	0.0412	0.0478	0.2715
	β	1	1.1614	0.1614	0.3196	0.7924	0.5	0.5738	0.0738	0.0827	0.4787
	θ	0.5	0.4743	-0.0257	0.2130	0.9124	0.7	0.6281	-0.0719	0.0671	0.6591
	λ	1	1.2050	0.2050	0.7241	3.2141	0.8	0.7705	-0.0295	0.2252	2.9567
$n=50$	α	2	2.0675	0.0675	0.0621	0.3502	1.5	1.5133	0.0133	0.0319	0.3024
	β	1	1.0758	0.0758	0.2385	0.6444	0.5	0.5641	0.0641	0.1539	0.3816
	θ	0.5	0.4815	-0.0185	0.1861	0.9618	0.7	0.6202	-0.0798	0.0552	0.3252
	λ	1	0.8462	-0.1538	0.6177	2.9721	0.8	0.7888	-0.0112	0.2413	2.6142
$n=100$	α	2	2.0561	0.0561	0.0464	0.2216	1.5	1.5031	0.0310	0.0151	0.1699
	β	1	1.0877	0.0877	0.1445	0.3892	0.5	0.5494	0.0494	0.0483	0.2502
	θ	0.5	0.4921	-0.0079	0.0706	0.7271	0.7	0.6594	-0.0406	0.0387	0.6048
	λ	1	1.0921	0.0921	0.5822	2.3921	0.8	0.7354	-0.0646	0.2166	2.1081
$n=200$	α	2	1.9792	-0.0208	0.0162	0.1648	1.5	1.4985	-0.0015	0.0076	0.1263
	β	1	1.0393	0.0393	0.0485	0.2704	0.5	0.5185	0.0185	0.0176	0.1519
	θ	0.5	0.4216	-0.0787	0.6224	0.7334	0.7	0.6762	-0.0238	0.0198	0.4513
	λ	1	1.0434	0.0434	0.1905	2.1947	0.8	0.7766	-0.0234	0.2347	1.6921
$n=500$	α	2	1.9875	-0.0125	0.0085	0.1147	1.5	1.4955	-0.0045	0.0033	0.0914
	β	1	1.0079	0.0079	0.0277	0.1807	0.5	0.5053	0.0053	0.0084	0.1059
	θ	0.5	0.4888	-0.0112	0.0412	0.4425	0.7	0.6894	-0.0106	0.0177	0.3191
	λ	1	0.9871	-0.0129	0.0927	1.5535	0.8	0.8302	0.0302	0.0962	1.2971

7. Two application examples

In this section, we present two applications of LPS^2 family of distributions using real-life data sets. In the applications, we use the Adequacy Model package version 1.0.8 available in the R programming language. The fit is compared to other distributions based on the maximized log-likelihood, the Kolmogorov-Smirnov test (K-S), Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC). Finally, we provide the histograms of the data sets to have a visual comparison of the fitted density functions.

Data set 1

The first set consists of the number of successive failures for the air conditioning system of each member in a fleet of 13 Boeing 720 airplanes. The pooled data, yielding a total of 213 observations, were first analyzed by Proschan (1963) and further discussed in Dahiya and Gurland (1972), Adamidis and Loukas (1998) and Tahmasbi and Rezaei (2008). Table 3 gives some descriptive statistics for the first data set. Figure 5a displays the Gaussian kernel density estimation for this data set.

For this data set, the new distributions, exponential geometric binomial (EGBD) and

Table 3: Descriptive statistics for data set 1.

n	Mean	Q_1	Median	Q_3	Mode	Variance	Skewness	Kurtosis	Min	Max
213	93.14	22	57	118	14	11398.47	2.11	7.92	1	603

Weibull geometric geometric (WGGD) distributions given by the following pdfs were fitted:

$$f_{EGBD}(x;\alpha,\beta,\theta,\lambda) = \frac{m\lambda\beta(1-\theta)e^{-\beta x}\left[1-\theta+(\theta+\lambda)e^{-\beta x}\right]^{m-1}}{\left[(1+\lambda)^m-1\right]\left\{1-\theta\left(1-e^{-\beta x}\right)\right\}^{m+1}},$$

and

$$f_{WGGD}(x;\alpha,\beta,\gamma) = \frac{\alpha\beta\,(1-\gamma)\,x^{\alpha-1}e^{-\beta x^{\alpha}}}{\left\{1-\gamma e^{-\beta x^{\alpha}}\right\}^2}$$

for $x>0$, $\alpha,\beta>0$, $\gamma,\theta<1$, $\lambda\in\mathbb{R}$. For comparison purposes, we also fit the generalization of Weibull distribution (GWD) (Shanker and Shukla, 2019), the generalization of generalized gamma distribution(GGD) (Shanker and Shukla, 2019), beta exponential (BE) (Nadarajah and Kotz, 2006) and odd Weibull (OW) (Cooray, 2006) distributions. Estimates of the parameters of the distributions, standard errors (in parentheses), log-likelihood function evaluated at the parameter estimates, K-S statistic and its p-value are shown in Table 4. Furthermore, to compare the models, the AIC, AICc, and BIC indices are obtained too. In general, the smaller the values of these criteria, the better the fit. According to these formal tests, the WGGD model has the largest likelihood, the smallest K-S statistic, the largest p-value, and the smallest values for all other indices, among all fitted models.

Figure 6a gives the graph of the estimated pdfs of the WGGD, EGBD, and other competitive models that are used to fit the data after replacing the unknown parameters included in each distribution by their MLEs. The fitted pdf of the WGGD distribution captures the observed histograms better than others for the data sets 1. This real example suggests that the three-parameter WGGD fits data set 1 very well when compared to the other distributions.

Data set 2

The second application takes into account the data related to the breaking stress of carbon fibres of 50 mm in length from Nicholas and Padgett (2006). This data set was used by Cordeiro and Lemonte (2011) which is given in Table 5. Table 6 gives some descriptive statistics for this data set. Figure 5b displays the Gaussian kernel density estimation for the second data set.

For the second data set, the new distributions, generalized half normal geometri Poisson distribution (GHNGPD) and WGPD were fitted:

$$f_{CHNGPD}(x;\alpha,\beta,\theta,\lambda) = \frac{\sqrt{\frac{2}{\pi}}\alpha\beta\lambda\,(1-\theta)\,x^{\alpha-1}e^{-\frac{1}{2}(\beta x^{\alpha})^2}}{\left[e^{\lambda}-1\right]\left\{1-2\theta\Phi(-\beta x^{\alpha})\right\}^2}\exp\left[\frac{2\lambda\,(1-\theta)\,\Phi(-\beta x^{\alpha})}{1-2\theta\Phi(-\beta x^{\alpha})}\right],$$

Table 4: Estimates and goodness-of-fit measures for the first data set.

Model	MLEs (standard errors)	Log-likelihood	K-S	p-value	AIC	AIC$_c$	BIC
EGBD	0.0067, 0.0403, 0.2201	-1175.871	0.0564	0.633	2357.74	2357.85	2367.82
SE	(0.0022, 0.3957, 0.2170)						
WGGD	1.1982, 0.0017, 0.7651	-1174.180	0.0504	0.765	2354.36	2354.47	2364.44
SE	(0.029, 2.14×10^{-5}, 0.0430)						
GWD	0.9395, 0.9219, 0.0168	-1177.586	0.0662	0.425	2361.17	2361.34	2371.26
SE	(1.0679, 0.0487, 0.0168)						
GGD	3.7958, 0.4419, 0.6943, 0.6911	-1174.514	0.0537	0.685	2357.03	2357.218	2370.47
SE	(2.8461, 0.1786, 0.2453, 0.2701)						
BE	1.0483, 2.2710, 0.0104	-1177.771	0.0637	0.475	2361.54	2361.73	2371.62
SE	(0.5925, 1.5206, 0.0058)						
OW	0.6667, 0.0469, 1.4838	-1176.062	0.0587	0.581	2358.12	2358.24	2368.20
SE	(0.1581, 0.0312, 0.3907)						

Table 5: Breaking stress of carbon fibres data.

0.39	0.85	1.08	1.25	1.47	1.57	1.61	1.61	1.69	1.80	1.84
1.87	1.89	2.03	2.03	2.05	2.12	2.35	2.41	2.43	2.48	2.50
2.53	2.55	2.55	2.56	2.59	2.67	2.73	2.74	2.79	2.81	2.82
2.85	2.87	2.88	2.93	2.95	2.96	2.97	3.09	3.11	3.11	3.15
3.15	3.19	3.22	3.22	3.27	3.28	3.31	3.31	3.33	3.39	3.39
3.56	3.60	3.65	3.68	3.70	3.75	4.20	4.38	4.42	4.70	4.90

and

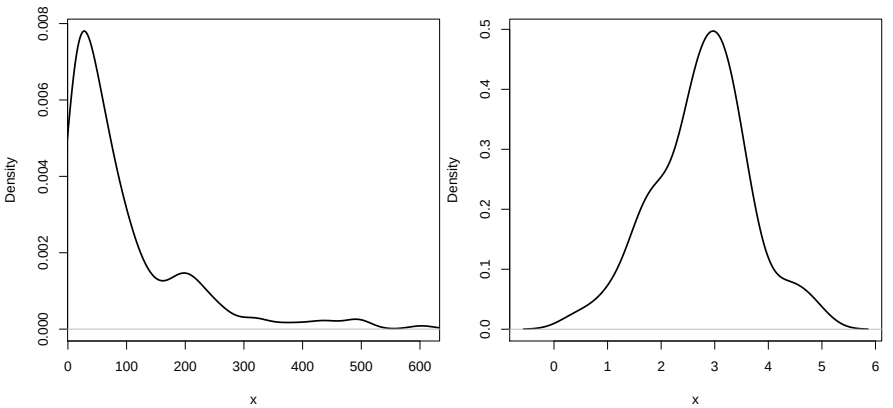
$$f_{WGPD}(x;\alpha,\beta,\theta,\lambda) = \frac{\alpha\beta\lambda(1-\theta)x^{\alpha-1}e^{-\beta x^{\alpha}}}{\left[e^{\lambda}-1\right]\left\{1-\theta e^{-\beta x^{\alpha}}\right\}^{2}} \exp\left[\frac{\lambda(1-\theta)e^{-\beta x^{\alpha}}}{1-\theta e^{-\beta x^{\alpha}}}\right]$$

for $x > 0$, $\alpha, \beta > 0$, $\theta < 1$, $\lambda \in \mathbb{R}$. We also fit the BE, beta Weibull (BW) (Famoye et al., 2005), Cauchy Weibull logistic (CWL) (Almheidat et al., 2015), Gumbel Weibull (GW) (Al-Aqtash et al., 2014) distributions to make a comparison with the new models. The parameter estimates, the log-likelihood values, the Kolmogorov-Smirnov statistics, and respective p-values are given in Table 7. Additionally, a comparison of these proposed distributions using the criteria, explained earlier, is presented.

It is observed that the WGPD distribution provides the best fit. In particular, we can see that the largest log-likelihood value, the largest p-value, the smallest AIC value, the smallest

Table 6: Descriptive statistics for data set 2.

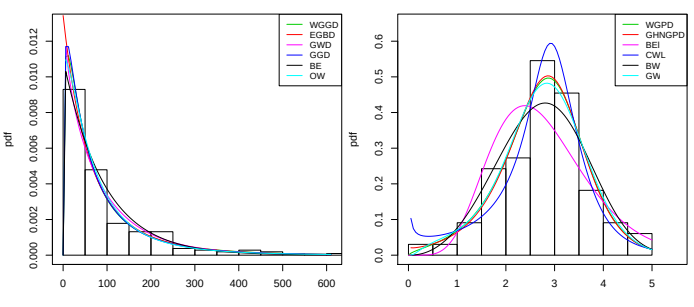
n	Mean	Q_1	Median	Q_3	Mode	Variance	Skewness	Kurtosis	Min	Max
66	2.178	2.178	2.853	3.278	1.61	0.795	-0.131	3.223	0.390	4.90



(a) Data set 1 (b) Data set 2

Figure 5: The Gaussian kernel density estimation for: (a) data set 1 (b) data set 2.

AICc value, and the smallest BIC value are obtained for the WGPD distribution. The fitted densities (with the respective histogram) are shown in Figure 6b. These indicate a good fit for the WGPD distribution for the second data set. It is clear from Tables 4 and 7 and also Figures 6a and 6b that the WGGD and WGPD models provide the best fits to these two real data sets.



(a) Data set 1 (b) Data set 2

Figure 6: Estimates of the density functions for the: (a) data set 1 (b) data set 2.

More on estimated hazard functions

The failure rate of a system usually depends on time, with the rate varying over the life cycle of the system. the hazard rate refers to the rate of failure for a system of a given age

Table 7: Estimates and goodness-of-fit measures for the second data set.

Model	MLEs (standard errors)	Log-likelihood	K-S	p-value	AIC	AIC$_c$	BIC
GHNGPD	0.8290, 1.0219, -37.8800, -0.8760	-84.849	0.0744	0.882	177.70	178.35	186.46
SE	(0.4745, 0.9104, 73.3213, 4.1871)						
WGPD	1.6669, 0.5589, -11.9420, -1.0240	-84.705	0.0723	0.902	177.41	178.07	186.17
SE	(0.8913, 0.9326, 29.3124, 2.8723)						
BE	0.1131, 7.5072, 20.9967	-91.223	0.1393	0.181	188.47	188.85	195.01
SE	(0.0170, 0.7642, 1.4865)						
CWL	2.1437, 7.9321, 2.9530	-86.989	0.1040	0.515	179.98	180.37	186.55
SE	(0.7221, 1.8887, 0.1083)						
BW	3.6790, 0.0136, 0.8820, 1.0594	-85.971	0.0830	0.786	179.94	180.60	188.70
SE	(0.7915, 0.0133, 0.3241, 1.2743)						
GW	3.4359, 5.5673, 2.4231, 1.1324	-84.834	0.0733	0.893	177.67	187.32	186.43
SE	(1.1494, 2.8064, 0.5078, 0.4524)						

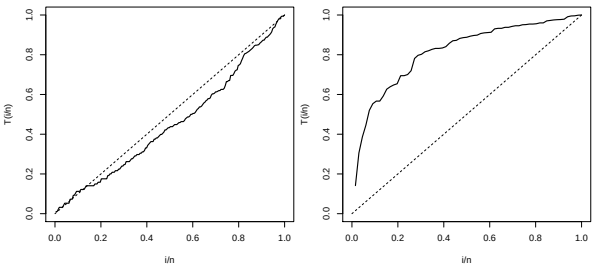
x and is defined as $h(x) = f(x)/S(x)$. Hazard rate provides an alternative characterization for the distribution of a random variable, especially when dealing with lifetime data and it is quite useful in defining and formulating a model. In this section, we focus on estimated hrfs as for the previous sections.

First, we provide the total time on test (TTT) transform procedure proposed by Aarset (1987) as a tool to identify the hazard behaviour of the distribution. The TTT-transform can illustrate the variety of the hazard rate curves for a lifetime distribution. If the empirical TTT-transform is convex and concave, the shape of the corresponding hrf is decreasing and increasing, respectively. If the TTT-transform is convex and concave, the hrf will have a bathtub shape. Finally, if the TTT-transform is concave and convex, a unimodal hrf will be more appropriate. Figure 7a shows that the TTT-plot for the first data set has a concave and convex shape. It indicates that the hrf has a unimodal shape. Figure 7b shows that the TTT-plot for the second data set has a concave shape. It indicates that the hrf has an increasing shape.

Graphs of the estimated hrfs are displayed in Figures 8 for data sets 1 and 2. Hence, the WGGD and WGPD distributions could be the appropriate models for the fitting of such data sets.

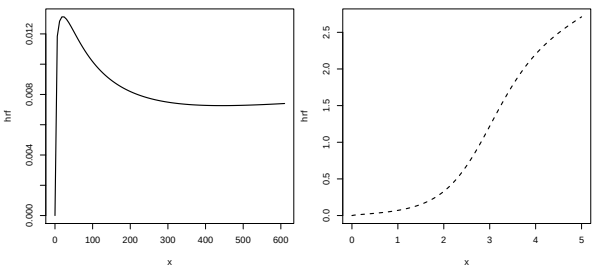
8. Conclusions

We introduce a new generalized class of lifetime distributions, called the LPS^2 family of distributions, by compounding a lifetime and twice power series distributions in a serial and parallel structure. The new models extend several distributions widely used in the lifetime



(a) Data set 1 (b) Data set 2

Figure 7: TTT-plot on the data sets 1 and 2.



(a) Data set 1 (b) Data set 2

Figure 8: Graphs of estimated hrf for the data sets 1 and 2.

literature such as the exponential power series, Weibull power series, and complementary of exponential power series distributions. The pdf of the new distributions can be expressed as a linear combination of baseline distributions and they have a hazard function that displays flexible behaviour. We provide a mathematical treatment of this family, including moments, quantiles, reliability functions, and moment generating function as well as the mean residual lifetime. The method of maximum likelihood was used to estimate the model parameters. We perform a Monte Carlo simulation study to assess the finite sample behavior of the maximum likelihood estimators. Some members of the LPS^2 family are fitted to two real data sets to illustrate the usefulness of the new distributions. They provide better fits than other competing models consistently.

References

Aarset, M. V., (1987). How to identify bathtub hazard rate. *IEEE Transactions on Reliability*, 36, pp. 106–108.

Adamidis, K., Dimitrakopoulou, T., and Loukas, S., (2005). On a generalization of the exponential geometric distribution. *Statistics and Probability Letters*, 73, pp. 259–269.

Adamidis, K., and Loukas, S., (1998). A lifetime distribution with decreasing failure rate. *Statistics and Probability Letters*, 39, pp. 35–42.

Al-Aqtash, R., Lee, C., and Famoye, L., (2014). Gumbel-Weibull Distribution: Properties and Applications. *Journal of Modern Applied Statistical Methods*, 13, pp. 201–225.

Al-Mheidat, M., Famoye, F., and Lee, C., (2015). Some generalized families of Weibull distribution: Properties and applications. *International Journal of Statistics and Probability*, 4, pp. 222–238.

Barreto-Souza, W., Morais, A., and Cordeiro, G.M., (2011). The Weibull-geometric distribution. *Journal of Statistical Computation and Simulation*, 81, pp. 645–657.

Barreto-Souza, W., Santos, A.H., and Cordeiro, G.M., (2010). The beta generalized exponential distribution. *Journal of Statistical Computation and Simulation*, 80, pp. 159–172.

Chahkandi M., and Ganjali, M., (2009). On some lifetime distributions with decreasing failure rate. *Computational Statistics and Data Analysis*, 53, pp. 4433–4440.

K. Cooray (2006). Generalization of the Weibull distribution: the odd Weibull family. *Statistical Modelling*, 6, pp. 265–277.

Cordeiro, G.M., and Lemonte, A., (2011). The β Birnbaum-Saunders distribution: An improved distribution for fatigue life modeling. *Computational Statistics and Data Analysis*, 55, pp. 1445–1461.

Eichhorn, S.J., and Davies, G.R., (2005). Modelling the crystalline deformation of native and regenerated cellulose. *Cellulose*, 13, pp. 291–307.

Famoye, F., Lee, C., and Olumolade O., (2005). The beta-Weibull distribution. *Journal of Statistical Theory and Application*, 4, pp. 122–136.

Flores, J., Borges, C., Cancho, V.G., and Louzada F., (2013). The complementary exponential power series distribution. *Brazilian Journal of Probability and Statistic*s, 27, pp. 565–584.

Greenwood, J.A., Landwehr, J.M., and Matalas, N.C., (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, 15, pp. 1049–1054.

Gurland, J., and Sethuraman, J.,(1994). Reversal of increasing failure rates when pooling failure data. *Technometrics*, 36, pp. 416–418.

Hassan, M.,Larson, L.E., Leung V.W., and Asbeck, P.M., (2012). A combined series-parallel hybrid envelope amplifier for envelope tracking mobile terminal RF power amplifier applications. *IEEE Journal of Solid-State Circuits*, 47, pp. 1185–1198.

Kazimierczuk, M.K., Thirunarayan, N., and Wang S., (1993). Analysis of series-parallel resonant converter. *IEEE transactions on aerospace and electronic systems*, 29, pp. 88–99.

Kenney, J., and Keeping., E., (1962). *Mathematics of Statistics*. Volume 1, Third edition, Van Nostrand, Princeton.

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

Morais, A.L., and Barreto-Souza, W., (2011). A compound class of Weibull and power series distributions. *Computational Statistics and Data Analysis*, 55, pp. 1410–1425.

Munteanu, B.G., Leahu, A., and Pârtachi, I., (2014). The max-Weibull power series distribution. *Analele UniversitY Oradea Fasc. Matematica*, 21, pp. 133–139.

Nadarajah, S., and Kotz, S., (2006). The beta exponential distribution. *Reliability Engineering & System Safety*, 91, pp. 689–697.

Nassar, M., Kumar, D., Dey, S., Cordeiro, G.M., and Afify, A.Z., (2019). The Marshall-Olkin alpha power family of distributions with applications. *Journal of Computational and Applied Mathematics*, 351, pp. 41–53.

Nicholas, M.D., and Padgett W.G., (2006). A bootstrap control chart for Weibull percentiles. *Quality and Reliability Engineering International*, 22, pp. 141–151.

Proschan, F., (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*, 5, pp. 375–383.

Ross, S.M., (2010). *Introduction to Probability Models*. Academic Press, Boston, 10th edition.

Shanker, R., Shukla, K.K., (2019). A generalization of Generalized Gamma distribution. *International Journal of Computational and Theoretical Statistics*, 6, pp. 33–42.

Shanker, R., Shukla, K.K., (2019). A generalization of Weibull distribution. *Reliability: Theory and Applications*, 14, pp. 57–70.

Smith, R.l., and Naylor, J.C., (1987). A comparison of maximum likelihood and Bayesian estimators for the three-parameter Weibull distribution. *Applied Statistics*, 36, pp. 358–369.

Tahmasbi, R., and Rezaei, S., (2008). A two-parameter lifetime distribution with decreasing failure rate. *Computational Statistics and Data Analysis*, 52, pp. 3889–3901

Xiong, W., Zhang, Y., and Yin, C., (2009). Optimal energy management for a series-parallel hybrid electric bus. *Energy Conversion and Management*, 50, pp. 1730–1738.

Appendix

Proof of (3) and (4)

Using the binomial expansion, we have

$$\begin{aligned} F(x;\boldsymbol{\xi}) &= 1-[A(\lambda)]^{-1}\sum_{u=1}^{\infty}a_u\lambda^u\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x,\boldsymbol{\varsigma}))\right\}^u \\ &= [A(\lambda)]^{-1}\sum_{u=1}^{\infty}a_u\lambda^u\left(1-\left\{1-[B(\theta)]^{-1}B(\theta\Pi(x,\boldsymbol{\varsigma}))\right\}^u\right) \\ &= [A(\lambda)]^{-1}\sum_{u=1}^{\infty}a_u\lambda^u\sum_{k=1}^{u}\binom{u}{k}(-1)^{k+1}[B(\theta)]^{-k}\left\{\sum_{z=1}^{\infty}b_z\theta^z[\Pi(x,\boldsymbol{\varsigma})]^z\right\}^k \\ &= [A(\lambda)]^{-1}\sum_{u=1}^{\infty}a_u\lambda^u\sum_{k=1}^{u}\binom{u}{k}(-1)^{k+1}[B(\theta)]^{-k}\sum_{j=0}^{\infty}l_{k,j}\theta^{k+j}[\Pi(x,\boldsymbol{\varsigma})]^{k+j}, \end{aligned}$$

where $j=z-1$ and $l_{k,j}=(jb_1)^{-1}\sum_{m=1}^{j}[m(j+1)-j]\,b_{m+1}l_{k,j-m}$. Then

$$F(x;\boldsymbol{\xi}) = \sum_{k=1}^{\infty}\sum_{j=0}^{\infty}\phi_{k,j}\Pi(x,\boldsymbol{\varsigma})^{k+j},$$

where

$$\phi_{k,j}=\phi_{k,j}(\theta,\lambda)=[A(\lambda)]^{-1}\sum_{u=k}^{\infty}\binom{u}{k}(-1)^{k+1}l_{k,j}a_u\lambda^u\theta^{k+j}[B(\theta)]^{-k}.$$

Finally, Equation (4) is obtained by using the direct differentiation of (3).

STATISTICS IN TRANSITION new series, March 2022
Vol. 23 No. 1, pp. 75–88, DOI 10.21307/stattrans-2022-005
Received – 21.12.2020; accepted – 21.11.2021

New improved Poisson and negative binomial item count techniques for eliciting truthful answers to sensitive questions

Barbara Kowalczyk[1], **Robert Wieczorkowski**[2]

ABSTRACT

Item count techniques (ICTs) are indirect survey questioning methods designed to deal with sensitive features. These techniques have gained the support of many applied researchers and undergone further theoretical development. Latterly in the literature, two new item count methods, called Poisson and negative binomial ICTs, have been proposed. However, if the population parameters of the control variable are not provided by the outside source, the methods are not very efficient. Efficiency is an important issue in indirect methods of questioning due to the fact that the protection of respondents' privacy is usually achieved at the expense of the efficiency of the estimation. In the present paper we propose new improved Poisson and negative binomial ICTs, in which two control variables are used in both groups, although in a different manner. In the paper we analyse best linear unbiased and maximum likelihood estimators of the proportion of the sensitive attribute in the population in the introduced new models. The theoretical findings presented in the paper are supported by a comprehensive simulation study. The improved procedure allowed the increase of the efficiency of the estimation compared to the original Poisson and negative binomial ICTs.

Key words: sensitive questions, indirect questioning methods, item count techniques, Poisson ICT, negative binomial ICT, EM algorithm

1. Methodology and questionnaire design

Reliable data on stigmatizing, socially unaccepted or illegal features are very hard to obtain in direct questioning. Many indirect methods of questioning have been developed to help in eliciting honest answers to sensitive questions and to eliminate the social desirability bias. Among them two methods are predominant: randomised response techniques (Warner 1965, Chaudhuri 2011, Imai 2015, Dihidar and

[1] SGH Warsaw School of Economics, Collegium of Economic Analysis, Poland.
E-mail: bkowal@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-5407-3438.
[2] Statistics Poland, Programming and Coordination of Statistical Surveys Department, Poland.
E-mail: R.Wieczorkowski@stat.gov.pl. ORCID: https://orcid.org/0000-0003-2706-4306.

Bhattacharya, 2017) and item count techniques (Miller, 1984, Blair and Imai, 2012, Chaudhuri and Christofides, 2007, Imai, 2011, Holbrook and Krosnick, 2010, Comsa and Postelnicu, 2013, Wolter and Laier, 2014, Kuha and Jackson, 2014, Trappman et al., 2014, Kowalczyk and Wieczorkowski, 2017, Krumpal et al., 2018). Item count techniques have many practical advantages (Tourangeau and Yan, 2007). They are very easy to implement, they do not require the use of any randomize device, and they are very easy to understand so the respondents realize how their privacy is being protected.

Latterly Tian et al. (2017) proposed new item count techniques, called Poisson and negative binomial ICTs. In their method (if the population parameters of the control variable are not given from the outside source) a sample of *n* elements is divided into a control group and a treatment group, of n_1 and n_2 elements respectively. Respondents in the control group are asked one neutral question with possible count outcomes 0,1,2,… An exemplary questionnaire might look like the following:

Q: How many times did you use an Uber last month? Your answer is ….

Respondents in the treatment group are presented with two questions: one exactly the same as in the control group, and one sensitive with possible outcomes 0 or 1. Respondents in the treatment group are asked to report only the sum of the two questions. An exemplary questionnaire might look like the following:

Q: How many times did you use an Uber last month?

S: Have you ever bribed an official? Assign number 1 if 'yes' (YES = 1) and number 0 if 'not' (NOT = 0).

Please report ONLY the sum of the two numbers. The sum is …

To increase efficiency of the estimation we propose a new item count method, which draws on the idea of the Poisson and negative binomial ICTs introduced by Tian et al. (2017) and advances the original method in order to attain greater efficiency of the estimation. Our improved technique incorporates the sensitive question in two groups and combines it with two different neutral questions. Below we describe the newly proposed methodology.

We divide the sample of *n* elements into the first and second treatment groups, of n_1 and n_2 elements respectively. In the first group respondents are asked one neutral question Q_1 with possible count outcomes 0,1,2,… Then respondents are presented with two questions, one neutral Q_2 with possible count outcomes 0,1,2,…, and one sensitive S with possible outcomes 0 or 1. To protect their privacy respondents are asked to report only the sum of their answers to questions Q_2 and S. They are never asked to report their answer to the sensitive question *S*. Below we give an exemplary questionnaire for the first treatment group.

Q_1: How many times did you use a taxi last month? Your answer is ….

Now, we show you two questions. Do not answer them until you read the end.

Q_2: How many times were you at the cinema last month?

S: Have you ever bribed an official? Assign number 1 if 'yes' (YES = 1) and number 0 if 'not' (NOT = 0).

Please report ONLY the sum of the two numbers. The sum is …

In the second treatment group neutral questions are switched. Therefore, an exemplary questionnaire for the second group is given below.

Q_2: How many times were you at the cinema last month? Your answer is ….

Now, we show you two questions. Do not answer them until you read the end.

Q_1: How many times did you use a taxi last month? Remember your number but do not reveal it.

S: Have you ever bribed an official? Assign number 1 if 'yes' (YES = 1) and number 0 if 'not' (NOT = 0).

Please report ONLY the sum of the two numbers. The sum is …

It is very important that the sensitive question is mentioned only once in each group and the respondents are never asked to answer the sensitive question directly. To assure complete privacy the two neutral questions should be unrelated with each other and unrelated with the sensitive question. It also ensures that the privacy protection level in the newly proposed methods is exactly the same as in the original Poisson and negative binomial ICTs.

2. Statistical model and estimation

2.1. Notation

Let $X^{(1)}$, $X^{(2)}$ denote control variables being the answers to the neutral questions Q_1 and Q_2 respectively, and let Z denote a Bernoulli distributed variable being the answer to the sensitive question *S*. To assure complete protection of the privacy we assume that $X^{(1)}, X^{(2)}, Z$ are independent. Let $P(Z = 1) = \pi$ be an unknown sensitive proportion under study. Let $Y^{(1)}$ denote an observed variable indicating the sum of answers to questions Q_2 and *S* in the first treatment group, i.e. $Y^{(1)} = X^{(2)} + Z$. Analogously, let $Y^{(2)}$ be an observed variable indicating the sum of answers to questions Q_1 and *S* in the second treatment group, i.e. $Y^{(2)} = X^{(1)} + Z$. In the first treatment group we have two vectors of observed variables: $\left(X_1^{(1)}, \dots, X_{n_1}^{(1)}\right)$ and $\left(Y_1^{(1)}, \dots, Y_{n_1}^{(1)}\right)$. In the second group vectors of observed variables are $\left(X_{n_1+1}^{(2)}, \dots, X_{n_1+n_2}^{(2)}\right)$ and $\left(Y_{n_1+1}^{(2)}, \dots, Y_{n_1+n_2}^{(2)}\right)$. Sensitive variable under study *Z* is not directly observable in this model. *Z* is a latent variable, which is in line with the principle of privacy protection.

2.2. Best linear unbiased estimator

We consider a linear estimator of the form

$$\hat{\pi} = \sum_{i=1}^{n_1} \alpha_i X_i^{(1)} + \sum_{i=1}^{n_1} \beta_i Y_i^{(1)} + \sum_{j=n_1+1}^{n_1+n_2} \gamma_i X_i^{(2)} + \sum_{i=n_1+1}^{n_1+n_2} \delta_i Y_i^{(2)}, \quad (1)$$

where $\alpha, \beta, \gamma, \delta$ are constants weight factors. We determine $\alpha, \beta, \gamma, \delta$ so as to minimize variance $Var(\hat{\pi})$ of the estimator $\hat{\pi}$ subject to the condition that this estimator is unbiased.

Conditions for unbiasedness are

$$\begin{cases} \sum_{i=1}^{n_1} \alpha_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i = 0 \\ \sum_{i=1}^{n_1} \beta_i + \sum_{j=n_1+1}^{n_1+n_2} \gamma_i = 0 \\ \sum_{i=1}^{n_1} \beta_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i = 1 \end{cases} \quad (2)$$

To achieve the smallest variance, the expression to be minimized is

$$Var(\hat{\pi}) - \lambda_1\left(\sum_{i=1}^{n_1} \alpha_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i\right) - \lambda_2\left(\sum_{i=1}^{n_1} \beta_i + \sum_{j=n_1+1}^{n_1+n_2} \gamma_i\right) - \\ -\lambda_3\left(\sum_{i=1}^{n_1} \beta_i + \sum_{i=n_1+1}^{n_1+n_2} \delta_i - 1\right) \quad (3)$$

The minimization leads to the best linear unbiased estimator (BLUE) of the sensitive population proportion π, which can be written in the final form

$$\hat{\pi} = w\left(\bar{Y}^{(2)} - \bar{X}^{(1)}\right) + (1-w)\left(\bar{Y}^{(1)} - \bar{X}^{(2)}\right) \quad (4)$$

where

$$w = \frac{Var\left(\bar{Y}^{(1)}\right) + Var\left(\bar{X}^{(2)}\right)}{Var\left(\bar{Y}^{(2)}\right) + Var\left(\bar{X}^{(1)}\right) + Var\left(\bar{Y}^{(1)}\right) + Var\left(\bar{X}^{(2)}\right)} \quad (5)$$

Variance of the BLUE estimator is

$$Var(\hat{\pi}) = \frac{\frac{1}{n_1 n_2}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right)\left(2Var\left(X^{(2)}\right) + \pi(1-\pi)\right)}{\frac{1}{n_1}\left(2Var\left(X^{(2)}\right) + \pi(1-\pi)\right) + \frac{1}{n_2}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right)} \quad (6)$$

For $n_1 = n_2 = 0.5n$ and for $Var\left(X^{(2)}\right) = Var\left(X^{(1)}\right)$ formula (6) simplifies to the form

$$Var(\hat{\pi}) = \frac{1}{n}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right) \quad (7)$$

For $n_1 = n_2 = 0.5n$ variance of the method of moment estimator in original Tian et al. (2017) Poisson and negative binomial ICTs with one neutral variable $X^{(1)}$ is

$$Var\left(\hat{\pi}^{orig}\right) = \frac{2}{n}\left(2Var\left(X^{(1)}\right) + \pi(1-\pi)\right) \quad (8)$$

From (7) and (8) it can be easily seen that for $n_1 = n_2 = 0.5n$ and for $Var\left(X^{(2)}\right) = Var\left(X^{(1)}\right)$ we get

$$Var(\hat{\pi}) = 0.5Var\left(\hat{\pi}^{orig}\right) \quad (9)$$

and the theoretical BLUE estimator in the improved model is more efficient than the method of moment estimator in the original model. Due to the fact that variances that

appear in formula (5) are not known in advance the theoretical BLUE estimator cannot be used directly. Therefore, we propose to use in practice the empirical BLUE estimator (EBLUE) of the form

$$\hat{\pi}^{emp} = \hat{w}^{emp}\left(\bar{Y}^{(2)} - \bar{X}^{(1)}\right) + \left(1 - \hat{w}^{emp}\right)\left(\bar{Y}^{(1)} - \bar{X}^{(2)}\right) \tag{10}$$

where

$$w^{emp} = \frac{\frac{1}{n_1}s^2\left(Y^{(1)}\right) + \frac{1}{n_2}s^2\left(X^{(2)}\right)}{\frac{1}{n_2}s^2\left(Y^{(2)}\right) + \frac{1}{n_1}s^2\left(X^{(1)}\right) + \frac{1}{n_1}s^2\left(Y^{(1)}\right) + \frac{1}{n_2}s^2\left(X^{(2)}\right)} \tag{11}$$

and $s^2\left(X^{(1)}\right)$, $s^2\left(X^{(2)}\right)$, $s^2\left(Y^{(1)}\right)$, $s^2\left(Y^{(2)}\right)$ are sample variances of observed variables $X^{(1)}$, $X^{(2)}$, $Y^{(1)}$, $Y^{(2)}$ respectively. Properties of the proposed EBLUE estimator of the sensitive proportion π are analyzed in Section 3.

2.3. Maximum likelihood estimation via EM algorithm

In our model, the sensitive variable under study $Z \sim Bernoulli(\pi)$ is not directly observable. In order to obtain maximum likelihood estimators in models with latent variables it is convenient to use expectation maximization (EM) algorithm introduced by Dempster et al. (1977) and further developed by, e.g. McLachlan and Krishnan (2008). EM algorithm has also become a standard tool for determining ML estimators when dealing with item count techniques, see, e.g. Imai (2011), Kuha and Jackson (2014), Tian et al. (2017). Complete log-likelihood function in our model is

$$\begin{aligned} &lnL_{com}\left(\pi, \theta_1, \theta_2; x^{(1)}, x^{(1)}, x^{(2)}, y^{(1)}, y^{(2)}, z\right) = \\ &= \sum_{i=1}^{n_1} lnp_{\theta_1}(x_i) + \sum_{j=n_1+1}^{n_1+n_2} lnp_{\theta_2}(x_j) + \\ &+ \sum_{i=1}^{n_1} z_i lnp_{\theta_2}(y_i - 1) + \sum_{j=n_1+1}^{n_1+n_2} z_j lnp_{\theta_1}(y_j - 1) + \\ &+ \sum_{i=1}^{n_1}(1 - z_i) lnp_{\theta_2}(y_i) + \sum_{j=n_1+1}^{n_1+n_2}(1 - z_j) lnp_{\theta_1}(y_j) + \\ &+ \sum_{j=1}^{n_1+n_2} z_j \ln \pi + \sum_{j=1}^{n_1+n_2}(1 - z_j)\ln(1 - \pi), \end{aligned} \tag{12}$$

where $p_{\theta_1}(x)$ and $p_{\theta_2}(x)$ are probability mass functions of the control variables $X^{(1)}$ and $X^{(2)}$ respectively. Conditional expectation computed in E-step of the EM algorithm is

$$\begin{aligned} &E_{\pi_0, \theta_{10}, \theta_{2o}}[lnL_{com}(\pi, \theta_1, \theta_2; y, Z | Y = y)] = \\ &= \sum_{i=1}^{n_1} lnp_{\theta_1}(x_i) + \sum_{i=n_1+1}^{n_1+n_2} lnp_{\theta_2}(x_j) + \\ &+ \sum_{i=1}^{n_1} \check{z}_i lnp_{\theta_2}(y_i - 1) + \sum_{j=n_1+1}^{n_1+n_2} \check{z}_j lnp_{\theta_1}(y_j - 1) + \\ &+ \sum_{i=1}^{n_1}(1 - \check{z}_i) lnp_{\theta_2}(y_i) + \sum_{j=n_1+1}^{n_1+n_2}(1 - \check{z}_j) lnp_{\theta_1}(y_j) + \\ &+ \sum_{j=1}^{n_1+n_2} \check{z}_j \ln \pi + \sum_{j=1}^{n_1+n_2}(1 - \check{z}_j)\ln(1 - \pi) \end{aligned} \tag{13}$$

where

$$\check{z}_i = E_{\pi_0,\theta_{2o}}\left(Z_i \middle| Y_i^{(1)} = y_i\right) = \frac{p_{\theta_{20}}(y_i-1)\pi_0}{p_{\theta_{20}}(y_i-1)\pi_0+p_{\theta_{20}}(y_i)(1-\pi_0)} \text{ for } i = 1,\dots,n_1 \tag{14}$$

$$\check{z}_j = E_{\pi_0,\theta_{10}}\left(Z_j \middle| Y_j^{(2)} = y_j\right) = \frac{p_{\theta_{10}}(y_j-1)\pi_0}{p_{\theta_{10}}(y_j-1)\pi_0+p_{\theta_{10}}(y_j)(1-\pi_0)} \text{ for } j = n_1+1,\dots,n_1+n_2 \tag{15}$$

To represent distribution of the count data Poisson and negative binomial distributions are commonly used. Therefore we consider three different cases: when both neutral variables follow Poisson distribution, when one neutral variable follows Poisson and the other neutral variable follows negative binomial distribution, and the last case, when both neutral variables follow negative binomial distributions.

Consider the first case where both neutral variables follow Poisson distribution. In this case we have $X^{(1)}\sim Poisson(\lambda_1)$, $X^{(2)}\sim Poisson(\lambda_2)$. Based on (12-15) we derive final iterative formulas for ML estimators via E-step and M-step of the EM algorithm as below.

E Step:

$$\check{z}_i = E\left(Z_i|Y_i^{(1)}\right) = \frac{y_i{}^{(1)}\pi}{y_i{}^{(1)}\pi+\lambda_2(1-\pi)} \text{ for } i = 1,\dots,n_1 \tag{16}$$

$$\check{z}_j = E\left(Z_j|Y_j^{(2)}\right) = \frac{y_j{}^{(2)}\pi}{y_j{}^{(2)}\pi+\lambda_1(1-\pi)} \text{ for } j = n_1+1,\dots,n_1+n_2 \tag{17}$$

M step:

$$\hat{\pi} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}\check{z}_i + \sum_{j=n_1+1}^{n_1+n_2}\check{z}_j\right) \tag{18}$$

$$\hat{\lambda}_1 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}x_i^{(1)} + \sum_{j=n_1+1}^{n_1+n_2}\left(y_j{}^{(2)} - \check{z}_j\right)\right) \tag{19}$$

$$\hat{\lambda}_2 = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1}(y_i{}^{(1)} - \check{z}_i) + \sum_{j=n_1+1}^{n_1+n_2}x_j^{(2)}\right) \tag{20}$$

When one variable follows Poisson distribution and the other one follows negative binomial distribution, say $X^{(1)}\sim Poisson(\lambda)$, $X^{(2)}\sim NB(r,p)$, first we assess parameter r based on the second treatment group

$$\hat{r} = \frac{\left(\bar{x}^{(2)}\right)^2}{S^2(X^{(2)})-\bar{x}^{(2)}} \tag{21}$$

and then we derive iterative formulas for the ML estimators via E-step and M-step of the EM algorithm as below.

E Step:

$$E\left(Z_i|Y_i^{(1)}\right) = \frac{y_i{}^{(1)}\pi}{y_i{}^{(1)}\pi+\left(y_i{}^{(1)}+r-1\right)p(1-\pi)} \text{ for } i = 1,\dots,n_1 \tag{22}$$

$$E\left(Z_j|Y_j^{(2)}\right) = \frac{y_j{}^{(2)}\pi}{y_j{}^{(2)}\pi+\lambda(1-\pi)} \text{ for } j = n_1+1,\dots,n_1+n_2 \tag{23}$$

M step:

$$\hat{\pi} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} z_i + \sum_{j=n_1+1}^{n_1+n_2} z_j\right) \tag{24}$$

$$\hat{\lambda} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} x_i^{(1)} + \sum_{j=n_1+1}^{n_1+n_2}\left(y_j{}^{(2)} - z_j\right)\right) \tag{25}$$

$$\hat{\boldsymbol{p}} = \frac{\left(\sum_{i=1}^{\boldsymbol{n_1}}(y_i{}^{(1)}-\boldsymbol{z_i})+\sum_{j=\boldsymbol{n_1}+1}^{\boldsymbol{n_1}+\boldsymbol{n_2}}\boldsymbol{x_j^{(2)}}\right)}{(\boldsymbol{n_1}+\boldsymbol{n_2})r+\left(\sum_{i=1}^{\boldsymbol{n_1}}(y_i{}^{(1)}-\boldsymbol{z_i})+\sum_{j=\boldsymbol{n_1}+1}^{\boldsymbol{n_1}+\boldsymbol{n_2}}\boldsymbol{x_j^{(2)}}\right)} \tag{26}$$

When both neutral variables follow negative binomial distribution, say $X^{(1)}\sim NB(r_1, p_1)$ and $X^{(2)}\sim NB(r_2, p_2)$, we first assess parameters r_1, r_2 based on the first and second treatment groups respectively by:

$$\hat{r_1} = \frac{\left(\bar{x}^{(1)}\right)^2}{S^2(X^{(1)})-\bar{x}^{(1)}} \tag{27}$$

$$\hat{r_2} = \frac{\left(\bar{x}^{(2)}\right)^2}{S^2(X^{(2)})-\bar{x}^{(2)}} \tag{28}$$

Next we derive formulas necessary to implement the EM algorithm.

E Step:

For $i = 1, \dots, n_1$:

$$E\left(Z_i|Y_i{}^{(1)}\right) = \frac{y_i{}^{(1)}\pi}{y_i{}^{(1)}\pi+\left(y_i{}^{(1)}+r_2-1\right)p_2(1-\pi)} \tag{29}$$

For $j = n_1 + 1, \dots, n_1 + n_2$:

$$E\left(Z_j|Y_j{}^{(2)}\right) = \frac{y_j{}^{(2)}\pi}{y_j{}^{(2)}\pi+\left(y_j{}^{(2)}+r_1-1\right)p_1(1-\pi)} \tag{30}$$

M step:

$$\hat{\pi} = \frac{1}{n_1+n_2}\left(\sum_{i=1}^{n_1} z_i + \sum_{j=n_1+1}^{n_1+n_2} z_j\right) \tag{31}$$

$$\hat{p}_1 = \frac{\left(\sum_{i=1}^{n_1} x_i^{(1)}+\sum_{j=n_1+1}^{n_1+n_2}(y_j{}^{(2)}-z_j)\right)}{(n_1+n_2)r_1+\left(\sum_{i=1}^{n_1} x_i^{(1)}+\sum_{j=n_1+1}^{n_1+n_2}(y_j{}^{(2)}-z_j)\right)} \tag{32}$$

$$\hat{p}_2 = \frac{\left(\sum_{i=1}^{n_1}(y_i{}^{(1)}-z_i)+\sum_{j=n_1+1}^{n_1+n_2} x_j^{(2)}\right)}{(n_1+n_2)r_2+\left(\sum_{i=1}^{n_1}(y_i{}^{(1)}-z_i)+\sum_{j=n_1+1}^{n_1+n_2} x_j^{(2)}\right)} \tag{33}$$

3. Simulation studies

To examine properties of the proposed improved Poisson and negative binomial ICTs and compare them with original Tian et al. (2017) design we conduct a comprehensive simulation study. For each set of model parameters separately, namely for $n = 500, 1000, 2000$ and for $\pi = 0.05, 0.1, 0.2, 0.3$, we generate n independent

variables $Z_1, Z_2, \dots, Z_n$ from *Bernoulli*(π) distribution. We use these once generated variables for all models considered in this section. Next, for each set of model parameters we generate $0.5n$ independent variables $X_1^{(1)}, \dots, X_{0.5n}^{(1)}$ from $Poisson(\lambda_1)$ distribution. These variables are used for both improved and original Poisson ICTs. For the improved Poisson ICT (Poisson-Poisson model) we additionally generate $0.5n$ independent variables $X_{0.5n+1}^{(2)}, \dots, X_n^{(2)}$ from $Poisson(\lambda_2)$ distribution. Next, we generate $0.5n$ independent variables $X_1^{(1)}, \dots, X_{0.5n}^{(1)}$ from $NB(r_1, p_1)$ distribution. We use these variables for both improved and original negative binomial ICTs. For the improved negative binomial ICT (NB-NB model) we additionally generate $0.5n$ independent variables $X_{0.5n+1}^{(2)}, \dots, X_n^{(2)}$ from $NB(r_2, p_2)$ distribution. Last but not least we generate $0.5n$ independent variables $X_1^{(1)}, \dots, X_{0.5n}^{(1)}$ from $Poisson(\lambda_1)$ and $0.5n$ independent variables $X_{0.5n+1}^{(2)}, \dots, X_n^{(2)}$ from $NB(r, p)$ distribution for the improved Poisson-NB model.

Based on the generated values we obtained n realizations of the two dimensional observable variables (X, Y) in the new models

$$(X_j, Y_j) = \begin{cases} \left(X_j^{(1)}, X_j^{(2)} + Z_j\right) & for\ j = 1, \dots, 0.5n \\ \left(X_j^{(2)}, X_j^{(1)} + Z_j\right) & for\ j = 0.5n + 1, \dots, n \end{cases},$$

and in the original Tian et al. (2017) models

$$Y_j = \begin{cases} X_j^{(1)} & for\ j = 1, \dots, 0.5n \\ X_j^{(1)} + Z_j & for\ j = 0.5n + 1, \dots, n \end{cases}.$$

Finally, we calculated EBLUE and ML estimators via EM algorithm according to formulas obtained in Section 2 and analogous MM and ML estimators according to formulas given in Tian et al. (2017). This process was replicated for each set of model parameters independently 10 000 times. In the simulation study we consider values $\pi \leq 0.3$. This corresponds to applications as the proportion of individuals possessing the sensitive feature is usually not very high in the general population.

The R codes used in our simulations are available at

https://github.com/rwieczor/ICT_Poisson_Negativebinomial.

In Table 1 root mean square error and bias of empirical best linear unbiased estimator is presented for different overall sample sizes, different sensitive proportions, and different models. It should be noted that obtained values of the RMSE of the EBLUE estimators are very close to the theoretical values $\sqrt{Var(\hat{\pi})}$, where $Var(\hat{\pi})$ is the variance of the theoretical BLUE estimator given in formula (7). RMSE and bias of maximum likelihood (ML) estimator is presented in Table 2. Naturally efficiency of the estimation increases (RMSE decreases) with the increase of the sample size. By

comparing Tables 1 and 2 it can be easily seen that ML estimators are more efficient than the corresponding EBLUE estimators. Advantage of ML over EBLUE estimators in terms of efficiency is especially highly visible for the small sample sizes and small values of π. Bias of the EBLUE estimators is very small. For ML estimators bias is visible for small values of π and small values of n.

Table 1. RMSE and BIAS (in parenthesis) of the EBLUE for different model parameters in the new model

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X^{(1)} \sim Poisson(2), X^{(2)} \sim Poisson(2)$				
n = 500	0.089 (0.001)	0.090 (-0.001)	0.091 (-0.001)	0.091 (0.000)
n = 1000	0.063 (0.001)	0.065 (0.001)	0.064 (0.001)	0.064 (0.000)
n = 2000	0.045 (0.000)	0.045 (0.000)	0.046 (0.000)	0.046 (0.001)
$X^{(1)} \sim Poisson(2), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.092 (0.000)	0.093 (0.000)	0.093 (-0.001)	0.093 (0.001)
n = 1000	0.065 (0.000)	0.065 (0.000)	0.067 (0.000)	0.066 (0.000)
n = 2000	0.046 (0.000)	0.046 (0.000)	0.046 (0.000)	0.046 (0.000)
$X^{(1)} \sim NB(r = 2, p = 0.4), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.096 (0.001)	0.095 (0.001)	0.097 (0.001)	0.096 (-.002)
n = 1000	0.068 (0.000)	0.067 (-0.001)	0.068 (0.001)	0.068 (0.001)
n = 2000	0.047 (0.000)	0.048 (-0.001)	0.048 (0.000)	0.048 (0.000)

Table 2. RMSE and BIAS (in parenthesis) of the ML estimators in the new model

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X^{(1)} \sim Poisson(2), X^{(2)} \sim Poisson(2)$				
n = 500	0.070 (0.017)	0.079 (0.006)	0.087 (-0.001)	0.086 (0.000)
n = 1000	0.052 (0.008)	0.061 (0.002)	0.062 (0.001)	0.060 (0.000)
n = 2000	0.040 (0.003)	0.044 (0.000)	0.044 (0.000)	0.043 (0.001)
$X^{(1)} \sim Poisson(2), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.067 (0.015)	0.076 (0.006)	0.080 (-0.002)	0.077 (-.002)
n = 1000	0.052 (0.007)	0.057 (0.002)	0.058 (-0.001)	0.054 (-.002)
n = 2000	0.039 (0.003)	0.043 (0.000)	0.041 (0.000)	0.038 (-.001)
$X^{(1)} \sim NB(r = 2, p = 0.4), X^{(2)} \sim NB(r = 2, p = 0.4)$				
n = 500	0.062 (0.013)	0.071 (0.004)	0.074 (-0.003)	0.070 (-.005)
n = 1000	0.049 (0.007)	0.054 (0.001)	0.054 (-0.001)	0.050 (-.002)
n = 2000	0.037 (0.003)	0.040 (-0.001)	0.038 (-0.001)	0.035 (-.001)

In Tables 3-4 we present RMSE of moments and ML estimators in original Tian et. al (2017) Poisson and negative-binomial ICTs. It should be noted that obtained values of the RMSE of moments estimators are very close to the theoretical values $\sqrt{Var(\hat{\pi}^{orig})}$, where $Var(\hat{\pi}^{orig})$ is given in formula (8). By determining sample sizes and the privacy protection level at the same level we can see that the new proposed models are more efficient. Gain in efficiency is achieved for all sample sizes and all values of π when comparing ML estimators in original and improved techniques and also when comparing MM with EBLUE estimators. It has to be emphasized that the new models resulted also in smaller bias when ML estimators are concerned.

Table 3. RMSE and BIAS (in parenthesis) of moments estimators in original Tian et al. (2017) Poisson and negative-binomial ICTs

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X \sim Poisson(2)$				
n = 500	0.128 (-.001)	0.128 (0.000)	0.128 (-0.002)	0.129 (-.001)
n = 1000	0.091 (0.001)	0.092 (0.000)	0.090 (0.001)	0.091 (0.000)
n = 2000	0.064 (0.000)	0.064 (0.000)	0.065 (0.000)	0.065 (0.001)
$X \sim NB(r = 2, p = 0.4)$				
n = 500	0.136 (0.003)	0.135 (0.002)	0.137 (-0.001)	0.136 (-.002)
n = 1000	0.094 (0.000)	0.095 (0.000)	0.096 (0.001)	0.096 (0.001)
n = 2000	0.068 (-0.001)	0.069 (-0.002)	0.069 (0.001)	0.068 (-.001)

Table 4. RMSE and BIAS (in parenthesis) of ML estimators in original Tian et. al (2017) Poisson and negative-binomial ICTs

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
$X \sim Poisson(2)$				
n = 500	0.095 (0.030)	0.105 (0.016)	0.117 (0.000)	0.120 (-.002)
n = 1000	0.071 (0.018)	0.081 (0.006)	0.086 (0.001)	0.086 (-.001)
n = 2000	0.052 (0.008)	0.06 (0.001)	0.063 (-0.001)	0.060 (0.001)
$X \sim NB(r = 2, p = 0.4)$				
n = 500	0.083 (0.024)	0.091 (0.008)	0.101 (-0.006)	0.098 (-.008)
n = 1000	0.063 (0.014)	0.071 (0.002)	0.075 (-0.002)	0.070 (-.004)
n = 2000	0.048 (0.005)	0.055 (-0.001)	0.055 (-0.001)	0.050 (-.002)

For further investigation let us consider succeeding model parameters and compare ML estimators in the improved and original Tian et al. (2017) Poisson ICT. Results of the simulation studies are given in Tables 5 and 6. In all cases both RMSE and BIAS of the ML estimators are visibly smaller when using newly proposed models.

Table 5. RMSE and BIAS (in parenthesis) of the ML estimators in the new model and original Poisson ICT

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim Poisson(1)$, $X^{(2)} \sim Poisson(1)$				
n = 500	0.053 (0.009)	0.06 (0.002)	0.063 (-0.001)	0.060 (0.000)
n = 1000	0.040 (0003)	0.045 (0.000)	0.044 (-0.001)	0.042 (-.001)
n = 2000	0.030 (0.001)	0.032 (-0.001)	0.031 (0.000)	0.030 (0.000)
Original model $X \sim Poisson(1)$				
n = 500	0.070 (0.017)	0.079 (0.006)	0.087 (-0.001)	0.084 (0.000)
n = 1000	0.053 (0.007)	0.060 (0.001)	0.062 (-0.001)	0.060 (-.001)
n = 2000	0.040 (0.003)	0.044 (0.000)	0.044 (-0.001)	0.042 (-.0010
New model $X^{(1)} \sim Poisson(3)$, $X^{(2)} \sim Poisson(3)$				
n = 500	0.083 (0.024)	0.093 (0.011)	0.106 (0.001)	0.106 (-.001)
n = 1000	0.062 (0.013)	0.071 (0.004)	0.076 (-0.002)	0.076 (-.001)
n = 2000	0.047 (0.005)	0.053 (0.001)	0.054 (0.000)	0.054 (0.000)
Original model $X \sim Poisson(3)$				
n = 500	0.116 (0.043)	0.124 (0.026)	0.141 (0.007)	0.146 (0.000)
n = 1000	0.084 (0.024)	0.093 (0.010)	0.104 (0.000)	0.108 (-.001)
n = 2000	0.061 (0.012)	0.070 (0.004)	0.076 (0.000)	0.075 (0.000)

Table 6. RMSE and BIAS (in parenthesis) of the ML estimators in the new model and original negative-binomial ICT

Sample size	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.2$	$\pi = 0.3$
New model $X^{(1)} \sim NB(r = 2, p = 0.6)$, $X^{(2)} \sim NB(r = 2, p = 0.6)$				
n = 500	0.094 (0.029)	0.102 (0.012)	0.115 (-0.002)	0.112 (-.007)
n = 1000	0.072 (0.017)	0.080 (0.004)	0.087 (-0.003)	0.081 (-.005)
n = 2000	0.055 (0.009)	0.062 (0.000)	0.062 (-0.002)	0.057 (-.002)
Original model $X \sim NB(r = 2, p = 0.6)$				
n = 500	0.124 (0.046)	0.131 (0.026)	0.147 (0.004)	0.153 (-.010)
n = 1000	0.096 (0.031)	0.103 (0.013)	0.116 (-0.003)	0.114 (-.009)
n = 2000	0.071 (0.016)	0.080 (0.003)	0.086 (-0.002)	0.082 (-.003)
New model $X^{(1)} \sim NB(r = 3, p = 0.5)$, $X^{(2)} \sim NB(r = 3, p = 0.5)$				
n = 500	0.098 (0.033)	0.108 (0.018)	0.118 (-0.001)	0.118 (-.004)
n = 1000	0.074 (0.019)	0.083 (0.008)	0.089 (-0.001)	0.085 (-.002)
n = 2000	0.056 (0.010)	0.064 (0.002)	0.064 (-0.001)	0.060 (-.001)
Original model $X \sim NB(r = 3, p = 0.5)$				
n = 500	0.133 (0.052)	0.139 (0.034)	0.153 (0.005)	0.160 (-.004)
n = 1000	0.101 (0.033)	0.108 (0.017)	0.118 (0.000)	0.120 (-.003)
n = 2000	0.073 (0.018)	0.083 (0.007)	0.089 (-0.001)	0.085 (-.003)

It is worth mentioning that in all comparisons we have set the overall sample size and privacy protection at the same level. Privacy protection is usually measured by the probability that the respondent possesses the sensitive attribute conditional on his or her answer. This probability was set to be the same in both compared methods by attaching the identical parameters to the control neutral variable associated with the sensitive one. In some surveys, however, asking a sensitive question – even indirectly – can be slightly more costly than asking the neutral one. In the new methods an indirect question about the sensitive variable is asked in the two groups and also two neutral questions are asked. Therefore, the newly proposed techniques can be slightly more costly in some situations, which also should be mentioned. However, this does not seem to apply to all surveys. Nevertheless, evident advantages of the newly proposed techniques in terms of efficiency and privacy protection should initiate its further development and application.

4. Conclusions

Item count techniques have attracted much attention among applied researchers. Methodology and theory of this method is still being developed, with a significant contribution by Tian et al. (2017), who introduced Poisson and negative binomial item count techniques. The two techniques allow for eliciting honest answers to sensitive questions, simplify the questionnaire design and theory. But this effect is achieved at the expense of the efficiency of the estimation, which is not high in the proposed techniques. In the paper three new models are proposed: Poisson-Poisson neutral questions ICT, Poisson-negative binomial neutral questions ICT, and negative binomial-negative binomial neutral questions ICT. Newly proposed methods maintain privacy of respondents at the same level regarding the sensitive question. At the same time the three newly proposed techniques increase efficiency of the estimation, which is very important in indirect methods of questioning.

References

Blair, G., Imai, K., (2012). Statistical Analysis of List Experiments. Polit Anal 20, pp. 47–77.

Chaudhuri, A., (2011). Randomized response and indirect questioning techniques in surveys, CRC Press, Boca Raton, FL.

Chaudhuri, A., Christofides, T. C., (2007). Item Count Technique in estimating the proportion of people with a sensitive feature. J Stat Plann Inference 137, pp. 589–593.

Comsa, M., Postelnicu C., (2013). Measuring Social Desirability Effects on Self-Reported Turnout Using the Item-Count Technique. Int J Public Opin Res 25, pp. 153–172.

Dempster, A.P., Laird, N. M., Rubin, D. B., (1977). Maximum-likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society Series B, Vol. 39, pp. 1–37.

DIHIDAR, K., BHATTACHARYA, M., (2017). Estimating sensitive population proportion using a combination of binomial and hypergeometric randomized responses by direct and inverse mechanism, Statistics in Transition new series, Vol. 18, No. 2, pp. 193–210.

Holbrook, A. L., Krosnick, J. A., (2010). Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique. Public Opin Quart 74, pp. 37–67.

Imai, K., (2011). Multivariate regression analysis for the item count technique, Journal of the American Statistical Association, Vol. 206, pp. 407–416.

Imai, K., (2015). Design and Analysis of the Randomized response Technique, Journal of the American Statistical Association, Vol. 110, No. 511, pp. 1304–1319.

Kowalczyk, B., Wieczorkowski, R., (2017). Comparing Proportions of sensitive Items in Two Populations when Using Poisson and Negative Binomial Item Count Techniques, Quantitative Methods in Economics, Vol. 18, pp. 68–77.

Krumpal, I., Jann, B., Korndörfer, M., Schmukle, S., (2018). Item Sum Double-List Technique: An Enhanced Design for Asking Quantitative Sensitive Questions, Survey Research Methods, Vol. 12, pp. 91–102.

Kuha, J., Jackson, J., (2014). The item count method for sensitive survey questions: modeling criminal behavior, Journal of the Royal Statistical Society Series C, Vol. 63, pp. 321–341.

Mclachlan, G. J., Krishnan, T., (2008). EM Algorithm and Extensions, Wiley Series in Probability and Statistics.

Miller Jd. (1994). A new survey technique for studying deviant behavior. PhD Thesis, The George Washington University, USA, 1984.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL https://www.R-project.org/.

Tian, G-L., Tang, M-L., Wu., Q, Liu Y., (2017). Poisson and negative binomial item count techniques for surveys with sensitive question, Statistical Methods in Medical Research, Vol. 26, pp. 931–947.

Tourangeau, R., Yan, T., (2007). Sensitive questions in surveys. Psychol Bull 133, pp. 859–883.

Trappman, M., Krumpal, I., Kirchner, A., Jann, B., (2014). Item Sum: A New Technique for Asking Quantitative Sensitive Questions, Journal of Survey Statistics and Methodology, Vol. 2, pp. 58–77.

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

Wolter, F., Laier, B., (2014). The Effectiveness of the Item Count Technique in Eliciting Valid Answers to Sensitive Questions. An Evaluation in the Context of Self-Reported Delinquency, Survey Research Methods, Vol. 8, pp. 153–168.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 89–108, DOI 10.21307/stattrans-2022-006
Received – 30.12.2020; accepted – 23.08.2021

The odd power generalized Weibull-G power series class of distributions: properties and applications

Broderick Oluyede[1], **Thatayaone Moakofi**[2], **Fastel Chipepa**[3]

ABSTRACT

We develop a new class of distributions, namely, the odd power generalized Weibull-G power series (OPGW-GPS) class of distributions. We present some special classes of the proposed distribution. Structural properties, have also been derived. We conducted a simulation study to evaluate the consistency of the maximum likelihood estimates. Moreover, two real data examples on selected data sets, to illustrate the usefulness of the new class of distributions. The proposed model outperforms several non-nested models on selected data sets.

Key words: Weibull-g distribution, power series, Poisson distribution, logarithmic distribution, maximum likelihood estimation.

1. Introduction

Existing distributions or a family of distributions cannot model all real lifetime data. Thus, there is a need to modify them by adding one or more parameters to gain flexibility. Some families of distributions available in the literature include the Weibull-G distribution by Bourguignon et al. (2014), the odd generalized half-logistic Weibull-G family of distributions by Chipepa et al. (2020a), the exponentiated generalized (EG) class of distributions by Cordeiro et al. (2013), beta-G family by Eugene et al. (2002), new power generalized Weibull-G family by Oluyede et al. (2021), the odd exponentiated half-logistic-G family of distributions by Afify et al. (2017), to mention a few.

Several generalized distributions proposed in the literature involving the power series include the exponentiated generalized power series class of distributions by Oluyede et al. (2020c), a new generalized Lindley-Weibull class of distributions by Makubate et al. (2020), the exponentiated power generalized Weibull power series family of distributions by Aldahlan et al. (2019), Weibull-power series distributions by Morais and Barreto-Souza (2011), complementary exponential power series by Flores et al. (2013), complementary extended Weibull-power series by Cordeiro and Silva (2014), Burr XII power series by Silva and Silva and Cordeiro (2015), extended Weibull-power series (EWPS) distribution by Silva et al. (2013).

[1]Botswana International University of Science and Technology, Botswana. E-mail: oluyedeo@biust.ac.bw. ORCID: https://orcid.org/0000-0002-9945-2255.
[2]Botswana International University of Science and Technology, Botswana.
E-mail: thatayaone.moakofi@studentmail.biust.ac.bw. ORCID: https://orcid.org/0000-0002-2676-7694.
[3]Botswana International University of Science and Technology, Botswana. E-mail: fastel.chipepa@gmail.com or chipepaf@biust.ac.bw. ORCID: https://orcid.org/0000-0001-6854-8740.

In this paper, we propose a new class of distributions, namely the odd power generalized Weibull-G power series (OPGW-GPS) class of distributions. An attractive feature about the model is that the extra parameter introduced have the capability to control both the weights at the tails of the density function. Also, the new class of distributions can model different types of failure rate functions that are available in different areas like reliability, engineering and biological studies. The new proposed distribution offers more flexibility in data modelling since the special cases exhibit more non-monotonic shapes for the hazard rate function compared to the other power series reviewed in this paper. Furthermore, the new class of distributions gives birth to more families of distributions by choosing any continuous probability distribution as a baseline distribution $G(x;\underline{\psi})$.

In a recent note, Moakofi et al. (2021) developed the odd power generalized Weibull-G (OPGW-G) family of distributions. The cumulative distribution function (cdf) and probability density function (pdf) of the OPGW-G distribution are given by

$$F(x;\alpha,\beta,\underline{\psi}) = 1-\exp\{1-(1+t)^{\beta}\} \tag{1}$$

and

$$f(x;\alpha,\beta,\underline{\psi}) = \frac{\alpha\beta(1+t)^{\beta-1}\exp\{1-(1+t)^{\beta}\}g(x;\underline{\psi})}{\left(1-G(x;\underline{\psi})\right)^{2}}\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha-1}, \tag{2}$$

respectively, where $t=\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}$, for $\alpha,\beta>0$ and parameter vector $\underline{\psi}$. In this note, we extend the OPGW-G family of distributions by compounding it with the power series distribution.

Let N be a zero truncated discrete random variable having a power series distribution, whose probability mass function (pmf) is given by

$$P(N=n)=\frac{a_n\theta^n}{C(\theta)}, n=1,2,3,..., \tag{3}$$

where $C(\theta)=\sum_{n=1}^{\infty}a_n\theta^n$ is finite, $\theta>0$ and $\{a_n\}_{n\geq 1}$ a sequence of positive real numbers. If we consider $X_{(1)}=\min(X_1,X_2,...,X_N)$, then the cumulative distribution function (cdf) and probability density function (pdf) of $X_{(1)}|N=n$ are defined by

$$F_{X_{(1)}|N=n}(x)=1-\frac{C(\theta S(x;\underline{\psi}))}{C(\theta)}, \tag{4}$$

and

$$f_{X_{(1)}|N=n}(x)=\frac{\theta g(x;\underline{\psi})C'(\theta S(x;\underline{\psi}))}{C(\theta)}, \tag{5}$$

where $S(x;\underline{\psi})$ is the survival function of the baseline distribution and $\underline{\psi}$ is a vector of parameters from the baseline distribution $g(x;\underline{\psi})$. The power series family of distributions includes binomial, Poisson, geometric and logarithmic distributions Johnson et al. (1994).

The rest of the paper is organized as follows: In Section 2, we present the new model and some of the statistical properties. We present some special cases of the proposed class of distributions in Section 3. A simulation study is presented in Section 4 and applications in Section 5, followed by concluding remarks.

2. The Model, Sub-Classes and Properties

In this section, we develop the new model, referred to as the odd power generalized Weibull-G power series (OPGW-GPS) class of distributions. Some statistical properties, including expansion of the density function, hazard rate function, quantile function, subclasses, moments, moment generating function and maximum likelihood estimation of model parameters are derived. Details on the derivations of other statistical properties are given in the Web-Appendix.

2.1. The Model

Using equation (4), the odd power generalized Weibull-G power series (OPGW-GPS) class of distributions denoted by OPGW-GPS(α,β,θ,ψ) has cdf and pdf given by

$$F_{OPGW-GPS}(x) = 1-\frac{C(\theta(\exp\{1-(1+t)^{\beta}\}))}{C(\theta)}, \tag{6}$$

and

$$\begin{aligned} f_{OPGW-GPS}(x) &= \frac{\theta\alpha\beta(1+t)^{\beta-1}g(x;\underline{\psi})}{\left(1-G(x;\underline{\psi})\right)^{2}}\exp\{1-(1+t)^{\beta}\}\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha-1} \\ &\times \frac{C'\left(\theta[\exp\{1-(1+t)^{\beta}\}]\right)}{C(\theta)}, \end{aligned} \tag{7}$$

respectively, where $t=\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}$, for $\alpha,\beta,\theta,x>0$ and parameter vector $\underline{\psi}$.

Table 1 below presents the special families of OPGW-GPS distribution when $C(\theta)$ is specified in equation (6).

Table 1: Special Families of the OPGW-GPS Distribution

Distribution	$C(\theta)$	a_n	cdf
OPGW-G Poisson	$e^{\theta}-1$	$(n!)^{-1}$	1-$\frac{\exp(\theta[\exp\{1-(1+t)^{\beta}\}])-1}{\exp(\theta)-1}$
OPGW-G Geometric	$\theta(1-\theta)^{-1}$	1	1-$\frac{(1-\theta)(\exp\{1-(1+t)^{\beta}\})}{(1-\theta[\exp\{1-(1+t)^{\beta}\}])}$
OPGW-G Logarithmic	$-\log(1-\theta)$	n^{-1}	1-$\frac{\log(\exp\{1-(1+t)^{\beta}\})}{\log(1-\theta)}$
OPGW-G Binomial	$(1+\theta)^m-1$	$\binom{m}{n}$	1-$\frac{(1+\theta[\exp\{1-(1+t)^{\beta}\}])^m-1}{(1+\theta)^m-1}$

2.2. Regularity Condition

We use the Kullback-Leibler distance between densities f_α, for $\alpha_1 \neq \alpha_2$ $D(f_1,f_2)=\int f(x|\alpha_1)\log\left(\frac{f(x|\alpha_1)}{f(x|\alpha_2)}\right)dx>0$. Hence, we obtain

$$\begin{aligned} D(f_1,f_2) &= \int f(x|\alpha_1)\left(\log\left[\frac{\alpha_1}{\alpha_2}\right]+(\alpha_1-\alpha_2)\log\left[\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right]\right)dx \\ &= \int f(x|\alpha_1)\left(\log\left[\frac{\alpha_1[1-G(x;\underline{\psi})]^{(\alpha_1-\alpha_2)}}{\alpha_2[G(x;\underline{\psi})]^{(\alpha_1-\alpha_2)}}\right]\right)dx, \end{aligned} \tag{8}$$

therefore, $D(f_1,f_2)>0$, for $\alpha_1\neq\alpha_2$ since $\log\left[\frac{\alpha_1[1-G(x;\underline{\psi})]^{(\alpha_1-\alpha_2)}}{\alpha_2[G(x;\underline{\psi})]^{(\alpha_1-\alpha_2)}}\right]>0$.

2.3. Quantile Function

Let X be a random variable with cdf defined by equation (6). The quantile function $Q_{OPGW-GPS}(u)$ is defined by $F_{OPGW-GPS}(Q_{OPGW-GPS}(u))=u, 0\leq u\leq 1$ so that the quantile function of the OPGW-GPS class of distributions is given by

$$Q_{OPGW-GPS}(u)=G^{-1}\left[\left(\left[\left(1-\log\left(\frac{C^{-1}\left[C(\theta)(1-u)\right]}{\theta}\right)\right)^{\frac{1}{\beta}}-1\right]^{\frac{-1}{\alpha}}+1\right)^{-1}\right]. \tag{9}$$

2.4. Expansion of Density

The pdf of the OPGW-GPS class of distributions is an infinite linear combination of exponentiated-G distribution expressed as

$$f_{OPGW-GPS}(x) = \sum_{m=0}^{\infty} w_{m+1}g_{m+1}(x;\underline{\psi}), \tag{10}$$

where $g_{m+1}(x;\underline{\psi}) = (m+1)\left(G(x;\underline{\psi})\right)^m g(x;\underline{\psi})$ is the Exp-G distribution with power parameter $(m+1)$ and

$$\begin{aligned} w_{m+1} &= \sum_{j,k,i,l=0}^{\infty}\sum_{n=1}^{\infty}\binom{j}{k}\binom{\beta(k+1)-1}{i}\binom{\alpha(i+1)-1}{l}\binom{-(\alpha(i+1)+1)+l}{m} \\ &\times \frac{\alpha\beta(-1)^{k+l+m}n^{j+1}a_n\theta^n}{C(\theta)j!}\frac{1}{m+1}. \end{aligned} \tag{11}$$

2.5. Moments and Generating Function

If X follows the OPGW-GPS distribution and $Y \sim Exp-G(m+1)$, then using equation (10) the p^{th} raw moment, μ'_p of the OPGW-GPS class of distributions is obtained as

$$\begin{aligned} \mu'_p = E(X^p) &= \int_{-\infty}^{\infty} x^p f(x)dx \\ &= \sum_{m=0}^{\infty} w_{m+1}E(Y^p), \end{aligned}$$

where w_{m+1} is given by equation (11). The moment generating function (MGF) $M(t) = E(e^{tX})$ is given by:

$$M_X(t) = \sum_{m=0}^{\infty} w_{m+1}M_Y(t),$$

where $M_Y(t)$ is the mgf of Y and w_{m+1} is given by equation (11).

2.6. Distribution of Order Statistics

Let $X_1, X_2, ..., X_n$ be a random sample from OPGW-GPS class of distributions and suppose $X_{1:n} < X_{2:n} < ... < X_{n:n}$ denote the corresponding order statistics. The pdf of the k^{th} order statistic is given by

$$f_{k:n}(x) = \frac{n!}{(k-1)!(n-k)!}\sum_{m=0}^{\infty}\sum_{l=0}^{n-k}\binom{n-k}{l}(-1)^l h_{m+1}g_{m+1}(x;\underline{\psi}), \tag{12}$$

where $g_{m+1}(x;\underline{\psi}) = (m+1)g(x;\underline{\psi})G^m(x;\underline{\psi})$ is an Exp-G with power parameter $m+1$ and the linear component

$$\begin{aligned} h_{m+1} &= \sum_{p,j,k,i,v=0}^{\infty}\sum_{n,z=1}^{\infty}\frac{na_n d_{z,p}\theta^{z+n}}{C^{z+1}(\theta)}\frac{(n+z)^j}{j!}\binom{k+l-1}{p}\binom{j}{k}\binom{\beta(k+1)-1}{i} \\ &\times \binom{\alpha(i+1)-1}{v}\binom{-(\alpha(i+1)+1)+v}{m}(-1)^{p+k+m}\alpha\beta\frac{1}{m+1}. \end{aligned} \tag{13}$$

2.7. Rényi Entropy

In this subsection, Rényi entropy for OPGW-GPS class of distributions is derived. An entropy is a measure of uncertainty or variation of a random variable. Rényi entropy by Rényi (1961) is a generalization of Shannon entropy by Shannon (1951). Rényi entropy for OPGW-GPS class of distributions is given by

$$I_R(v) = \frac{1}{1-v}\log\left(\sum_{m=0}^{\infty} w^* e^{(1-v)I_{REG}}\right), \tag{14}$$

where $I_{REG} = \int_0^{\infty}[(1+m/v)g(x;\underline{\psi})G^{m/v}(x;\underline{\psi})]^v dx$ is Rényi entropy for an Exp-G distribution with power parameter $(m/v+1)$ and

$$\begin{aligned} w^* &= \sum_{j,k,i,l,m=0}^{\infty}\sum_{n=1}^{\infty}\frac{d_{v,n}\theta^{v+n-1}}{\left(C(\theta)\right)^v}(v+(n-1))^j(\alpha\beta)^v\binom{j}{k}\binom{\beta(k+v)-v}{i}\frac{(-1)^{k+l+m}}{j!} \\ &\times \binom{\alpha(i+v)-1}{l}\binom{-(\alpha(i+v)+v)+l}{m}\frac{1}{(1+m/v)^v}. \end{aligned} \tag{15}$$

Consequently, Rényi entropy for OPGW-GPS class of distributions can be obtained from Rényi entropy of the Exp-G distribution.

2.8. Maximum Likelihood Estimation

We obtain the maximum likelihood estimates of the parameters of the OPGW-GPS class of distributions in this section. Let $X_i \sim OPGW-GPS(\alpha,\beta,\theta,\underline{\psi})$ and $\Delta = (\alpha,\beta,\theta,\underline{\psi})^T$ be the parameter vector. The log-likelihood $\ell = \ell(\Delta)$ based on a random sample of size n is given by

$$\begin{aligned} \ell(\Delta) &= n\ln[\theta\alpha\beta] + (\beta-1)\sum_{i=1}^{n}\ln[1+t] - n\ln[C(\theta)] + \sum_{i=1}^{n}(1-(1+t)^{\beta}) \\ &+ (\alpha-1)\sum_{i=1}^{n}\ln\left[\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right] + \sum_{i=1}^{n}\ln\left[C'\left(\theta\left[\exp\left(1-[1+t]^{\beta}\right)\right]\right)\right] \\ &+ \sum_{i=1}^{n}\ln\left[g(x;\underline{\psi})\right] - 2\sum_{i=1}^{n}\ln\left[\left(1-G(x;\underline{\psi})\right)^2\right], \end{aligned}$$

where $t = \left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}$. The maximum likelihood estimates of the parameters, denoted by $\hat{\Delta}$ is obtained by solving the nonlinear equation $(\frac{\partial\ell_n}{\partial\alpha}, \frac{\partial\ell_n}{\partial\beta}, \frac{\partial\ell_n}{\partial\theta}, \frac{\partial\ell_n}{\partial\underline{\psi}_k})^T = \mathbf{0}$, using a numerical method such as the Newton-Raphson procedure. The multivariate normal distribution $N_{q+3}(\underline{0}, J(\hat{\Delta})^{-1})$, where the mean vector $\underline{\mathbf{0}} = (0,0,0,\underline{0})^T$ and $J(\hat{\Delta})^{-1}$ is the observed Fisher information matrix evaluated at $\hat{\Delta}$, can be used to construct confidence intervals and confidence regions for the individual model parameters and for the survival and hazard rate functions.

3. Some Special Classes of the OPGW-GPS Class of Distributions

In this section, special classes of OPGW-GPS class of distributions are presented by specifying the baseline distribution to be Weibull and log-logistic distributions, respectively. We considered the power series distributions Poisson and Logarithmic for each selected baseline distribution. The cdf and pdf of the Weibull distribution are given by $G(x;\lambda) = 1-\exp\left(-x^{\lambda}\right)$ and $g(x;\lambda) = \lambda x^{\lambda-1}\exp\left(-x^{\lambda}\right)$, for $\lambda > 0$, and $x > 0$. Furthermore, the log-logistic distribution has cdf and pdf given by $G(x;\lambda) = 1-\left(1+x^{\lambda}\right)^{-1}$ and $g(x;\lambda) = \lambda x^{\lambda-1}\left(1+x^{\lambda}\right)^{-2}$, for $\lambda > 0$, and $x > 0$.

3.1. Odd Power Generalized Weibull-Weibull Poisson (OPGW-WP) Distribution

The cdf and pdf of the OPGW-WP distribution are given by

$$F_{OPGW-WP}(x) = 1-\frac{\exp(\theta[\exp\{1-(1+z)^{\beta}\}]-1)}{\exp(\theta)-1},$$

and

$$\begin{aligned} f_{OPGW-WP}(x) &= \theta\alpha\beta(1+z)^{\beta-1}\left(\frac{\left(1-\exp\{-x^{\lambda}\}\right)}{\exp\{-x^{\lambda}\}}\right)^{\alpha-1}\exp\{1-(1+z)^{\beta}\} \\ &\times \frac{\lambda x^{\lambda-1}\exp\{-x^{\lambda}\}}{\exp\{-x^{\lambda}\}^{2}}\frac{\exp\{\theta[\exp\{1-(1+z)^{\beta}\}]\}}{(\exp(\theta)-1)}, \end{aligned}$$

respectively, where $z = \left(\frac{1-\exp\{-x^{\lambda}\}}{\exp\{-x^{\lambda}\}}\right)^{\alpha}$, for α,β,λ and $\theta > 0$.

Figure 1 shows the plots of pdfs and hrfs of the OPGW-WP distribution. The pdf can take various shapes that include uni-modal, reverse-J, left skewed and right-skewed. Furthermore, the hazard rate functions (hrfs) for the OPGW-WP distribution exhibit increasing, reverse-J, bathtub, and upside bathtub shapes.

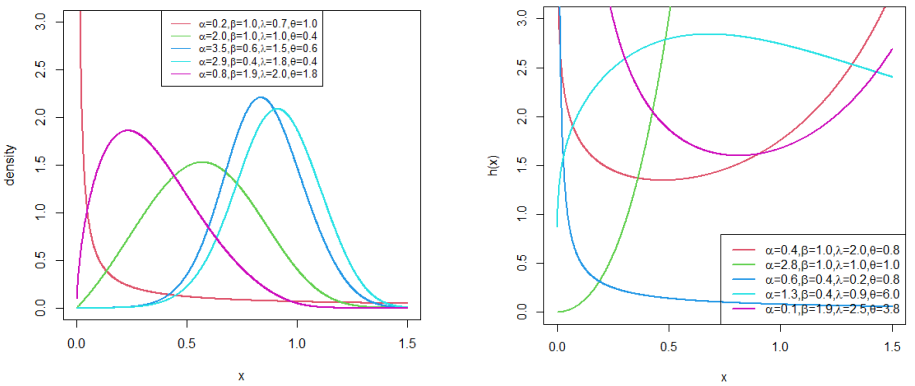


Figure 1: Plots of the pdf and hrf for the OPGW-WP distribution

3.2. Odd Power Generalized Weibull-Weibull Logarithmic (OPGW-WLoG) Distribution

The cdf and pdf of the OPGW-WLoG distribution are given by

$$F_{OPGW-WLoG}(x) = 1 - \frac{\log[1-\theta(\exp\{1-(1+z)^{\beta}\})]}{\log[1-\theta]},$$

and

$$f_{OPGW-WLoG}(x) = \theta\alpha\beta(1+z)^{\beta-1}\left(\frac{1-\exp\{-x^{\lambda}\}}{\exp\{-x^{\lambda}\}}\right)^{\alpha-1}\exp\{1-[1+z]^{\beta}\} \times \frac{\lambda x^{\lambda-1}\exp\{-x^{\lambda}\}}{\exp\{-x^{\lambda}\}^2}\frac{(1-\theta[\exp\{1-(1+z)^{\beta}\}])^{-1}}{-\log[1-\theta]},$$

respectively, for $\alpha, \beta, \lambda > 0$ and $0 < \theta < 1$.

Figure 2 shows the plots of pdfs and hrfs of the OPGW-WLoG distribution. The pdf can take various shapes that include uni-modal, reverse-J, left or right-skewed. Furthermore, the hazard rate functions (hrfs) for the OPGW-WLoG distribution exhibit increasing, reverse-J, bathtub, upside-down bathtub, and upside-down bathtub follwed by bathtub shapes.

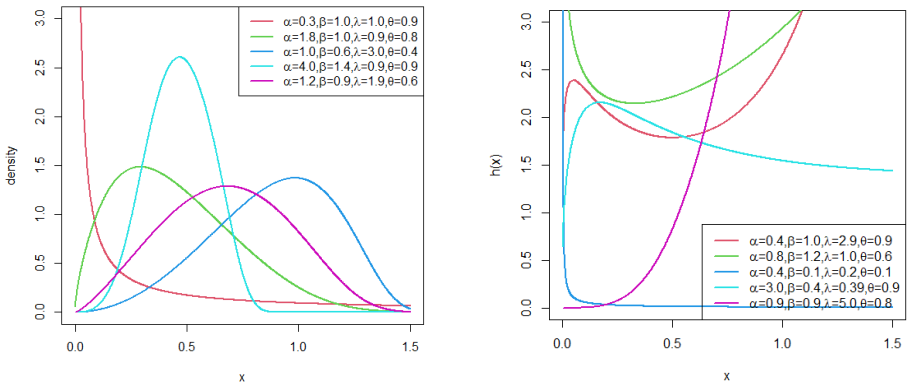


Figure 2: Plots of the pdf and hrf for the OPGW-WLoG distribution

3.3. Odd Power Generalized Weibull-Log-Logistic Poisson (OPGW-WLLoGP) Distribution

The cdf and pdf of the OPGW-LLoGP distribution are given by

$$F_{OPGW-LLoGP}(x) = 1-\frac{\exp\{\theta[\exp\{1-(1+w)^{\beta}\}]\}-1}{\exp(\theta)-1},$$

and

$$\begin{aligned} f_{OPGW-LLoGP}(x) &= \theta\alpha\beta(1+w)^{\beta-1}\left(\frac{1-(1+x^{\lambda})^{-1}}{(1+x^{\lambda})^{-1}}\right)^{\alpha-1}\exp\{1-(1+w)^{\beta}\} \\ &\times \frac{\lambda x^{\lambda-1}(1+x^{\lambda})^{-2}}{(1+x^{\lambda})^{-2}}\frac{\exp\{\theta[\exp\{1-(1+w)^{\beta}\}]\}}{\exp\{\theta\}-1}, \end{aligned}$$

respectively, where $w=\left(\frac{1-(1+x^{\lambda})^{-1}}{(1+x^{\lambda})^{-1}}\right)^{\alpha}$, for α,β,λ and $\theta>0$.

Figure 3 shows the plots of the pdfs and hrfs of the OPGW-LLoGP distribution. The pdf can take various shapes that include almost-symmetric, reverse-J, left or right-skewed. The hazard rate functions (hrfs) for the OPGW-LLoGP distribution exhibit increasing, reverse-J, bathtub and upside-down bathtub shapes.

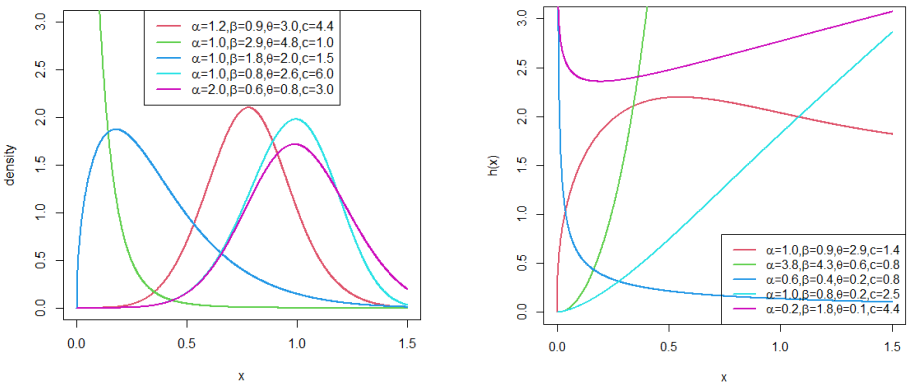


Figure 3: Plots of the pdf and hrf for the OPGW-LLoGP distribution

3.4. Odd Power Generalized Weibull-Log-Logistic Logarithmic (OPGW-LLoGLoG) Distribution

The cdf and pdf of the OPGW-LLoGLoG distribution are given by

$$F_{OPGW-LLoGLoG}(x) = 1-\frac{\log[1-\theta(\exp\{1-(1+w)^{\beta}\})]}{\log[1-\theta]},$$

and

$$\begin{aligned} f_{OPGW-LLoGLoG}(x) &= \theta\alpha\beta(1+w)^{\beta-1}\left(\frac{1-(1+x^{\lambda})^{-1}}{(1+x^{\lambda})^{-1}}\right)^{\alpha-1}\frac{\lambda x^{\lambda-1}(1+x^{\lambda})^{-2}}{(1+x^{\lambda})^{-2}} \\ &\times \exp\{1-(1+w)^{\beta}\}\frac{(1-\theta[\exp\{1-(1+w)^{\beta}\}])^{-1}}{-\log[1-\theta]}, \end{aligned}$$

respectively, for $\alpha,\beta,\lambda > 0$ and $0 < \theta < 1$.

Figure 4 shows the pdfs of the OPGW-LLoGLoG distribution. The pdf can take various shapes that include unimodal, reverse-J, left or right-skewed. Furthermore, the hazard rate functions (hrfs) for the OPGW-LLoGLoG distribution exhibit increasing, reverse-J, bathtub, upside-down bathtub, and upside-down bathtub follwed by bathtub shapes.

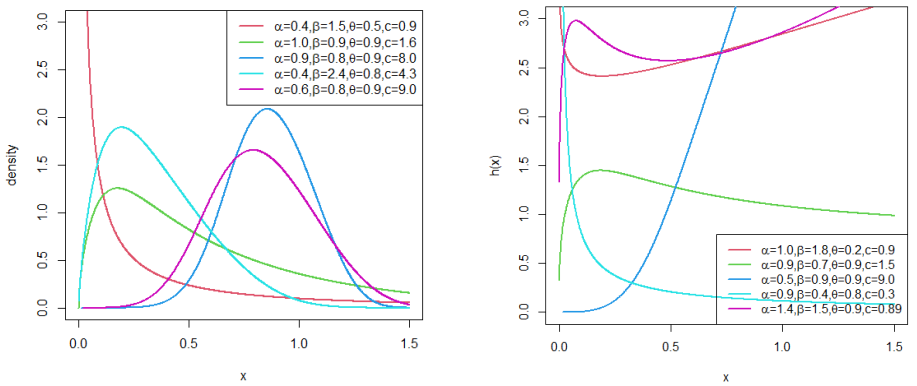


Figure 4: Plots of the pdf and hrf for the OPGW-LLoGLoG distribution

4. Simulation Study

In this section, the performance of the OPGW-WP distribution is examined by conducting various simulations for different sizes (n=25, 50, 100, 200, 400, 800 and 1000) via the R package. We simulate $N = 1000$ samples for the true parameters values given in Table 2. The table lists the mean MLEs of the model parameters along with the respective bias and root mean squared errors (RMSEs). The precision of the MLEs is discussed by means of the following measures: mean, mean square error (MSE) and average bias.

The estimated parameter values in Table 2 indicate that the estimates are quite stable and, more importantly, are close to the true parameter values for these sample sizes. The simulation study shows that the maximum likelihood method is appropriate for estimating the OPGW-WP model parameters. In fact, the means of the parameters tend to be closer to the true parameter values when n increases. The bias and RMSE for the estimated parameter, say, $\hat{\theta}$, are given by:

$$Bias(\hat{\theta}) = \frac{\sum_{i=1}^{N} \hat{\theta}_i}{N} - \theta, \quad \text{and} \quad RMSE(\hat{\theta}) = \sqrt{\frac{\sum_{i=1}^{N} (\hat{\theta}_i - \theta)^2}{N}},$$

respectively.

5. Inference

We present two real data examples in this section, to illustrate the importance of the OPGW-WP distribution. We compared the OPGW-WP distribution to various models. We estimate model parameters using the maximum likelihood estimation technique via the **nlm** package in R Software (2014). Model performance was assessed using the **Adequacy-**

Table 2: Monte Carlo Simulation Results for OPGW-WP Distribution: Mean, RMSE and Average Bias

		$\alpha=0.5,\beta=1.1,\lambda=1.1,c=1.1$			$\alpha=1.0,\beta=1.0,\lambda=1.0,c=0.9$		
	n	Mean	RMSE	Bias	Mean	RMSE	Bias
	25	0.5335	0.9724	0.0335	1.0068	1.4337	0.0068
	50	0.4575	0.4222	-0.0425	0.7489	0.7986	-0.2511
	100	0.4224	0.2242	-0.0776	0.7243	0.5732	-0.2757
α	200	0.4464	0.1718	-0.0536	0.7645	0.4457	-0.2355
	400	0.4505	0.1216	-0.0495	0.8235	0.3418	-0.1765
	800	0.4662	0.0888	-0.0338	0.8991	0.2600	-0.1009
	1000	0.4757	0.0779	-0.0243	0.9316	0.2332	-0.0684
	25	0.8481	0.3930	-0.2519	0.7820	0.3946	-0.2180
	50	0.8847	0.3324	-0.2153	0.8130	0.3100	-0.1870
	100	0.9469	0.2894	-0.1531	0.8612	0.2648	-0.1388
β	200	1.0077	0.2298	-0.0923	0.9268	0.2081	-0.0732
	400	1.0337	0.1653	-0.0663	0.9621	0.1488	-0.0379
	800	1.0636	0.1172	-0.0364	0.9873	0.0951	-0.0127
	1000	1.0737	0.0913	-0.0263	0.9907	0.0798	-0.0093
	25	2.8782	2.0091	1.3782	3.0075	2.9668	2.0075
	50	2.4305	1.4320	0.9305	2.5782	2.2024	1.5782
	100	2.2442	1.1478	0.7442	2.1886	1.7264	1.1886
λ	200	1.9503	0.8102	0.4503	1.7452	1.1217	0.7452
	400	1.8119	0.5772	0.3119	1.4616	0.7597	0.4616
	800	1.6766	0.3879	0.1766	1.2337	0.4634	0.2337
	1000	1.6274	0.3250	0.1274	1.1665	0.3893	0.1665
	25	2.6701	2.6410	1.5701	2.9833	3.9787	2.0833
	50	2.3715	1.8348	1.2715	2.5426	2.1073	1.6426
	100	2.0826	1.7178	0.9826	2.2745	2.0097	1.3745
c	200	1.6754	1.3790	0.5754	1.7600	1.5762	0.8600
	400	1.4741	0.9297	0.3741	1.4197	1.1446	0.5197
	800	1.3180	0.6384	0.2180	1.1350	0.7098	0.2350
	1000	1.2605	0.5092	0.1605	1.0595	0.5971	0.1595
		$\alpha=1.1,\beta=1.5,\lambda=0.9,c=1.1$			$\alpha=1.0,\beta=0.9,\lambda=1.0,c=0.9$		
	25	0.9414	1.3843	-0.1586	0.8803	1.0931	-0.1197
	50	0.8392	0.9418	-0.2608	0.7589	0.8029	-0.2411
	100	0.8497	0.8119	-0.2503	0.7232	0.5468	-0.2768
α	200	0.8751	0.6733	-0.2249	0.7517	0.4154	-0.2483
	400	0.9233	0.5315	-0.1767	0.8211	0.3367	-0.1789
	800	1.0137	0.4518	-0.0863	0.8871	0.2439	-0.1129
	1000	1.0340	0.4206	-0.0660	0.9209	0.2185	-0.0791
	25	1.8679	0.7760	0.3679	0.6926	0.3115	-0.2074
	50	1.7416	0.5930	0.2416	0.7050	0.2879	-0.1950
	100	1.6756	0.4820	0.1756	0.7525	0.2565	-0.1475
β	200	1.6521	0.3916	0.1521	0.7895	0.2225	-0.1105
	400	1.6211	0.3205	0.1211	0.8284	0.1683	-0.0716
	800	1.5767	0.2530	0.0767	0.8686	0.1063	-0.0314
	1000	1.5605	0.2270	0.0605	0.8756	0.0936	-0.0244
	25	2.7571	2.7198	1.8571	2.6099	2.3171	1.6099
	50	2.5146	2.3418	1.6146	2.3455	1.8754	1.3455
	100	2.2513	2.0824	1.3513	1.9934	1.3621	0.9934
λ	200	1.8883	1.6188	0.9883	1.6985	0.9996	0.6985
	400	1.5498	1.2004	0.6498	1.4514	0.7178	0.4514
	800	1.2673	0.8589	0.3673	1.2390	0.4494	0.2390
	1000	1.1892	0.7238	0.2892	1.1732	0.3883	0.1732
	25	2.0189	4.2126	0.9189	2.7413	3.1552	1.8413
	50	1.7098	2.4929	0.6098	2.6147	2.1555	1.7147
	100	1.4663	1.0845	0.3663	2.3388	2.0413	1.4388
c	200	1.2365	0.7964	0.1365	2.0201	1.8711	1.1201
	400	1.1102	0.7889	0.0102	1.6168	1.4018	0.7168
	800	1.0299	0.5098	-0.0701	1.2507	0.8697	0.3507
	1000	1.0471	0.4529	-0.0529	1.1567	0.7870	0.2567

Model package in R software R Software (2014) and the following goodness-of-fit statistics were considered: Cramer-von-Mises (W^*) and Andersen-Darling (A^*), -2loglikelihood (-2 log L), Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (AICC), Bayesian Information Criterion (BIC), Kolmogorov-Smirnov (K-S) statistic (and its p-value), and sum of squares (SS). The model with the smallest values of the goodness-of-fit statistics and a bigger p-value for the K-S statistic is regarded as the best model.

The OPGW-WP distribution was compared to the following models: odd Weibull-Topp-Leone-log-logistic Poisson (OW-TL-LLoGP), odd Weibull-Topp-Leone-log-logistic geometric (OW-TL-LLoGG) and odd Weibull-Topp-Leone-log-logistic logarithmic (OW-TL-LLoGL) by Oluyede et al. (2020b), exponentiated half-logistic-power generalized Weibull-log-logistic (EHL-PGW-LLoG) by Oluyede et al. (2020a), odd exponentiated half-logistic-Burr XII (OEHL-BXII) by Aldahlan and Afify (2018), exponentiated half-logistic odd Weibull-Topp-Leone-log logistic (EHLOW-TL-LLoG) by Chipepa et al. (2020a), odd generalized half-logistic Weibull-Weibull (OGHLW-W) by Chipepa et al. (2020b), odd log-logistic exponentiated Weibull (OLLEW) by Afify et al. (2018), Kumaraswamy odd Lindley-log logistic (KOL-LLoG) by Chipepa et al. (2019) and Kumaraswamy-Weibull (Kw-W) by Cordeiro et al. (2010). The pdfs of the non-nested models are

$$\begin{aligned} f_{OW-TL-LLoGP}(x;\alpha,\lambda,\gamma,\theta) &= \frac{2\theta\gamma\alpha\lambda x^{\lambda-1}(1+x^{\lambda})^{-3}[1-(1+x^{\lambda})^{-2}]^{\gamma\alpha-1}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]^{\alpha+1}} \\ &\times \exp\left\{-\left[\frac{[1-(1+x^{\lambda})^{-2}]^{\gamma}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]}\right]^{\alpha}\right\} \\ &\times \frac{\exp\left(\theta\left(\exp\left\{-\left[\frac{[1-(1+x^{\lambda})^{-2}]^{\gamma}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]}\right]^{\alpha}\right\}\right)\right)}{\exp(\theta)-1}, \end{aligned}$$

for $\alpha,\lambda,\gamma,\theta>0$,

$$\begin{aligned} f_{OW-TL-LLoGG}(x;\alpha,\lambda,\gamma,\theta) &= \frac{2(1-\theta)\gamma\alpha\lambda x^{\lambda-1}(1+x^{\lambda})^{-3}[1-(1+x^{\lambda})^{-2}]^{\gamma\alpha-1}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]^{\alpha+1}} \\ &\times \exp\left\{-\left[\frac{[1-(1+x^{\lambda})^{-2}]^{\gamma}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]}\right]^{\alpha}\right\} \\ &\times \left(1-\left(\theta\left(\exp\left\{-\left[\frac{[1-(1+x^{\lambda})^{-2}]^{\gamma}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]}\right]^{\alpha}\right\}\right)\right)\right)^{-2}, \end{aligned}$$

for $\alpha,\lambda,\gamma>0$ and $0<\theta<1$,

$$\begin{aligned} f_{OW-TL-LLoGL}(x;\alpha,\lambda,\gamma,\theta) &= \frac{2\theta\gamma\alpha\lambda x^{\lambda-1}(1+x^{\lambda})^{-3}[1-(1+x^{\lambda})^{-2}]^{\gamma\alpha-1}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]^{\alpha+1}} \\ &\times \exp\left\{-\left[\frac{[1-(1+x^{\lambda})^{-2}]^{\gamma}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]}\right]^{\alpha}\right\} \\ &\times \frac{\left(1-\left(\theta\left(\exp\left\{-\left[\frac{[1-(1+x^{\lambda})^{-2}]^{\gamma}}{[1-(1-(1+x^{\lambda})^{-2})\gamma]}\right]^{\alpha}\right\}\right)\right)\right)^{-1}}{-\log(1-\theta)}, \end{aligned}$$

for $\alpha,\lambda,\gamma>0$ and $0<\theta<1$,

$$\begin{aligned} f_{EHL-PGW-LLoG}(x;\alpha,\beta,\delta,c) &= 2\alpha\beta\delta\left[1+\left(\frac{1-(1+x^{c})^{-1}}{(1+x^{c})^{-1}}\right)^{\alpha}\right]^{\beta-1} e^{\left(1-\left[1+\left(\frac{1-(1+x^{c})^{-1}}{(1+x^{c})^{-1}}\right)^{\alpha}\right]^{\beta}\right)} \\ &\times \left((1+x^{c})^{-1}\right)^{-(\alpha+3)}\left(1+e^{\left(1-\left[1+\left(\frac{1-(1+x^{c})^{-1}}{(1+x^{c})^{-1}}\right)^{\alpha}\right]^{\beta}\right)}\right)^{-2} \\ &\times \left[\frac{1-e^{\left(1-\left[1+\left(\frac{1-(1+x^{c})^{-1}}{(1+x^{c})^{-1}}\right)^{\alpha}\right]^{\beta}\right)}}{1+e^{\left(1-\left[1+\left(\frac{1-(1+x^{c})^{-1}}{(1+x^{c})^{-1}}\right)^{\alpha}\right]^{\beta}\right)}}\right]^{\delta-1} cx^{c-1}\left(1-(1+x^{c})^{-1}\right)^{\alpha-1}, \end{aligned}$$

for $\alpha,\beta,\delta,c>0$,

$$f_{OEHLBXII}(x;\alpha,\lambda,a,b)=\frac{2\alpha\lambda abx^{a-1}\exp(\lambda[1-(1+x^{a})^{b}])(1-\exp(\lambda[1-(1+x^{a})^{b}]))^{\alpha-1}}{(1+x^{a})^{-b-1}(1+\exp(\lambda[1-(1+x^{a})^{b}]))^{\alpha+1}},$$

for $\alpha,\lambda,a,b>0$,

$$\begin{aligned} f_{EHLOW-TL-BXII}(x;\alpha,\beta,\delta,\lambda,\gamma) &= \frac{4\alpha\beta\delta\lambda\gamma x^{\lambda-1}(1+x^{\lambda})^{-2\gamma-1}[1-(1+x^{\lambda})^{-2\gamma}]^{\alpha\beta-1}}{(1-[1-(1+x^{\lambda})^{-2\gamma}]^{\alpha})^{\beta+1}} \\ &\times \exp(-t)\big(1+\exp(-t)\big)^{-2}\left[\frac{1-\exp(-t)}{1+\exp(-t)}\right]^{\delta-1}, \end{aligned}$$

where $t=\left[\frac{[1-(1+x^{\lambda})^{-2\gamma}]^{\alpha}}{1-[1-(1+x^{\lambda})^{-2\gamma}]^{\alpha}}\right]^{\beta}$, for $\alpha,\beta,\delta,\lambda,\gamma>0$ (We obtain the EHLOW-TL-LLoG distribution from the EHLOW-TL-BXII distribution by setting $\gamma=1$),

$$f_{OGHLW-W}(x;\alpha,\beta,\lambda,\gamma)=\frac{2\alpha\beta\lambda\gamma x^{\gamma-1}e^{-\lambda x^{\gamma}}(1-e^{-\lambda x^{\gamma}})^{\beta-1}\exp\left\{-\alpha\left[\frac{1-e^{-\lambda x^{\gamma}}}{e^{-\lambda x^{\gamma}}}\right]^{\beta}\right\}}{e^{-(\beta+1)\lambda x^{\gamma}}\left(1+\exp\left\{-\alpha\left[\frac{1-e^{-\lambda x^{\gamma}}}{e^{-\lambda x^{\gamma}}}\right]^{\beta}\right\}\right)^{2}},$$

for $\alpha,\beta,\lambda,\gamma>0$,

$$f_{OLLEW}(x;\alpha,\beta,\gamma,\theta)=\frac{\theta\beta\gamma x^{\beta-1}e^{-(x/\alpha)^{\beta}}[1-e^{-(x/\alpha)^{\beta}}]^{\gamma\theta-1}(1-[1-e^{-(x/\alpha)^{\beta}}]^{\gamma})^{\theta-1}}{\alpha\beta([1-e^{-(x/\alpha)^{\beta}}]^{\theta\gamma}+(1-[1-e^{-(x/\alpha)^{\beta}}]^{\gamma})^{\theta})^{2}},$$

for $\alpha,\beta,\lambda,\gamma,\theta>0$,

$$\begin{aligned} f_{KOL-LLoG}(x;a,b,\lambda,c) &= ab\left[\frac{\lambda^2}{(1+\lambda)}\frac{cx^{c-1}}{(1+x^c)^{-1}}\exp(-\lambda z)\right] \\ &\times \left[1-\frac{\lambda+((1+x^c)^{-1})}{(1+\lambda)((1+x^c)^{-1})}\exp(-\lambda z)\right]^{a-1} \\ &\times \left(1-\left[1-\frac{\lambda+((1+x^c)^{-1})}{(1+\lambda)((1+x^c)^{-1})}\exp(-\lambda z)\right]^{a}\right)^{b-1}, \end{aligned}$$

where $z=\frac{(1-(1+x^c)^{-1})}{((1+x^c)^{-1})}$, $a,b,\lambda,c>0$, and

$$f_{Kw-W}(x;a,b,\alpha,\beta)=ab\alpha^{\beta}x^{\beta-1}e^{-(\alpha x)^{\beta}}(1-e^{-(\alpha x)^{\beta}})^{a-1}(1-(1-e^{-(\alpha x)^{\beta}})^{a})^{b-1},$$

for $a,b,\alpha,\beta>0$.

Data analysis results are shown in Tables 3 and 4. A histogram of data, fitted densities and probability plots are shown in Figures 5 and 6.

5.1. Carbon Fibres Data

The data set consists of 66 observations on breaking stress of carbon fibres (Gba). The data set was reported by Nichols and Padgett (2006). The observations are: 3.70, 2.74, 2.73, 2.50, 3.60, 3.11, 3.27, 2.87, 1.47, 3.11, 4.42, 2.41, 3.19, 3.22, 1.69, 3.28, 3.09, 1.87, 3.15, 4.90, 3.75, 2.43, 2.95, 2.97, 3.39, 2.96, 2.53, 2.67, 2.93, 3.22, 3.39, 2.81, 4.20, 3.33, 2.55, 3.31, 3.31, 2.85, 2.56, 3.56, 3.15, 2.35, 2.55, 2.59, 2.38, 2.81, 2.77, 2.17, 2.83, 1.92, 1.41, 3.68, 2.97, 1.36, 0.98, 2.76, 4.91, 3.68, 1.84, 1.59, 3.19, 1.57, 0.81, 5.56, 1.73, 1.59, 2.00, 1.22, 1.12, 1.71, 2.17, 1.17, 5.08, 2.48, 1.18, 3.51, 2.17, 1.69, 1.25, 4.38, 1.84, 0.39, 3.68, 2.48, 0.85, 1.61, 2.79, 4.70, 2.03, 1.80, 1.57, 1.08, 2.03, 1.61, 2.12, 1.89, 2.88, 2.82, 2.05, 3.65.

The estimated variance-covariance matrix is

$$\begin{bmatrix} 0.1944 & -1.0502\times10^{-3} & -0.0340 & -0.1122 \\ -0.0010 & 3.7606\times10^{-5} & -0.0008 & -0.0094 \\ -0.0340 & -8.5370\times10^{-4} & 0.0629 & 0.2165 \\ -0.1122 & -9.4433\times10^{-3} & 0.2165 & 4.2044 \end{bmatrix}$$

and the 95% confidence intervals for the model parameters are given by
$\alpha\in[1.1232\pm0.8643]$, $\beta\in[0.0096\pm0.0120]$, $\lambda\in[2.8341\pm0.4918]$ and $\theta\in[4.3616\pm4.0189]$.

Table 3: MLEs and goodness-of-fit statistics

Model	Estimates				Statistics							
	α	β	λ	θ	$-2\log L$	AIC	$AICC$	BIC	W^*	A^*	K-S	p-value
OPGW-WP	1.1232	0.0096	2.8341	4.3616	282.3	290.3	290.7	300.7	0.0629	0.3926	0.0609	0.8526
	(0.4409)	(0.0061)	(0.2509)	(2.0504)								
	α	λ	γ	θ								
OW-TL-LLoGP	6.3768	0.2383	4.3496	14.3196	282.6	290.6	291.0	301.0	0.0681	0.3966	0.0650	0.7924
	(8.7731)	(0.2573)	(3.4799)	(27.1783)								
OW-TL-LLoGG	2.3977	0.5925	5.4260	3.0075×10^{-13}	282.9	290.9	291.3	301.3	0.0725	0.4300	0.0636	0.8133
	(5.9558)	(1.0870)	(6.9440)	(2.0297)								
OW-TL-LLoGL	3.0041	0.4891	4.6641	1.0180×10^{-10}	282.8	290.8	291.2	301.2	0.0684	0.4186	0.0615	0.8438
	(3.0503)	(0.4251)	(2.8576)	(0.0010)								
	α	β	δ	c								
EHL-PGW-LLoG	1.2499	0.6264	2.6141	1.4037	286.8	294.8	295.2	305.2	0.1568	0.7964	0.1003	0.2664
	(63.3087)	(0.2445)	(0.9515)	(71.1000)								
	α	λ	θ	-								
OEHL-BXII	0.3078	0.0019	11.9671	0.4005	318.6	326.6	327.1	337.1	0.2041	1.4189	0.1301	0.0679
	(0.0616)	(0.0024)	(0.0016)	(0.0666)								
	b	β	δ	c								
EHLOW-TL-LLoG	3.8346	2.3341	1.3504	0.4819	282.4	290.4	290.8	300.8	0.0626	0.3766	0.0618	0.8392
	(5.5094)	(6.3418)	(0.8344)	(1.2822)								
	α	β	λ	γ								
OGHLW-W	2.4257×10^{-5}	0.4640	18.7820	0.2151	287.0	295.0	295.4	305.4	0.0699	0.5971	0.0635	0.8141
	(7.2507×10^{-6})	(4.5353×10^{-3})	(1.1192×10^{-4})	(0.0122)								
	α	β	γ	θ								
OLLEW	3.4848	2.5562	0.6938	1.4692	282.4	290.4	290.8	300.8	0.0659	0.3865	0.0631	0.8208
	(3.1200)	(1.3638)	(1.4331)	(1.5554)								
	a	b	λ	c								
KOL-LLoG	2.1807	8.9816	0.2946	1.1641	282.6	290.6	291.0	301.0	0.0684	0.3994	0.0646	0.7982
	(5.7138)	(75.5774)	(0.5355)	(2.5688)								
	a	b	α	β								
Kw-W	73.5730	3.6270×10^{3}	109.0600	0.1408	282.9	290.9	291.3	301.3	0.0804	0.4446	0.0688	0.7313
	(6.3506)	(0.0017)	(0.8936)	(0.0063)								

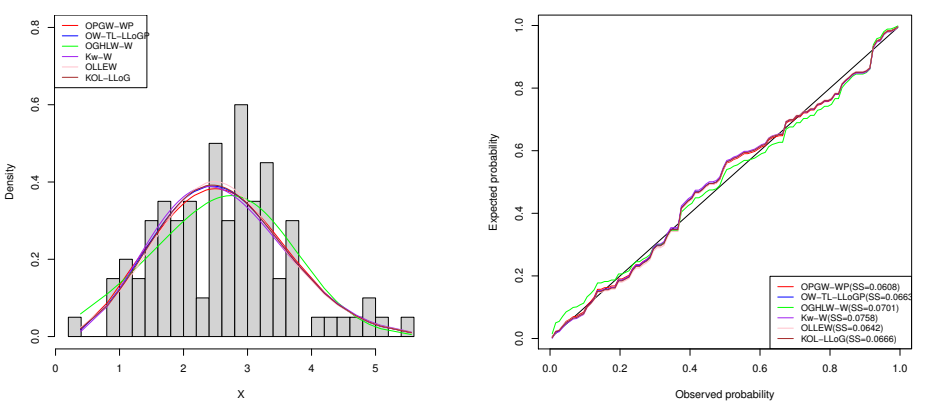


Figure 5: Fitted pdfs and probability plots for carbon fibres data set

Table 3 shows results for the various models fitted for carbon fibres data set. From the given results, we conclude that the OPGW-WP distribution is a good model compared to the selected models since it has the lowest values for the goodness-of-fit statistics: $-2\log L$, AIC, $AICC$, BIC, A^*, W^* and K-S (and the largest p-value for the K-S statistic). Also, from fitted densities and probability plots shown in Figure 5, we observe that the OPGW-WP model fit the data set better than the other models because it has the lowest value for the SS statistic.

5.2. Strengths of 1.5 cm Glass Fibres Data

The second data set represents strengths of 1.5 cm glass fibres. The data set was also analysed by Bourguignon et al. (2014) and Chipepa et al. (2020c). The data are 0.55, 0.93, 1.25, 1.36, 1.49, 1.52, 1.58, 1.61, 1.64, 1.68, 1.73, 1.81, 2.00, 0.74, 1.04, 1.27, 1.39, 1.49, 1.53, 1.59, 1.61, 1.66, 1.68, 1.76, 1.82, 2.01, 0.77, 1.11, 1.28, 1.42, 1.50, 1.54, 1.60, 1.62, 1.66, 1.69, 1.76, 1.84, 2.24, 0.81, 1.13, 1.29, 1.48, 1.50, 1.55, 1.61, 1.62, 1.66, 1.70, 1.77, 1.84, 0.84, 1.24, 1.30, 1.48, 1.51, 1.55, 1.61, 1.63, 1.67, 1.70, 1.78, 1.89.

The estimated variance-covariance matrix is

$$\begin{bmatrix} 0.0318 & -5.4363\times10^{-4} & -0.0540 & -0.0432 \\ -0.0005 & 9.0349\times10^{-5} & -0.0029 & -0.0192 \\ -0.0540 & -2.9252\times10^{-3} & 0.4700 & 0.4033 \\ -0.0432 & -1.9221\times10^{-2} & 0.4033 & 7.5440 \end{bmatrix}$$

and the 95% confidence intervals for the model parameters are given by
$\alpha \in [0.4327 \pm 0.3500]$, $\beta \in [0.0149 \pm 0.0186]$, $\lambda \in [6.6097 \pm 1.3438]$ and $\theta \in [6.02271 \pm 5.3834]$.

Table 4: MLEs and goodness-of-fit statistics

	Estimates				Statistics							
Model	α	β	λ	θ	$-2\log L$	*AIC*	*AICC*	*BIC*	W^*	A^*	K-S	p-value
OPGW-WP	0.43272	0.0149	6.6097	6.0271	25.3	33.3	34.0	41.9	0.1070	0.6061	0.1195	0.3298
	(0.1786)	(0.0095)	(0.6856)	(2.7466)								
	α	λ	γ	θ								
OW-TL-LLoGP	54.4878	0.0638	2.7087	488.4785	30.4	38.4	39.1	47.0	0.2382	1.3088	0.1527	0.1058
	(12.3996)	(0.0176)	(0.0756)	(0.5924)								
OW-TL-LLoGG	4.7219	0.6701	3.5777	1.8173×10^{-9}	31.1	39.1	39.8	47.7	0.2572	1.4103	0.1636	0.0686
	(2.4099)	(0.2953)	(0.6537)	(0.5338)								
OW-TL-LLoGL	4.1499	0.7465	3.7542	5.2222×10^{-8}	31.3	39.3	40.0	47.9	0.2634	1.4437	0.1642	0.0669
	(3.0503)	(0.4251)	(2.8576)	(0.0010)								
	α	β	δ	c								
EHL-PGW-LLoG	2.0377	0.6397	2.1273	1.7395	39.3	47.3	48.0	55.9	0.4178	2.2961	0.2077	0.0087
	(0.2532)	(0.1854)	(0.6324)	(0.2966)								
	α	λ	θ	-								
OEHL-BXII	0.3225	0.0030	11.8172	0.8356	50.3	58.3	59.0	66.9	0.2417	1.3747	0.1423	0.1558
	(0.0670)	(0.0036)	(0.0075)	(0.1347)								
	b	β	δ	c								
EHLOW-TL-LLoG	1.1293	0.1464	4.3716	7.8796	34.9	42.9	43.6	51.4	0.3373	1.8409	0.1868	0.0246
	(0.7335)	(0.0736)	(1.0252)	(3.8531)								
	α	β	λ	γ								
OGHLW-W	3.0734×10^{-5}	0.5007	16.9910	0.4785	27.1	35.1	35.8	43.6	0.1372	0.7816	0.1284	0.2500
	(3.2131×10^{-6})	(2.1967×10^{-9})	(6.4740×10^{-11})	(6.2517×10^{-10})								
	α	β	γ	θ								
OLLEW	1.9920	8.7485	0.3021	1.6872	28.0	36.0	36.7	44.6	0.1864	1.0314	0.1320	0.2223
	(0.2975)	(3.9396)	(0.2668)	(0.7436)								
	a	b	λ	c								
KOL-LLoG	0.5532	25.4210	0.0038	5.9116	27.4	35.4	36.1	44.0	0.1507	0.8450	0.1293	0.2429
	(0.0577)	(6.0680×10^{-6})	(0.0012)	(0.0066)								
	a	b	α	β								
Kw-W	1.04695	641.1439	0.2009	5.5116	30.4	38.4	39.1	47.0	0.2372	1.3038	0.1522	0.1082
	(3.9411)	(0.0539)	(0.0228)	(20.5514)								

Furthermore, from the results shown in Table 4, we conclude that the OPGW-WP distribution is indeed a better model compared to several selected models since it is associated with the lowest values for all the the goodness-of-fit statistics (and the largest p-value for the K-S statistic). We also observe from Figure 6 that the OPGW-WP model fit the data set better than the other models that were considered.

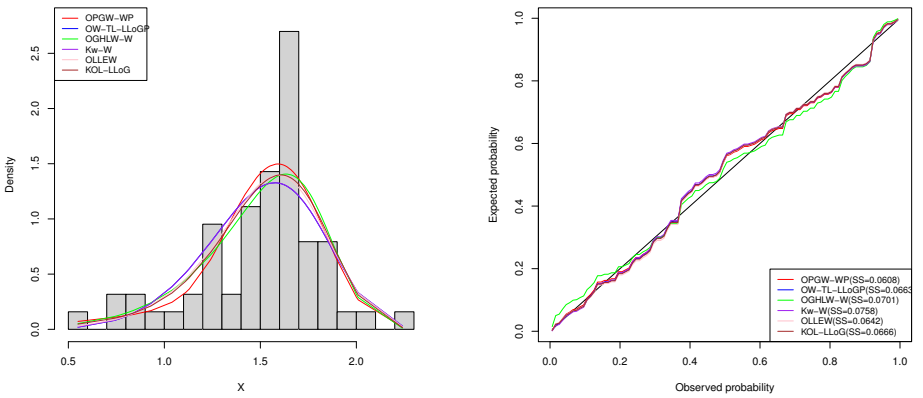


Figure 6: Fitted pdfs and probability plots for glass fibres data set

References

Afify, A. Z., Alizadeh, M., Zayed, M., Ramires, T. G. and Louzada, F., (2018). The Odd Log-Logistic Exponentiated Weibull Distribution: Regression Modelling, Properties, and Applications. *Iranian Journal of Science and Technology*, 42(4), pp. 2273–2288.

Afify, A. Z., Altun, E., Alizadeh, M., Ozel, G., and Hamedani, G. G., (2017). The Odd Exponentiated Half-Logistic-G Family: Properties, Characterizations and Applications. *Chilean Journal of Statistics*, 8(2), pp. 65–91.

Aldahlan, M. and Afify, A. Z., (2018). The odd exponentiated half-logistic Burr XII distribution. *Pakistan Journal of Statistics and Operation Research*, 14(2), pp. 305–317.

Aldahlan, M. A., Jamal, F., Chesneau, C., Elbatal, I., and Elgarhy, M., (2019). Exponentiated Power Generalized Weibull Power Series Family of Distributions: Properties, Estimation and Applications. *PLoS ONE* 15(3): e0230004. https://doi.org/10.1371/journal.pone.0230004.

Bourguignon, M., Silva, R. B. and Cordeiro, G. M., (2014). The Weibull-G Family of Probability Distributions. *Journal of Data Science,* 12, pp.53–68.

Chipepa, F., Oluyede, B. and Wanduku, D., (2020). The Exponentiated Half Logistic Odd Weibull-Topp-Leone-G Family of Distributions: Model, Properties and Applications. *Journal of Statistical Modelling: Theory and Applications,* 2(1), pp. 15–38.

Chipepa, F., Oluyede, B. and Makubate, B., (2020). The Odd Generalized Half-Logistic Weibull-G Family of Distributions: Properties and Applications. *Journal of Statistical Modeling: theory and Applications,* 1(1), 65–89.

Chipepa, F., Oluyede, B. and Makubate, B., (2020). The Topp-Leone-Marshall-Olkin-G Family of Distributions with Applications. *International Journal of Statistics and Probability,* 9(4), pp. 15-32. doi:10.5539/ijsp.v9n4p15.

Chipepa, F., Oluyede, B. and Makubate, B., (2019). A New Generalized Family of Odd Lindley-G Distributions with Application. *International Journal of Statistics and Probability,* 8(6), pp. 1–22. doi:10.5539/ijsp.v8n6p1.

Cordeiro, G. M. and Silva, R. B., (2014). The Complementary Extended Weibull Power Series Class of Distributions. *Ciência e Natura*, 36(3).

Cordeiro, G. M., Ortega, E. M. M., and da Cunha, D. C. C., (2013). The Exponentiated Generalized Class of Distributions. *Journal of Data Science,* 11, pp. 1–27.

Cordeiro, G. M. Ortega, E. M. M. and Nadarajaah, S., (2010). The Kumaraswamy Weibull Distribution with Application to Failure Data. *Journal of the Franklin Institute*, 347, pp. 1399–1429.

Eugene, N., Lee, C., and Famoye, F., (2002). Beta-Normal Distribution and its Applications. *Communications in Statistics: Theory and Methods*, 31, pp. 497–512.

Flores, J., Borges, P., Cancho, V. G., and Louzada, F., (2013). The Complementary Exponential Power Series Distribution. *Brazilian Journal of Probability and Statistics,* 27(4), pp. 565–584.

Gradshetyn, I. S., and Ryzhik, I. M., (2000). Tables of Integrals, Series and Products, Sixth Edition, Academic Press, San Diego.

Jamal, F., Reyad, H. M., Nasir, M. A., Chesneau, C., Shah, M. A. A. and Ahmed S. O., (2019). Topp-Leone Weibull-Lomax Distribution: Properties, Regression Model and Applications, *hal-02270561*.

Johnson, N. L., Kotz, S., and Balakrishnan, N., (1994). Continuous Distributions, Volume 1, John Wiley & Sons, New York, NY.

Makubate, B., Moakofi, T., and Oluyede, B., (2020). A new Generalized Lindley-Weibull Class of Distributions: Theory, Properties and Applications, *Mathematica Slovaka,* 71(1), No. 1, pp. 211–234.

Moakofi, T., Oluyede, B., Chipepa, F and Makubate, B., (2021). Odd Power Generalized Weibull-G Family of Distributions: Properties and Applications, *Journal of Statistical Modelling: Theory and Applications,* 2(1), pp. 121–142.

Morais, A. L. and Barreto-Souza, W., (2011). A Compound Class of Weibull and Power Series Distributions. *Computational Statistics and Data Analysis,* 55(3), pp. 1410-1425.

Nichols, M. D. and Padgett, W. J. A., (2006). A Bootstrap Control Chart for Weibull Percentiles. *Quality and Reliability Engineering International*, 22, pp. 141–151.

Oluyede, B., Moakofi, T., Chipepa, F and Makubate, B., (2021). A New Power Generalized Weibull-G Family of Distributions: Properties and Applications. *Journal of Statistical Modelling: Theory and Applications,* 1(2), pp. 167–191.

Oluyede, B., Chipepa, F. and Wanduku, D., (2020). The Exponentiated Half Logistic-Power Generalized Weibull-G Family of Distributions: Model, Properties and Applications. *Eurasian Bulletin of Mathematics,* 3(3), pp. 134–161.

Oluyede, B., Chipepa, F. and Wanduku, D., (2020). The Odd Weibull-Topp-Leone-G Power Series Family of Distributions: Model, Properties and Applications. *Journal of Nonlinear Sciences and Applications,* 14, pp. 268–286.

Oluyede, B., Fagbamigbe, A., Mashabe, B., Makubate, B., and Wanduku, D., (2020). The Exponentiated Generalized Power Series Family of distributions: Theory Properties and Applications. *Heliyon*, 6(8), e04653.

R Development Core Team, (2014). A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria.

Rényi, A., (1961). On measures of Entropy and Information, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1. The Regents of the University of California.

Shannon, C. E., (1951). Prediction and Entropy of Printed English, *The Bell System Technical Journal*, 30(1), pp. 50–64.

Silva, R. B., Bourguignon, M., Dias, C. R. B. and Cordeiro, G. M., (2013). The Compound Class of Extended Weibull Power Series Distributions. *Computational Statistics and Data Analysis,* 58, pp. 352–367.

Silva, R. B. and Cordeiro, G. M., (2015). The Burr-XII Power Series Distributions: A New Compounding Family. *Brazilian Journal of Probability and Statistics,* 29(3), pp. 565–589.

Appendix

Useful expansions

$$\exp\left(n\left(1-\left[1+\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}\right]^{\beta}\right)\right) = \sum_{j=0}^{\infty}\frac{\left(n\left(1-\left[1+\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}\right]^{\beta}\right)\right)^{j}}{j!},$$

$$\left(1-\left[1+\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}\right]^{\beta}\right)^{j} = \sum_{k=0}^{\infty}\binom{j}{k}(-1)^{k}\left[1+\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}\right]^{\beta k},$$

and

$$\left[1+\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha}\right]^{\beta(k+1)-1} = \sum_{i=0}^{\infty}\binom{\beta(k+1)-1}{i}\left(\frac{G(x;\underline{\psi})}{1-G(x;\underline{\psi})}\right)^{\alpha i}.$$

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1 pp. 109–128, DOI 10.21307/stattrans-2022-007
Received – 05.09.2020; accepted – 28.06.2021

A modified robust confidence interval for the population mean of distribution based on deciles

Moustafa Omar Ahmed Abu-Shawiesh[1], **Juthaphorn Sinsomboonthong**[2], **Bhuiyan Mohammad Golam Kibria**[3]

ABSTRACT

The confidence interval is an important statistical estimator of population location and dispersion parameters. The paper considers a robust modified confidence interval, which is an adjustment of the Student's t confidence interval based on the decile mean and decile standard deviation for estimating the population mean of a skewed distribution. The efficiency of the proposed interval estimator is evaluated on the basis of an extensive Monte Carlo simulation study. The coverage ratio and average width of the proposed confidence interval are compared with certain existing and widely used confidence intervals. The simulation results show that, in general, the proposed interval estimator's performance is highly effective. For illustrative purposes, three real-life data sets are analyzed, which, to a certain extent, support the findings obtained from the simulation study. Thus, we recommend that practitioners use the robust modified confidence interval for estimating the population mean when the data are generated by a normal or skewed distribution.

Key words: robust confidence interval, decile mean, decile mean standard deviation, decile mean standard error, Monte Carlo simulation

1. Introduction

The normality assumption is the basis for many developed statistical theories. One of these theories is the estimation theory for constructing the confidence interval developed by Neyman (1937). However, in real life a lot of the data do not follow normality assumption and data are not mound shaped, rather they are skewed; that is, there is a lack of symmetry of the distribution about the mean. Skewed data may harm our results. Skewness is considered either positive or negative based on the direction

[1] Department of Mathematics, Faculty of Science, The Hashemite University, P. O. Box 330127, Zarqa 13133, Jordan. E-mail: mabushawiesh@hu.edu.jo. ORCID: https://orcid.org/0000-0003-0618-7833.

[2] Department of Statistics, Faculty of Science, Kasetsart University, Bangkok 10900, Thailand. E-mail: fscijps@ku.ac.th. ORCID: https://orcid.org/0000-0002-3375-5982.

[3] Department of Mathematics and Statistics, Florida International University, University Park, Miami FL 33199, USA. E-mail: kibriag@fiu.edu. ORCID: https://orcid.org/0000-0002-6073-1978.

and nature of the distribution. It has long been known that when sampling from a skewed population, with small sample sizes, the usual frequentist confidence intervals for the population mean (μ) have poor coverage properties (Meeden, 1999). The positively skewed data, for example, are common in various fields of modelling such as in psychology (Cain et al., 2016), health science (Baklizi and Kibria, 2009; Banik and Kibria, 2010; Ghosh and Polansky, 2016), environmental science (Mudelsee and Alkio, 2007), biological science (McDonald, 2014), engineering science and others. A confidence interval is an interval estimator that will capture the true parameter value in repeated samples. Abu-Shawiesh et al. (2019) defined confidence interval as a range of values that provides the user with an understanding of how precise the estimates of a parameter are. In practice, it is usual to use normal theory to construct a confidence interval for making inferences about the population mean (μ). Unfortunately, the confidence interval based on this theory suffers when samples come from skewed or non-normal populations. Therefore, it is important to construct a confidence interval of a population mean (μ) that is not limited by the assumption of population normality (Miller and Penfield, 2005). Several other methods have been described in the literature, such as transformation methods and bootstrap methods, to obtain an acceptable coverage rate and small interval width with skewed distribution and small sample sizes (Meeden, 1999; Shi and Kibria, 2007; Ghosh and Polansky, 2016). There are various methods in the literature in which confidence intervals are obtained for the population mean (μ). In practice, it is often possible to work with smaller sample sizes. In such cases, Student's t confidence interval can be preferred instead of the classical confidence interval, but it requires an assumption of normality. Luh and Guo (2001) argued that "since violation of the normality assumption may be fairly common in applied research, robust and efficient alternatives to deal with the problem are needed". Therefore, it is essential to use robust estimators which are less affected by outliers or small departures from the model assumptions (Sindhumol et al., 2016). Johnson (1978) proposed a modification of the Student's t confidence interval for skewed distributions. Since Johnson (1978), many researchers have obtained confidence intervals for population mean of a skewed distribution (Chen, 1995; Meeden, 1999; Kibria, 2006; Shi and Kibria, 2007; Baklizi, 2008; Abu-Shawiesh et al., 2009; Baklizi and Kibria, 2009; Abu-Shawiesh et al., 2011; Pek et al., 2017; Abu-Shawiesh et al., 2018; Abu-Shawiesh and Saghir, 2019; Akyuz and Abu-Shawiesh, 2020; Sinsomboonthong et al., 2020).

In this paper, we compare various methods for constructing a confidence interval for the population mean (μ) when data are normally or non-normally distributed and propose a new robust confidence interval. This proposed confidence interval is an adjustment of the Student's t confidence interval based on the decile mean and the decile mean standard deviation. Since a theoretical comparison is not possible, we investigate the performance of the proposed confidence interval by using a Monte Carlo simulation study and its implementation with three real-life data sets.

2. The decile mean (DM) and the decile mean standard deviation (SD_{DM})

The sample mean ($\bar{x}$) and sample standard deviation (s) are the most popular and frequently used classical estimators of the location and scale parameters of a probability distribution. However, they are unreliable in the presence of skewed distributions. In this paper, we estimate them with well-known and simple robust estimators of location and scale. They are the decile mean (DM) for location and the decile mean standard deviation (SD_{DM}) for scale. Furthermore, the standard error of the decile mean standard deviation (SD_{DM}) is defined.

2.1. The Deciles (D_m)

The central tendency of a data set is a measure of the location or most typical value of the data set. There are various types of descriptive statistics, such as sample mean, sample median and sample trimmed mean that can be chosen as a measure of the central tendency; under a well-behaved normal distribution, they possess some desirable properties. But there is evidence that they may perform poorly and not as well as expected in the presence of skewed distributions. Rana et al. (2012) proposed a new measure of central tendency based on deciles called the decile mean (DM). This measure is fairly robust as it automatically discards extreme observations or outliers from both tails but at the same time is more informative than the sample median in every respect. Let $X_1, X_2, \dots, X_n$ be independent identically distributed (iid) observations from a given population with mean (μ) and standard deviation (σ); then the deciles, which are a measure of position, are the values (nine in number) of the variable that divide any ordered data set $X_{(1)}, X_{(2)}, \dots, X_{(n)}$ into ten equal parts so that each part represents $\frac{1}{10}$ of the sample or population, and are denoted by $D_1, D_2, \dots, D_9$. The fifth decile (D_5) is equal to the sample median (MD). The deciles determine the values for 10%, 20% … and 90% of the data set. Now assume that $x_1, x_2, \dots, x_n$ be the sample observations of the random sample $X_1, X_2, \dots, X_n$, then the deciles can be calculated as follows:

(1) Order the observations $x_1, x_2, \dots, x_n$ according to the magnitude of the values to get the ordered data set $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.

(2) To find the value of the m^{th} sample decile where $m = 1, 2, \dots, 9$, the following simple formula can be used:

$$D_m = x_{\left(\left[\frac{m\,(n+1)}{10}\right]\right)} \text{ observation} \tag{1}$$

where n is the total number (sample size) of observations.

2.2. The Decile Mean (DM)

The decile mean, denoted by DM for the random sample $X_1, X_2, \ldots, X_n$, can be calculated by summing all the deciles $D_1, D_2, \ldots, D_9$ and dividing the sum by the number of deciles. Thus, the formula to find the decile mean (DM) from 9 deciles is as follows:

$$DM = \frac{\sum_{i=1}^{9} D_i}{9} = \frac{D_1 + D_2 + \cdots + D_9}{9} \tag{2}$$

The main advantage of the decile mean (DM) is that it is less sensitive to extreme values than any other existing measures; also, it depends on 80% of a sample. Let X distributed like $X_1, X_2, \ldots, X_n$. Then D_m is a number for which $P(X < D_m) \leq \frac{m}{10} \leq P(X \leq D_m)$. If X has absolutely continuous distribution function $F(x) = P(X \leq x)$ then $F(D_m) = P(X \leq D_m) = \frac{m}{10}$. It is termed a robust estimator in this regard. Rana et al. (2012) used the bootstrap method to investigate the sampling distribution of the newly proposed decile mean (DM) with three other popular and commonly used measures of location, i.e., the sample mean, median and trimmed mean, and found that the newly proposed decile mean (DM) has the following properties:

(i) The distribution of the sample decile is quite normal in shape and irrespective of the presence of outliers.
(ii) The bias and standard error of sample decile mean (DM) are very small, and among the four compared estimators, this estimator appears to be the best in every respect.
(iii) The results presented show that all four estimators are biased, but this bias is the least for the sample decile mean (DM).

Both the bootstrap and simulation study demonstrate that the sample decile mean (DM) is a more accurate measure of central tendency or location in terms of possessing smaller bias and lower standard errors in a variety of situations, and hence can be recommended to be used as an effective measure of central tendency or location.

2.3. Decile Mean Standard Deviation (SD_{DM}) and Standard Error (SE_{DM})

Decile mean standard deviation (SD_{DM}) is a robust measure of dispersion proposed by Doullah (2018) as an alternative to the sample standard deviation (S). Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a given population with mean (μ) and standard deviation (σ); then the decile mean standard deviation (SD_{DM}) can be calculated by using the following formula:

$$SD_{DM} = \sqrt{\frac{\text{The sum of the 9 deciles of } (X_i - DM)^2}{9-1}} = \sqrt{\frac{1}{8}\sum_{i=1}^{9}(D_i - D_m)^2} \tag{3}$$

Doullah (2018) also defined the standard error of the decile mean standard deviation (SD_{DM}), denoted by SE_{DM} , to be computed as follows:

$$SE_{DM} = \frac{SD_{DM}}{\sqrt{n}} \tag{4}$$

3. Methods for estimation the confidence interval for the population mean

In this section, the used methods and the proposed robust method for confidence interval of the population mean (μ) for normal and non-normal distributions are introduced. Let $X_1, X_2, \ldots, X_n$ be iid random sample of size n from a population with mean (μ) and standard deviation (σ). Our purpose is to find an interval estimate for the population mean (μ) with a specific level of confidence. Several methods have been suggested in the literature to find the confidence interval for μ. These are (a) the parametric approach, (b) the modified t approach, (c) the nonparametric approach and (d) the bootstrap approach, among others. In this study, we concentrate on (a) and (b) approaches only. The (1 – α) 100% confidence intervals for the population mean (μ) by different approaches are presented below.

3.1. The Parametric t-Approach

The parametric method to construct the (1 – α) 100% confidence interval for the population mean (μ) is the most used approach because it is well understood, simple and widely used to construct such the confidence interval. Under this approach, we consider two confidence interval methods. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a normal distribution with mean (μ) and variance (σ^2); that is, $X_1, X_2, \ldots, X_n \sim N(\mu,\ \sigma^2)$. Then, the (1 – α) 100% confidence interval for the population mean (μ) given by Student (1908) and known as the Student's t confidence interval for a small sample size n ($n \leq 30$) and unknown population standard deviation (σ) can be constructed as follows:

$$C.I. = \bar{X} \pm t_{\left(\frac{\alpha}{2},\ n-1\right)} \frac{S}{\sqrt{n}} \tag{5}$$

where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$, $S = \sqrt{(n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ and $t_{(\alpha/2,n-1)}$ is the upper $\alpha/2$ percentage point of the Student's t-distribution with $(n-1)$ degrees of freedom. Now, since the Student's t confidence interval depends on the normality assumption, it may not be the best confidence interval and may not perform as well as expected in the presence of skewed distributions. DiCicco and Efron (1996) and Boos and Hughes-Oliver (2000) stated that the Student's t confidence interval is not very robust and can be quite inaccurate in practice for non-normal data.

3.2. The Modified Parametric t-Approach

If the iid random sample $X_1, X_2, \ldots, X_n$ is from a non-normal distribution, the distribution of the t-statistic is not a Student's t distribution. In particular, the skewness of a non-normal distribution has a large impact on the validity of the Student's t-distribution; see, for example, Yanagihara and Yuan (2005). Several methods for constructing the (1 – α) 100% confidence interval for the population mean (μ) have been proposed to remove the effect of skewness by modifying the t-statistic. Here, we briefly review the most important of these methods.

3.2.1. The Johnson t-Approach

Based on the first term of the inverse Cornish–Fisher expansion, Johnson (1978) proposed the following confidence interval estimator for the population mean (μ):

$$C.I. = \left[\bar{X} + \frac{\hat{\mu}_3}{6nS^2}\right] \pm t_{\left(\frac{\alpha}{2},n-1\right)} \frac{S}{\sqrt{n}} \tag{6}$$

where $\hat{\mu}_3 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^3}{n}$ is the estimator of the third central moment of the population (μ_3). Kibria (2006) concluded it appears that the width of the Student's t and Johnson-t confidence intervals is the same.

3.2.2. The Chen t-Approach

Using the Edgeworth expansion, Chen (1995) modified the CLT approach and proposed the following confidence interval estimator for the population mean (μ):

$$C.I. = \bar{X} \pm \left[t_{\left(\frac{\alpha}{2},n-1\right)} + \frac{\hat{\gamma}\left(1+2t^2_{\left(\frac{\alpha}{2},n-1\right)}\right)}{6\sqrt{n}} + \frac{\hat{\gamma}^2\left(t_{\left(\frac{\alpha}{2},n-1\right)} + 2t^2_{\left(\frac{\alpha}{2},n-1\right)}\right)}{9\,n} \right] \frac{S}{\sqrt{n}} \tag{7}$$

where $\hat{\gamma} = \frac{\hat{\mu}_3}{S^3}$ is the estimate of the coefficient of skewness.

3.2.3. The Yanagihara and Yuan t-Approach

To reduce the effect of the mean bias as well as population skewness, Yanagihara and Yuan (2005) proposed the following confidence interval estimator for the population mean (μ):

$$C.I. = \left[\bar{X} + \frac{(S\,\hat{k}_3)}{\left((4n)\left(2+\frac{15}{n}\right)\right)}\right] \pm t_{\left(\frac{\alpha}{2},n-1\right)} \frac{S}{\sqrt{n}} \tag{8}$$

where $\hat{k}_3 = \frac{\left(\sum_{i=1}^{n}(X_i-\bar{X})^3/n\right)}{\left(\sum_{i=1}^{n}(X_i-\bar{X})^2/n\right)^{3/2}}$.

3.2.4. The Shi and Kibria Mad t-Approach

In terms of using the sample median (MD) rather than the sample mean ($\bar{X}$) for defining the sample standard deviation, Shi and Kibria (2007) proposed another confidence interval estimator for the population mean (μ), as follows:

$$C.I. = \bar{X} \pm t_{(\frac{\alpha}{2}, n-1)} \frac{\tilde{S}_2}{\sqrt{n}} \tag{9}$$

where $\tilde{S}_2 = \frac{1}{n} \sum_{i=1}^{n} |X_i - \bar{X}|$ is the sample mean absolute deviation (Mad).

3.2.5. The Abu-Shawiesh, Banik and Kibria AADM t-Approach

Abu-Shawiesh et al. (2018) proposed a modification of the Student's t confidence interval for the population mean (μ) of a skewed distribution, called AADM-t confidence interval estimator and expressed as follows:

$$C.I. = \bar{X} \pm t_{(\frac{\alpha}{2}, n-1)} \frac{AADM}{\sqrt{n}} \tag{10}$$

where $AADM = \frac{\sqrt{\pi/2}}{n} \sum_{i=1}^{n} |X_i - MD|$ is the average absolute deviation from the sample median (Gastwirth, 1982). Gastwirth (1982) stated that the AADM is an asymptotically normally distributed, consistent estimator of the population standard deviation (σ) and almost surely converges to it.

3.3. The Confidence Interval Based on Resampling Approach

Efron and Tibshirani (1993) recommended resampling approach to generate a large number of independent bootstrap samples $x^{*1}, x^{*2}, \ldots, x^{*B}$ for a random sample from an unknown distribution with population mean. A bootstrap sample $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ is obtained by randomly resampling n times with replacement from the original data sample $x_1, x_2, \ldots, x_n$. Then, the (1 – α) 100% confidence interval based on bootstrap percentile for the population mean (μ) can be constructed as follows: this approach performs resampling technique B times and let $\hat{\mu}^{*1}, \hat{\mu}^{*2}, \ldots, \hat{\mu}^{*B}$ be the estimator of parameter μ for each independent bootstrap sample $x^{*1}, x^{*2}, \ldots, x^{*B}$. If $\hat{\mu}^*$ is a random variable drawn from the normal distribution with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$, then the (1 – α) 100% bootstrap percentile confidence interval estimator for the population mean (μ) can be expressed in the form of equation (11) and (12) as follows:

$$CL = 100 \cdot {}^{\alpha}/_{2}{}^{th} \; percentile \; of \; \hat{\mu}^{*}'s \; distribution \tag{11}$$

$$UCL = 100 \cdot (1 - {}^{\alpha}/_{2})^{th} \; percentile \; of \; \hat{\mu}^{*}'s \; distribution \tag{12}$$

The bootstrap confidence interval produce a good coverage ratio for interval estimation as shown in the study by DiCiccio and Efron (1996), Marinho et al. (2018), Ghosh and Polansky (2016).

3.4. The Proposed Robust $DMSD_{DM}$ t-Approach

In this section, we propose a robust modification of the Student's t confidence interval for the population mean (μ) of a skewed population. It is a simple adjustment of the Student's t confidence interval and can be obtained with the following steps:

Step 1: Select a random sample of size (n), $X_1, X_2, \ldots, X_n$, from the probability distribution of the random variable X.

Step 2: Calculate the sample decile mean (DM), which is given by equation (2).

Step 3: Calculate the decile mean standard deviation (SD_{DM}) and the standard error of the decile mean standard deviation (SE_{DM}), which are given by equations (3) and (4).

Step 4: The lower confidence limit (LCL) and the upper confidence limit (UCL) for the $(1-\alpha)100\%$ proposed robust confidence interval estimator–$DMSD_{DM}$-t confidence interval–of the population mean (μ) for the skewed distribution can be calculated as follows:

$$LCL = DM - t_{(\frac{\alpha}{2},n-1)} \frac{SD_{DM}}{\sqrt{n}} \tag{13}$$

$$UCL = DM + t_{(\frac{\alpha}{2},n-1)} \frac{SD_{DM}}{\sqrt{n}} \tag{14}$$

where $\boldsymbol{t}_{(\frac{\alpha}{2},\boldsymbol{n-1})}$ is the upper α/2 percentage point of the Student's t-distribution with $(\boldsymbol{n-1})$ degrees of freedom.

4. The simulation study

Since a theoretical comparison among these confidence intervals is not possible, a simulation study is conducted. All the simulation results are performed by SAS programming version 9.4.

4.1. Performance Evaluation

A Monte Carlo simulation study is presented in this section to compare the performance of eight confidence interval estimators for the population mean of three distributions. We consider a set of possible useful confidence intervals and compare them with the proposed robust method, aiming to confirm that it is appropriate for estimating the population mean (μ) of a skewed distribution. To make comparisons

among confidence intervals, the coverage ratio (CR) and average width (AW) of the confidence intervals are considered as the performance criteria. A smaller width indicates a better confidence interval when the coverage ratios are the same level. Further, the higher coverage ratio indicates a better confidence interval when the widths of intervals are the same level. The sample sizes of n = 10, 20, 30, 40, 50 and 100 were randomly generated 100,000 times. For each set of samples, 95% confidence intervals were constructed for the considered methods and the construction of bootstrap percentile confidence intervals for the population mean are generated resampling 1,000 times for each situation. The coverage ratio (CR) and the average width (AW) of the confidence intervals are obtained using the following two formulas:

$$CR = \frac{\#(L \le \theta \le U)}{100{,}000} \quad \text{and} \quad AW = \frac{\sum_{i=1}^{100{,}000} (U_i - L_i)}{100{,}000} \tag{15}$$

4.2. Probability Distributions for the Simulation Study

To study the effect of skewness and compare the performance of the eight confidence interval estimators for the population mean (μ) of the distribution, two cases for the simulation observations, namely normal and skewed distributions, are considered in this study.

Case (a): Normal Distribution

The normal distribution is symmetric and has no skewness. The probability density function (pdf) of a normal distribution with mean μ and standard deviation σ, $N(\mu, \sigma^2)$, is given as follows:

$$f(x;\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} ;\ -\infty < x < \infty\,,\ -\infty < \mu < \infty\,,\ \sigma > 0 \tag{16}$$

In the simulation algorithm of this study, the population mean μ and the population standard deviation σ are set as $\mu = 20$ and $\sigma = 5, 10, 20$.

Case (b): Skewed Distributions

The skewness of a probability distribution refers to the departure of the distribution from symmetry. A distribution with longer tail on the left is negative skewed, and a distribution with longer tail on the right is positive skewed (Sharma et al., 2009). For skewed distributions, we simulate observations from two probability distributions with varying degrees of skewness as follows:

(i) The chi-square distribution, $\chi^2_{(k)}$, where k is the number of degrees of freedom with probability density function (pdf), is given as follows:

$$f(x;\, k) = \begin{cases} \frac{1}{\Gamma(k/2)\, 2^{k-1}}\, x^{(k/2)-1}\, e^{-x/2} & ,\quad x > 0 \\ 0 & ,\quad otherwise \end{cases} \tag{17}$$

The mean and the variance of the chi-square distribution are given by $\mu = k$ and $\sigma^2 = 2k$. The coefficient of skewness of the distribution is $\sqrt{8/k}$. In the simulation algorithm of this study, the parameter k for the chi-square distribution is set as k = 5, 10, 50.

(ii) The triangular distribution, $Tr(a, b, c)$, involves parameters a, b and c, where a is the minimum value, b is the maximum value and c is the most likely value (mode). The triangular distribution is selected for this study as it can be used to model both positive and negative skewed distributions. The probability density function (pdf) for the triangular distribution is given as follows:

$$f(x; a, b, c) = \begin{cases} 0 & , \quad x < a \\ \frac{2(x-a)}{(b-a)(c-a)} & , \quad a \le x \le c \\ \frac{2(b-x)}{(b-a)(b-c)} & , \quad c < x \le b \\ 0 & , \quad x > b \end{cases} \tag{18}$$

The mean and variance of the triangular distribution, $Tr(a, b, c)$, are given by $\mu = \frac{a+b+c}{3}$ and $\sigma^2 = \frac{a^2+b^2+c^2-ab-ac-bc}{18}$. The skewness coefficient of the triangular distribution is given by $\frac{\sqrt{2}(a+b-2c)(2a-b-c)(a-2b+c)}{5(a^2+b^2+c^2-ab-ac-bc)^{3/2}}$. In the simulation algorithm of this study, we simulate observations from $Tr(0, 1, 0.05)$, $Tr(0, 1, 0.5)$ and $Tr(0, 1, 0.95)$ to represent the positive, symmetric and negative cases of the triangular distribution, respectively. Table 4.1 shows the specific distributions and their skewness coefficients used in this simulation study.

Table 4.1. Coefficients of skewness for the studied simulation probability distributions

Probability Distributions	Parameters	Coefficients of Skewness
$N(\mu, \sigma^2)$	$\mu = 20, \sigma = 5, 10, 20$	0
$\chi^2_{(k)}$	$k = 5$	1.2649
	$k = 10$	0.8944
	$k = 50$	0.4000
$Tr(0, 1, c)$	$c = 0.05$	0.5607
	$c = 0.50$	0
	$c = 0.95$	-0.5607

4.3. The Simulation Study Results

The simulation results for all studied cases are shown in Tables 4.2 to 4.10. The performance of 95% confidence intervals of the population mean for the eight methods are as follows: in the case of normally distributed data as shown in Tables 4.2 to 4.4, it is observed that the coverage ratio of $DMSD_{DM}$-t confidence interval is slightly under

0.95 for all sample sizes. However, the coverage ratio of Bootstrap-percentile confidence interval is only slightly under 0.95 for a small sample size, but it close to 0.95 for the large sample sizes. In addition, the coverage ratios of five intervals–Student's t, Johnson-t, Chen-t, YY-t and AADM-t–are close to 0.95 for all sample sizes. Further, the coverage ratio of Mad-t confidence interval is more under than the nominal level when compared with the proposed interval. When the performance of confidence intervals is compared in terms of the average width, the five methods in which the coverage ratio is close to 0.95 tend to have no difference in average width for any sample size or any of the normal distributed data. Although, the coverage ratio of the proposed method is slightly lower than that of the five methods, the average width of this proposed interval is smaller than that of the five intervals for all sample sizes and it is smaller than the average width of Bootstrap-percentile for the large sample size.

Table 4.2. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of normal distribution with $\mu = 20$ and $\sigma = 5$

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9506	0.9506	0.9477	0.9505	0.8855	0.9438	0.9260	0.9013
	AW	7.0	7.0	7.1	7.0	5.4	6.8	6.2	5.7
20	CR	0.9496	0.9495	0.9482	0.9495	0.8826	0.9459	0.9174	0.9266
	AW	4.6	4.6	4.6	4.6	3.6	4.6	4.1	4.2
30	CR	0.9494	0.9495	0.9489	0.9495	0.8819	0.9464	0.9142	0.9345
	AW	3.7	3.7	3.7	3.7	2.9	3.7	3.3	3.5
40	CR	0.9499	0.9499	0.9493	0.9499	0.8820	0.9481	0.9138	0.9382
	AW	3.2	3.2	3.2	3.2	2.5	3.2	2.8	3.0
50	CR	0.9501	0.9502	0.9499	0.9501	0.8818	0.9483	0.9129	0.9409
	AW	2.8	2.8	2.8	2.8	2.2	2.8	2.5	2.7
100	CR	0.9501	0.9501	0.9501	0.9501	0.8818	0.9490	0.9119	0.9449
	AW	2.0	2.0	2.0	2.0	1.6	2.0	1.8	1.9

Table 4.3. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of normal distribution with $\mu = 20$ and $\sigma = 10$

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9506	0.9506	0.9477	0.9505	0.8855	0.9437	0.9260	0.9013
	AW	13.9	13.9	14.1	13.9	10.8	13.6	12.5	11.4
20	CR	0.9496	0.9495	0.9482	0.9495	0.8826	0.9458	0.9173	0.9266
	AW	9.2	9.2	9.3	9.2	7.3	9.1	8.2	8.4
30	CR	0.9494	0.9494	0.9489	0.9495	0.8820	0.9464	0.9142	0.9345
	AW	7.4	7.4	7.4	7.4	5.9	7.3	6.6	7.0
40	CR	0.9499	0.9500	0.9493	0.9499	0.8821	0.9481	0.9138	0.9381
	AW	6.4	6.4	6.4	6.4	5.0	6.3	5.6	6.1
50	CR	0.9501	0.9502	0.9499	0.9501	0.8819	0.9483	0.9129	0.9409
	AW	5.7	5.7	5.7	5.7	4.5	5.6	5.0	5.5
100	CR	0.9501	0.9501	0.9500	0.9501	0.8818	0.9490	0.9119	0.9449
	AW	4.0	4.0	4.0	4.0	3.1	3.9	3.5	3.9

Table 4.4. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of normal distribution with $\mu = 20$ and $\sigma = 20$

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	Student-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9506	0.9506	0.9477	0.9506	0.8856	0.9437	0.9260	0.9013
	AW	27.8	27.8	28.3	27.8	21.7	27.1	25.0	22.8
20	CR	0.9496	0.9495	0.9482	0.9495	0.8826	0.9458	0.9173	0.9266
	AW	18.5	18.5	18.6	18.5	14.6	18.2	16.4	16.8
30	CR	0.9494	0.9494	0.9489	0.9495	0.8819	0.9464	0.9142	0.9345
	AW	14.8	14.8	14.9	14.8	11.7	14.7	13.1	13.9
40	CR	0.9499	0.9500	0.9493	0.9499	0.8821	0.9481	0.9138	0.9382
	AW	12.7	12.7	12.7	12.7	10.1	12.6	11.3	12.1
50	CR	0.9501	0.9502	0.9499	0.9501	0.8818	0.9483	0.9129	0.9409
	AW	11.3	11.3	11.3	11.3	9.0	11.2	10.0	10.9
100	CR	0.9501	0.9502	0.9500	0.9501	0.8818	0.9490	0.9119	0.9449
	AW	7.9	7.9	7.9	7.9	6.3	7.9	7.0	7.8

In the case of data are generated from two skewed probability distributions–chi-square and triangular distributions–with varying degrees of skewness, the performance of 95% confidence intervals of the population means are shown in Tables 4.5 to 4.10. If the coefficient of skewness for chi-square distribution is equal to 1.2649 or 0.8944, then the coverage ratio of $DMSD_{DM}$-t tends to decrease when the sample size increases. However, if the coefficient of skewness for this distribution is equal to 0.4000, then the coverage ratio of $DMSD_{DM}$-t is slightly under 0.95, and it tends to be at the same level irrespective of the sample size. Moreover, the coverage ratio of five intervals – Student's t, Johnson-t, Chen-t, YY-t and AADM-t – is close to the specified confidence coefficient level, while that of the Mad-t confidence interval is more under than the nominal level for all sample sizes when it is compared with the proposed method. For both positive and negative coefficients of skewness for triangular distribution, the coverage ratio of $DMSD_{DM}$-t tends to decrease for a large sample size. Moreover, this coverage ratio of $DMSD_{DM}$-t tends to be the same level and slightly under 0.95 for each sample size when coefficient of skewness equals zero. Additionally, the coverage ratio of five intervals – Student's t, Johnson-t, Chen-t, YY-t and AADM-t – is close to 0.95 for all sample sizes and all coefficients of skewness for triangular distributions. When considering all of the distributions in this study, it is found that the coverage ratios of the proposed confidence interval are close to the nominal level and greater than this of the Bootstrap-percentile confidence interval for a small sample size, and the average width of these two methods tends to be no difference for all sample sizes.

Table 4.5. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of Chi-square distribution with df = 5

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9299	0.9311	0.9445	0.9306	0.8670	0.9233	0.8959	0.8869
	AW	4.3	4.3	5.2	4.3	3.3	4.2	3.8	3.5
20	CR	0.9366	0.9379	0.9563	0.9375	0.8687	0.9305	0.8650	0.9159
	AW	2.9	2.9	3.4	2.9	2.2	2.8	2.4	2.6
30	CR	0.9406	0.9417	0.9601	0.9413	0.8683	0.9350	0.8481	0.9281
	AW	2.3	2.3	2.7	2.3	1.8	2.2	1.9	2.2
40	CR	0.9420	0.9429	0.9619	0.9426	0.8669	0.9358	0.8288	0.9322
	AW	2.0	2.0	2.3	2.0	1.5	1.9	1.6	1.9
50	CR	0.9433	0.9440	0.9621	0.9437	0.8693	0.9364	0.8181	0.9353
	AW	1.8	1.8	2.0	1.8	1.4	1.7	1.4	1.7
100	CR	0.9472	0.9477	0.9633	0.9477	0.8685	0.9404	0.7580	0.9429
	AW	1.2	1.2	1.4	1.2	1.0	1.2	1.0	1.2

Table 4.6. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of Chi-square distribution with df = 10

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9391	0.9397	0.9483	0.9393	0.8749	0.9325	0.9099	0.8926
	AW	6.2	6.2	7.1	6.2	4.8	6.0	5.5	5.0
20	CR	0.9423	0.9430	0.9566	0.9428	0.8750	0.9376	0.8894	0.9200
	AW	4.1	4.1	4.6	4.1	3.2	4.0	3.5	3.7
30	CR	0.9452	0.9458	0.9599	0.9456	0.8755	0.9410	0.8801	0.9313
	AW	3.3	3.3	3.7	3.3	2.6	3.2	2.8	3.1
40	CR	0.9458	0.9461	0.9611	0.9459	0.8764	0.9424	0.8715	0.9357
	AW	2.8	2.8	3.1	2.8	2.2	2.8	2.4	2.7
50	CR	0.9460	0.9462	0.9606	0.9462	0.8741	0.9418	0.8643	0.9374
	AW	2.5	2.5	2.8	2.5	2.0	2.5	2.1	2.4
100	CR	0.9475	0.9476	0.9596	0.9475	0.8742	0.9441	0.8349	0.9434
	AW	1.8	1.8	1.9	1.8	1.4	1.7	1.5	1.7

Table 4.7. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of Chi-square distribution with df = 50

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9469	0.9470	0.9494	0.9471	0.8831	0.9397	0.9213	0.8985
	AW	13.9	13.9	14.9	13.9	10.8	13.5	12.4	11.3
20	CR	0.9487	0.9488	0.9546	0.9488	0.8822	0.9452	0.9121	0.9257
	AW	9.2	9.2	9.8	9.2	7.3	9.1	8.1	8.4
30	CR	0.9502	0.9503	0.9569	0.9502	0.8813	0.9465	0.9087	0.9349
	AW	7.4	7.4	7.8	7.4	5.8	7.3	6.5	7.0
40	CR	0.9504	0.9504	0.9571	0.9505	0.8825	0.9476	0.9063	0.9389
	AW	6.4	6.4	6.6	6.4	5.0	6.3	5.6	6.1
50	CR	0.9485	0.9484	0.9555	0.9484	0.8806	0.9465	0.9028	0.9389
	AW	5.7	5.7	5.9	5.7	4.5	5.6	5.0	5.5
100	CR	0.9483	0.9484	0.9540	0.9484	0.8795	0.9470	0.8946	0.9431
	AW	4.0	4.0	4.1	4.0	3.1	3.9	3.5	3.9

Table 4.8. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of triangular distribution with a = 0, b = 1 and c = 0.05

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9405	0.9420	0.9528	0.9412	0.8848	0.9361	0.9063	0.9030
	AW	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.3
20	CR	0.9450	0.9464	0.9597	0.9458	0.8898	0.9465	0.8920	0.9275
	AW	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
30	CR	0.9466	0.9477	0.9601	0.9473	0.8935	0.9505	0.8865	0.9357
	AW	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.2
40	CR	0.9489	0.9500	0.9614	0.9496	0.8964	0.9543	0.8829	0.9406
	AW	0.1	0.1	0.2	0.1	0.1	0.2	0.1	0.1
50	CR	0.9472	0.9478	0.9584	0.9475	0.8951	0.9533	0.8766	0.9401
	AW	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
100	CR	0.9493	0.9498	0.9584	0.9496	0.8978	0.9576	0.8600	0.9460
	AW	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 4.9. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of triangular distribution with a = 0, b = 1 and c = 0.50

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9474	0.9479	0.9452	0.9477	0.8880	0.9432	0.9201	0.9025
	AW	0.3	0.3	0.3	0.3	0.2	0.3	0.3	0.2
20	CR	0.9502	0.9509	0.9495	0.9507	0.8900	0.9501	0.9140	0.9294
	AW	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
30	CR	0.9491	0.9497	0.9486	0.9494	0.8881	0.9503	0.9114	0.9362
	AW	0.2	0.2	0.2	0.2	0.1	0.2	0.1	0.1
40	CR	0.9492	0.9497	0.9490	0.9496	0.8886	0.9513	0.9107	0.9393
	AW	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
50	CR	0.9500	0.9504	0.9502	0.9503	0.8895	0.9528	0.9102	0.9420
	AW	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
100	CR	0.9501	0.9503	0.9502	0.9502	0.8914	0.9541	0.9113	0.9458
	AW	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Table 4.10. Coverage ratio (CR) and average width (AW) of the 95% confidence intervals for the population mean of triangular distribution with a = 0, b = 1 and c = 0.95

n	Performance Measures	Confidence Interval Methods							
		Student-t	Johnson-t	Chen-t	YY-t	Mad-t	AADM-t	$DMSD_{DM}$-t	Bootstrap
10	CR	0.9411	0.9429	0.9249	0.9420	0.8858	0.9371	0.9062	0.9033
	AW	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
20	CR	0.9451	0.9464	0.9280	0.9460	0.8906	0.9462	0.8923	0.9281
	AW	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
30	CR	0.9466	0.9478	0.9310	0.9474	0.8921	0.9503	0.8847	0.9358
	AW	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.2
40	CR	0.9487	0.9496	0.9344	0.9492	0.8944	0.9538	0.8816	0.9411
	AW	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1
50	CR	0.9491	0.9501	0.9362	0.9497	0.8969	0.9553	0.8790	0.9424
	AW	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
100	CR	0.9493	0.9500	0.9403	0.9497	0.8979	0.9580	0.8609	0.9461
	AW	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

5. Real data applications

In this section, three real-life examples from normal and skewed distributions are analyzed to illustrate the applications of the proposed robust confidence interval.

5.1. Load at failure data

The first data set was obtained from Berndt (1989). The data describe the results of tensile adhesion tests (in megapascals) on 22 U-700 alloy specimens: 19.8, 10.1, 14.9, 7.5, 15.4, 15.4, 15.4, 18.5, 7.9, 12.7, 11.9, 11.4, 11.4, 14.1, 17.6, 16.7, 15.8, 10.5, 8.8, 13.6, 11.9, and 11.4. The Kolmogorov-Smirnov (K-S) goodness-of-fit test for normality for this data set has a p-value (p-value > 0.150) greater than $\alpha = 0.05$. We conclude that the data are in excellent agreement with a normal distribution with skewness = 0.07, kurtosis = -0.68, mean = 13.305 and standard deviation = 3.369.

Table 5.1. The 95% confidence intervals for the population mean of load at failure

Methods	Estimated Confidence Interval Limits		Width
	Lower Limit	Upper Limit	
Student-t	11.8108	14.7983	2.9875
Johnson-t	11.8125	14.7999	2.9874
Chen-t	11.7948	14.8143	3.0195
YY-t	11.8118	14.7992	2.9874
Mad-t	12.0611	14.5480	2.4869
AADM-t	11.7461	14.8630	3.1169
$DMSD_{DM}$-t	11.9306	14.6138	2.6832
Bootstrap	12.0159	14.6750	2.6591

The 95% CI for the population mean (μ) for load specimen failure is studied. The considered confidence intervals and their corresponding width have been given in Table 5.1. From Table 5.1, the 95% estimated confidence interval for population mean (μ) of load specimen failure, which is constructed using AADM-t method, gives the largest width, whereas the 95% of Mad-t confidence interval gives the smallest width and the secondary width is constructed for the 95% confidence interval using the $DMSD_{DM}$-t and Bootstrap-percentile methods. Therefore, the results from this real-life example as shown in Table 5.1 support the simulation study in Section 4.

5.2. Psychotropic drug exposure data

To study the average use of psychotropic drugs among non-antipsychotic drug users, the number of psychotropic drug users was reported for a random sample of $n = 20$ from different categories of drugs. The following data represent the number of users (Johnson and McFarland, 1993): 43.4, 24, 1.8, 0, 0.1, 170.1, 0.4, 150, 31.5, 5.2, 35.7, 27.3, 5, 64.3, 70, 94, 61.9, 9.1, 38.8, and 14.8. The data are checked and found to be positively

skewed with skewness = 1.57, kurtosis = 2.06, mean = 42.37 and standard deviation = 48.43. The considered confidence intervals and their corresponding width are given in Table 5.2.

Table 5.2. The 95% confidence intervals for the average use of psychotropic drugs

Methods	Estimated Confidence Interval Limits		Width
	Lower Limit	Upper Limit	
Student-t	19.8445	64.8955	45.0509
Johnson-t	20.3850	65.4359	45.0509
Chen-t	13.4694	71.2706	57.8013
YY-t	20.1629	65.2139	45.0509
Mad-t	25.7607	58.9793	33.2185
AADM-t	22.7839	61.9561	39.1722
$DMSD_{DM}$-t	19.3406	50.2838	30.9432
Bootstrap	23.7750	66.1800	42.4050

From Table 5.2, the 95% estimated confidence interval for the average use of psychotropic drugs, which is constructed by using the Chen-t method, gives the largest width and differs from other methods, whereas the 95% of $DMSD_{DM}$-t confidence interval gives the shortest width, followed by Mad-t and AADM-t confidence intervals. Since this data set is positively skewed, we conclude that the results in Table 5.2 support the simulation results in the case of positively skewed distribution of this study.

5.3. Long jump distance data

The following data represent the results of the final points scores reported for 40 players in long jump distance in meters (International Olympic Committee, 2019): 8.11, 8.11, 8.09, 8.08, 8.06, 8.03, 8.02, 7.99, 7.99, 7.97, 7.95, 7.92, 7.92, 7.92, 7.89, 7.87, 7.84, 7.79, 7.79, 7.77, 7.76, 7.72, 7.71, 7.66, 7.62, 7.61, 7.59, 7.55, 7.53, 7.5, 7.5, 7.42, 7.38, 7.38, 7.26, 7.25, 7.08, 6.96, 6.84, 6.55. The data are checked and found to be negatively skewed with skewness = −1.16, kurtosis = 1.20, mean = 7.6745 and standard deviation = 0.37. The considered confidence intervals and their corresponding width have been given in Table 5.3. From Table 5.3, the 95% estimated confidence interval for the population mean (μ) of the final points scores in long jump distance in meters, which is constructed by using Student's t, Johnson-t and YY-t methods, gives the same value of the largest width, whereas the 95% of $DMSD_{DM}$-t confidence interval gives the smallest width and the secondary width is constructed by using the Mad-t confidence interval. Since this data set is negatively skewed, we conclude that the results in Table 5.3 support the simulation results in the case of negatively skewed distribution of this study.

Table 5.3. The 95% confidence intervals for the population mean of the final scores for long jump distance in meters

Methods	Estimated Confidence Interval Limits		Width
	Lower Limit	Upper Limit	
Student-t	7.5528	7.7962	0.2434
Johnson-t	7.5512	7.7945	0.2434
Chen-t	7.5668	7.7822	0.2154
YY-t	7.5517	7.7951	0.2434
Mad-t	7.5793	7.7697	0.1903
AADM-t	7.5587	7.7903	0.2316
$DMSD_{DM}$-t	7.6242	7.8029	0.1787
Bootstrap	7.5553	7.7798	0.2245

6. Summary and concluding remarks

The proposed confidence interval, $DMSD_{DM}$-t, is an adjustment of the Student's t confidence interval based on the decile mean and the decile mean standard deviation. In addition, the simulation results show that in many cases the proposed confidence interval performs better than the existing estimators when observations are sampled from both normal and skewed distributions. Even though the Mad-t confidence interval tends to provide the smallest average width in the case of observations sampled from the normal distribution, the coverage ratio of this tends to be more under the nominal level when compared with the proposed confidence interval. That is, the performance of the $DMSD_{DM}$-t method is better than the Mad-t method for both coverage ratio and average width because the coverage ratio of the $DMSD_{DM}$-t confidence interval tends to be slightly below the nominal level. Although the coverage ratio of the proposed interval is slightly lower than that of the five intervals – Student's t, Johnson-t, Chen-t, YY-t and AADM-t – and the average width of this proposed interval is smaller than that of the five intervals, especially for a small sample size and observations sampled from the normal distribution. In the case of skewed distributions, such as observations sampled from chi-square distribution with a small coefficient of skewness, the average width of the proposed interval is also smaller than that of the five intervals – Student's t, Johnson-t, Chen-t, YY-t and AADM-t – even though the coverage ratio of the proposed interval is slightly lower than that of the five intervals. The bootstrap estimator for a confidence interval is reliable at least for n not very small.

Acknowledgement

Authors are thankful to anonymous reviewers and Editor-in-Chief for their valuable comments and suggestions, which certainly improve the quality and presentation of the paper. Author Moustafa Abu-Shawiesh wants to dedicate this paper to his most favourite teacher, Prof. Dr. Mokhtar Abdullah from Malaysia, for his constant inspiration during student life and affection that motivated him to achieve this present position.

References

Abu-Shawiesh, M. O. A., Al-Athari, F. M., Kittani, H. F., (2009). Confidence interval for the mean of a contaminated normal distribution, *Journal of Applied Sciences*, Vol. 9(15), pp. 2835–2840.

Abu-Shawiesh, M. O. A., Banik, S., Kibria, B. G., (2011). A simulation study on some confidence intervals for the population standard deviation, *SORT-Statistics and Operations Research Transactions*, Vol. 35(2), pp. 83–102.

Abu-Shawiesh, M. O. A., Banik, B., Kibria, B. M. G., (2018). Confidence Intervals based on absolute deviation for population mean of a positively skewed distribution, *International Journal of Computational and Theoretical Statistics*, Vol. 5(1), pp. 1–13.

Abu-Shawiesh, M. O. A., Saghir, A., (2019). Robust confidence intervals for the population mean alternatives to the Student-t confidence interval, *Journal of Modern Applied Statistical Methods*, Vol. 18(1), pp. 1–21.

Akyüz, H. E., Abu-Shawiesh, M. O. A., (2020). A robust confidence interval based on modified trimmed standard deviation for the mean of positively skewed populations, *Electronic Journal of Applied Statistical Analysis*, Vol. 13(1), pp. 164–182.

Baklizi, A., Kibria, B. M. G., (2009). One and two sample confidence intervals for estimating the mean of skewed populations: An empirical comparative study, *Journal of Applied Statistics*, Vol. 36(6), pp. 601—609.

Banik, S., Kibria, B. M. G., (2010). Comparison of some parametric and nonparametric type one sample confidence intervals for estimating the mean of a positively skewed distribution, *Communications in Statistics-Simulation and Computation*, Vol. 39(2), pp. 361–389.

Berndt, C. C., (1989). Instrumented tensile adhesion tests on plasma sprayed thermal barrier coatings, *Journal of Materials Engineering*, Vol. 11(4), pp. 275–282.

Boos, D., And Hughes-Oliver, J., (2000). How large does n have to be for Z and t intervals, *The American Statistician*, Vol. 54(2), pp. 121–128.

Cain, M. K., Zhang, Z., Yuan, K. H., (2016). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation, *Behavior Research Methods*, Vol. 49(5), pp. 1716–1735.

Chen, L., (1995). Testing the mean of skewed distributions, *Journal of the American Statistical Association*, Vol. 90(430), pp. 767–772.

Diciccio, T. J., Efron, B., (1996). Bootstrap confidence intervals, *Statistical Science*, Vol. 11(3), pp. 189–228.

Doulah, M. S. U., (2018). Alternative measures of standard deviation coefficient of variation and standard error, *International Journal of Statistics and Applications*, Vol. 8(6), pp. 309–315.

Efron, B., Tibshirani, R. J., (1993). *An Introduction to the Bootstrap*, London: Chapman & Hall.

Gastwirth, J. L., (1982). Statistical Properties of a Measure of Tax Assessment Uniformity, *Journal of the Statistical Inference and Planning*, Vol. 6(1), pp. 1–12.

Ghosh, S., Polansky, A. M., (2016). New bootstrap confidence interval for means of a positively skewed distributions, *Communications in Statistics-Theory and Methods*, Vol. 45(23), pp. 6915—6927.

International Olympic Committee, (2019). London 2012 long jump men – Olympic athletics. Retrieved from https://www.olympic.org/london2012/athletics/long-jump-men.

Johnson, N. J., (1978). Modified t-tests and confidence intervals for asymmetrical population, *Journal of the American Statistical Association*, Vol. 73(363), pp. 536–544.

Johnson, R. E., Mcfarland, B. H., (1993). Antipsychotic drug exposure in a health maintenance organization, *Medical Care*, Vol. 31(5), pp. 432–444.

Kibria, B. M. G., (2006). Modified confidence intervals for the mean of the asymmetric distribution, *Pakistan Journal of Statistics*, Vol. 22(2), pp. 109–120.

Luh, W., Guo, J., (2001). Transformation works for non-normality? On one-sample transformation trimmed t methods, *British Journal of Mathematical and Statistical Psychology*, Vol. 54(2), pp. 227–236.

Marinho, P. R. D., Cordeiro, G. M., Pena, R., Alizadeh, M., Bourguignon, M., (2018). The exponentiated logarithmic generated family of distributions and the evaluation of the confidence intervals by percentile bootstrap, *Brazilian Journal of Probability and Statistics*, Vol. 32(2), pp. 281–308.

Mcdonald, J. H., (2014). *Handbook of Biological Statistics*, 3rd ed., Maryland: Sparky House Publishing.

Meeden, G., (1999). Interval estimators for the population mean for skewed distributions with a small sample size, *Journal of Applied Statistics*, Vol. 26(1), pp. 81–96.

Miller, J. M., Penfield, R. D., (2005). Using the score method to construct asymmetric confidence intervals: An SAS program for content validation in scale development, *Behavior Research Methods*, Vol. 37(3), pp. 450–452.

Mudelsee, M., Alkio, M., (2007). Quantifying effects in two-sample environmental experiments using bootstrap confidence intervals, *Environmental Modelling and Software*, Vol. 22(1), pp. 84–96.

Neyman, J., (1937). Outline of a theory of statistical estimation based on the classical theory of probability, *Philosophical Transactions of the Royal Society Series A*, Vol. 236(767), pp. 333–380.

Pek, J., Wong, A.C.M., Wong, O., (2017). Confidence intervals for the mean of non-normal distribution: Transform or not to transform, *Open Journal of Statistics*, Vol. 7(3), pp. 405–421.

Rana, S., Doulah, M. S. U., Midi, H., Imon, A. H. M. R., (2012). Decile mean: a new robust measure of central tendency, *Chiang Mai Journal of Sciences*, Vol. 39(3), pp. 478–485.

Sharma, K. K., Kumar, A., Chaudhary, A., (2009). *Statistics in Management Studies*. India: Krishna Media (P) Ltd.

Shi, W., Kibria, B. M. G., (2007). On some confidence intervals for estimating the mean of a skewed population, *International Journal of Mathematical Education in Science and Technology*, Vol. 38(3), pp. 412–421.

Sindhumol, M. R., Srinivasan, M. M., Gallo, M., (2016). Robust control charts based on modified trimmed standard deviation and Gini's mean difference, *Journal of Applied Quantitative Methods*, Vol. 11(3), pp. 18–30.

Sinsomboonthong, J., Abu-Shawiesh, M. O. A., Kibria, B. M. G., (2020). Performance of robust confidence intervals for estimating population mean under both non-normality and in presence of outliers, *Advances in Science, Technology and Engineering Systems Journal*, Vol. 5(3), pp. 442–449.

Student, (1908). The probable error of a mean, Biometrika, Vol. 6(1), pp. 1–25.

Yanagihara, H., Yuan, K. H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis, *British Journal of Mathematical and Statistical Psychology*, Vol. 58(2), pp. 209–237.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 129–152, DOI 10.21307/stattrans-2022-008
Received – 05.05.2020; accepted – 13.11.2020

Estimation procedures for reliability functions of Kumaraswamy-G Distributions based on Type II Censoring and the sampling scheme of Bartholomew

Aditi Chaturvedi[1], **Surinder Kumar**[2]

ABSTRACT

In this paper, we consider Kumaraswamy-G distributions and derive a Uniformly Minimum Variance Unbiased Estimator (UMVUE) and a Maximum Likelihood Estimator (MLE) of the two measures of reliability, namely $R(t) = P(X > t)$ and $P = P(X > Y)$ under Type II censoring scheme and sampling scheme of Bartholomew (1963). We also develop interval estimates of the reliability measures. A comparative study of the different methods of point estimation has been conducted on the basis of simulation studies. An analysis of a real data set has been presented for illustration purposes.

Key words: interval estimation, Kumaraswamy-G distributions, Monte-Carlo simulation, point estimation.

1. Introduction

The Kumaraswamy (Kum) distribution is widely applied to model the random phenomenon having finite lower and upper bounds, e.g.-, the height of individuals, atmospheric temperatures, hydrological data such as daily rain fall, daily stream flow, etc. The distribution was first defined by Kumaraswamy (1976, 1978). Nadarajah (2008) demonstrated that the distribution may be viewed as a special case of three parameter Beta distribution. Several other unimodal distributions can also be approximated by Kumaraswamy's distribution [See, Kumaraswamy (1980) and Ponnambalam *et al.* (2001)]. Garg (2009) studied the generalized order statistics from the Kum distribution. Jones (2009) explored the background and genesis of the Kum distribution and demonstrated some similarities and differences between the beta and Kum distributions. He highlighted several advantages of the Kum distribution over the beta distribution. In hydrology and related areas, the Kum distribution has received considerable interest [See, Sundar and Subbiah (1989), Fletcher and Ponnambalam (1996), Seifi *et al.* (2000), Ponnambalam *et al.* (2001) and Ganji *et al.* (2006)]. Sindhu *et al.* (2013) focused on Bayesian and non-Bayesian estimation for the shape parameter of the Kum distribution under Type-II censored samples.

Eldin *et al.* (2014) obtained the MLE's and Bayes estimators for the parameters of the Kum distribution under general progressive Type II censoring. Mameli (2015) propose a new generalization of the skew-normal distribution, referred to as the Kum skew-normal

[1]Corresponding Author. Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow, India. E-mail: caditic@gmail.com.
[2]Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow, India. E-mail: surinderntls@gmail.com.

distribution. He demonstrated that this new distribution is computationally more tractable than the Beta skew-normal distribution proposed by Mameli and Musio (2013). Kızılaslan and Nadar (2016) considered the Kum distribution, when the lower record values along with the number of observations following the record values (inter-record times) were observed, and derived the maximum likelihood and Bayes estimators for estimating the parameters of the distribution as well as for the future record values prediction. Dey *et al.* (2017) focussed on Bayesian and non-Bayesian estimation of multicomponent stress–strength reliability when both step and strength follow the Kum distribution with common shape parameter. Dey *et al.* (2018) considered and investigated performance of ten different frequentist approaches for estimation of parameters of Kum distribution, namely, maximum likelihood estimators, moments estimators, L-moments estimators, percentile based estimators, least squares estimators, weighted least squares estimators, maximum product of spacings estimators, Cramér–von-Mises estimators, Anderson–Darling estimators and right tailed Anderson–Darling estimators.

In recent years, a large amount of literature has been developed regarding the generalization of classical distributions. For some of the citations, one may refer to Hassan *et al.* (2020) and the references therein. Cordeiro and Castro (2011) introduced a new Kumaraswamy generalized (Kum-G) family of distributions and discussed its basic statistical properties. They mentioned that the Kum-G family of densities has ability of fitting skewed data and allows for greater flexibility of its tails. The distribution generalizes the modelling ability of the Kumaraswamy distribution and can be widely applied in many areas of engineering and biology. Nadarajah *et al.* (2012) derived simple representation for the Kum-G family of distributions as a linear combination of exponentiated distributions and studied its general properties. They obtained MLEs of its parameters and discussed its bivariate extension as well. Tamandi and Nadarajah (2016) developed maximum spacing estimation procedure for the parameters of Kum-G distribution. Kundu and Chowdhary (2018) compared the minimums of two independent and heterogeneous samples each following Kum-G distribution with respect to usual stochastic ordering and hazard rate ordering. They also established likelihood ratio ordering between the minimum order statistics for heterogeneous multiple-outlier Kum-G random variables with the same parent distribution function. Kumari *et al.* (2019) provided characterization ofthe Kum-G distribution based on record values and obtained point and interval estimates of two measures of reliability function $R(t) = P(X > t)$ and $P = P(X > Y)$ based on records. They considered two types of point estimators, namely UMVUE's and MLE's and developed procedures for testing hypotheses related to various parametric functions. Chaturvedi and Bhatnagar (2020) developed classical and preliminary test estimators for measures of reliability of the Kum-G distribution under progressive Type II censoring.

The purpose of the present paper is to extend the results of Kumari *et al.* (2019) for the cases of Type II censoring and the sampling scheme proposed by Bartholomew. Considering the Kum-G distribution, we develop UMVUE's and MLE's for the reliability functions, $R(t)$ and P. For deriving UMVUE's, we followed the approach proposed by Chaturvedi and Tomer (2003), which saves tedious and time-consuming calculation of stress–strength function. The paper is organized as follows: In Section 2, we provide point estimators and exact confidence intervals for the q^{th} power of parameter α, for $q \in (-\infty, +\infty)$, and for functions

$R(t)$ and P based on Type II censoring scheme. In Section 3, based on the sampling scheme proposed by Bartholomew (1963), the point estimators for the α^q, $R(t)$ and P are provided. In Section 4, we present findings of simulation studies followed by real data analysis in Section 5. We end with a brief set of conclusions in Section 6. Proofs of some important results can be found in the Appendix.

2. Estimation based on Type II Censoring Scheme

A random variable X is said to follow the Kumaraswamy (1980) distribution if its pdf is given by

$$f(x;\alpha,\beta) = \alpha\beta x^{\beta-1}(1-x^{\beta})^{\alpha-1};\ 0<x<1,\ \alpha,\beta>0. \tag{1}$$

Considering the complete sample case, Nadar *et al.* (2014) have obtained the estimator of P for the distribution given in (1) assuming the parameter 'β' to be common for the two distributions.

A random variable X follow Kumaraswamy-G distributions [Cordeiro and Castro (2011)], if its pdf is of the form

$$f(x;\alpha,\beta) = \alpha\beta g(x)G^{\beta-1}(x)[1-G^{\beta}(x)]^{\alpha-1};\ x>0,\ \alpha,\beta>0, \tag{2}$$

where $g(x)$ denotes the pdf of $G(x)$, α and β are the shape parameters of the Kum-G distribution.

It is to be noted that the distribution given in (2) reduces to the Kumaraswamy distribution when $G(x) = x$.

2.1. UMVUE's and MLE's of $\alpha^q, R(t)$ and P Based on Type II Censoring

Suppose 'n' items are put on a test and the test is terminated after the first 'r' ordered observations are recorded. Let us denote by $0 < X_{(1)} \le X_{(2)} \le ... \le X_{(r)}$, $0 < r < n$, the lifetimes of first r failures. Obviously $(n-r)$ items survived until $X_{(r)}$. Here, we provide an important lemma, which will be helpful in proving the main results of this section.

***Lemma 1** Let

$$S_{(r)} = -\left[\sum_{i=1}^{r} ln\left\{1-G^{\beta}(x_i)\right\} + (n-r)ln\left\{1-G^{\beta}(x_r)\right\}\right],$$

then, $S_{(r)}$ is complete and sufficient for the Kum-G distribution (2). Moreover, the pdf of $S_{(r)}$ is given by

$$g_{S_{(r)}}(s;\alpha) = \frac{1}{\Gamma(r)} s^{r-1}\alpha^{r}\exp\{-\alpha s\}, s>0, \alpha>0, r>0, \tag{3}$$

*The proof of Lemma 1 is available from the corresponding author on request.

where, $\Gamma(\cdot)$ denotes the Gamma function.

In the following theorems, we provide the UMVUEs of α^q, $R(t)$ and P, based on Type II censoring scheme and under the assumption that β is known.

Theorem 1 For $q \in (-\infty, \infty)$, the UMVUE of α^q is given by:

$$\widetilde{\alpha}_{II}^{q} = \begin{cases} \frac{\Gamma(r)}{\Gamma(r-q)} S_{(r)}^{-q}; & r-q>0 \\ 0; & \text{otherwise.} \end{cases}$$

Proof. From (3),

$$E\left(S_{(r)}^{-q}\right) = \frac{\Gamma(r-q)}{\Gamma(r)}\alpha^q, r > q, \tag{4}$$

and the theorem follows from Lehmann-Scheffe theorem [see Rohatgi & Saleh(2012)]. Let us write the pdf (2) as follows

$$f(x;\alpha,\beta) = \frac{\alpha\beta g(x)G^{\beta-1}(x)}{1-G^{\beta}(x)} \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} \left\{-ln(1-G^{\beta}(x))\right\}^i \alpha^i,$$

then the following Corollary straight away follows from Theorem 1.

Corollary 1 *The UMVUE of the sampled pdf at a specified point x is:*

$$\widetilde{f}_{II}(x;\alpha,\beta) = \begin{cases} \frac{\beta g(x)G^{\beta-1}(x)}{B(1,r-1)S_{(r)}(1-G^{\beta}(x))} \left(1+\frac{ln(1-G^{\beta}(x))}{S_{(r)}}\right)^{r-2}; \\ \qquad\qquad -ln(1-G^{\beta}(x)) < S_{(r)} \\ 0; \qquad\qquad \textit{otherwise}, \end{cases}$$

where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function.

Theorem 2 The UMVUE of $R(t)$ at a specified point t is

$$\widetilde{R}(t)_{II} = \begin{cases} \left[1+\frac{ln(1-G^{\beta}(t))}{S_{(r)}}\right]^{r-1}; & -ln(1-G^{\beta}(t)) < S_{(r)} \\ 0; & \text{otherwise.} \end{cases}$$

Proof. Using Corollary 1, we have

$$\widetilde{R}(t)_{II} = \int_t^{\infty} \frac{\beta g(x)G^{\beta-1}(x)}{B(1,r-1)S_{(r)}(1-G^{\beta}(x))} \left(1+\frac{ln(1-G^{\beta}(x))}{S_{(r)}}\right)^{r-2} dx,$$

and the result follows by substituting $\frac{-ln(1-G^{\beta}(x))}{S_{(r)}}$=v.

Let X and Y be two independent random variables following the classes of distributions $f_1(x;\alpha_1,\beta_1)$ and $f_2(y;\alpha_2,\beta_2)$, respectively, where

$$f_1(x;\alpha_1,\beta_1)=\alpha_1\beta_1 g(x)G^{\beta_1-1}(x)(1-G^{\beta_1}(x))^{\alpha_1-1};\ x>0,\ \alpha_1,\beta_1>0 \tag{5}$$

and

$$f_2(y;\alpha_2,\beta_2)=\alpha_2\beta_2 h(y)H^{\beta_2-1}(y)(1-H^{\beta_2}(y))^{\alpha_2-1};\ y>0,\ \alpha_2,\beta_2>0. \tag{6}$$

Let n items on X and m items on Y are put on a life test and the termination numbers for X and Y are r and r', respectively. Let us define

$$S_{(r)}=-\left[\sum_{i=1}^{r} ln(1-G^{\beta_1}(x_i))+(n-r)(ln(1-G^{\beta_1}(x_r))\right]$$

and

$$T_{(r')}=-\left[\sum_{j=1}^{r'} ln(1-H^{\beta_2}(y_j))+(m-r')(ln(1-H^{\beta_2}(y_{r'}))\right].$$

In the following theorem, we obtain the UMVUE of P.

***Theorem 3** The UMVUE of P, when X and Y belong to different family of distributions, is given by

$$\widetilde{P}_{II}=\begin{cases}\displaystyle\int_{z=0}^{c}\frac{1}{B(1,r'-1)}\left[1+\frac{ln\left\{1-G(H^{-1}(1-e^{-zT_{(r')}}))^{\beta_1/\beta_2}\right\}}{S_{(r)}}\right]^{r-1}\\ (1-z)^{r'-2}dz;\quad if\ G^{-1}\left\{(1-e^{-S_{(r)}})^{1/\beta_1}\right\}\le H^{-1}\left\{(1-e^{-T_{(r')}})^{1/\beta_2}\right\}\\ \\ \displaystyle\int_{z=0}^{1}\frac{1}{B(1,r'-1)}\left[1+\frac{ln\left\{1-G(H^{-1}(1-e^{-zT_{(r')}}))^{\beta_1/\beta_2}\right\}}{S_{(r)}}\right]^{r-1}(1-z)^{r'-2}dz;\\ if\ G^{-1}\left\{(1-e^{-S_{(r)}})^{1/\beta_1}\right\}>H^{-1}\left\{(1-e^{-T_{(r')}})^{1/\beta_2}\right\},\end{cases}$$

where $c=-T^{-1}ln\left[1-H\left\{G^{-1}(1-e^{-S_{(r)}})^{\beta_2/\beta_1}\right\}\right]$.

Along the lines of Theorem 3, we can easily prove the following Corollary.

Corollary 2 *The UMVUE of P, when X and Y belong to same family of distributions, i.e., when* $G(\cdot)=H(\cdot)$ *and* $\beta_1=\beta_2$*, is given by*

**The proof of Theorem 3 is available from the corresponding author on request.*

$$\widetilde{P}_{II} = \begin{cases} \frac{1}{B(1,r'-1)} \sum_{i=0}^{r'-2} (-1)^i \binom{r'-2}{i} \left(\frac{S_{(r)}}{T_{(r')}}\right)^i B(i+1,r); & S_{(r)} \leq T_{(r')} \\ \\ \frac{1}{B(1,r'-1)} \sum_{j=0}^{r-1} (-1)^j \binom{r-1}{j} \left(\frac{T_{(r')}}{S_{(r)}}\right)^j B(j+1,r'-1); & S_{(r)} > T_{(r')}. \end{cases}$$

Using (2), the joint pdf of $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(r)}$ is given by

$$h(x_{(1)},x_{(2)},...,x_{(r)};\alpha,\beta) = \frac{n!}{(n-r)!}\alpha^r\beta^r \prod_{i=1}^{r} \frac{g(x_{(i)})G^{\beta-1}(x_{(i)})}{1-G^{\beta}(x_{(i)})} \exp(-\alpha S_{(r)}) \tag{7}$$

It can be easily seen from (7) that the MLE of α^q based on Type II censoring is

$$\widehat{\alpha}_{II}^q = \left(\frac{r}{S_{(r)}}\right)^q. \tag{8}$$

From (2) and invariance property of maximum likelihood estimators, the MLE of $f(x)$ is given by

$$\widehat{f(x)}_{II} = \frac{r}{S_{(r)}}\beta g(x)G^{\beta-1}(x)\left[1-G^{\beta}(x)\right]^{\frac{r}{S_{(r)}}-1}.$$

Similarly, using the invariance property of MLE, the MLE of $R(t)$ is given by

$$\widehat{R(t)}_{II} = \left(1-G^{\beta}(t)\right)^{\frac{r}{S_{(r)}}}. \tag{9}$$

The MLE of P, when X and Y belong to different family of distributions, is given by

$$\widehat{P}_{II} = \int_{z=0}^{1} \left[1-G^{\beta_1}\left\{H^{-1}(z^{1/\beta_2})\right\}\right]^{\frac{r}{S_{(r)}}} \frac{r'}{T_{(r')}}(1-z)^{\frac{r'}{T_{(r')}}-1}.$$

The MLE of P, when X and Y belong to same family of distributions, i.e.-, when $G(\cdot)=H(\cdot)$ and $\beta_1=\beta_2$ is given by

$$\widehat{P}_{II} = \frac{r'S_{(r)}}{r'S_{(r)}+rT_{(r')}}. \tag{10}$$

2.2. Exact Confidence Intervals for α, $R(t)$ and P based on Type II Censoring

We consider the problem of constructing a two-sided confidence interval for α. The confidence interval is obtained by using pivotal quantity $2\alpha S_{(r)}$. If we define $\chi^2(\nu)$ as the value of χ^2 such that

$$P(\chi^2 > \chi^2(\delta)) = \int_{\chi^2(\delta)}^{\infty} P(\chi^2)d\chi^2 = \delta, \tag{11}$$

where $P(\chi^2)$ is the pdf of χ^2 distribution with 2r degrees of freedom, then by using the fact that $2\alpha S_{(r)} \sim \chi^2_{2r}$, the confidence interval is given by

$$P\left(\frac{\chi^2\left(1-\frac{\delta}{2}\right)}{2S_{(r)}} \leq \alpha \leq \frac{\chi^2\left(\frac{\delta}{2}\right)}{2S_{(r)}}\right) = 1-\delta, \tag{12}$$

where $\chi^2\left(\frac{\delta}{2}\right)$ and $\chi^2\left(1-\frac{\delta}{2}\right)$ are obtained by using (11). Thus, for known β, $100(1-\delta)\%$ confidence interval for α is given by

$$\left(\frac{\chi^2\left(1-\frac{\delta}{2}\right)}{2S_{(r)}}, \frac{\chi^2\left(\frac{\delta}{2}\right)}{2S_{(r)}}\right).$$

Further, for $q<0$, the confidence interval for α^q is given by

$$\left(\left(\frac{\chi^2\left(\frac{\delta}{2}\right)}{2S_{(r)}}\right)^q, \left(\frac{\chi^2\left(1-\frac{\delta}{2}\right)}{2S_{(r)}}\right)^q\right)$$

and for $q>0$, the confidence interval for α^q is given by

$$\left(\left(\frac{\chi^2\left(1-\frac{\delta}{2}\right)}{2S_{(r)}}\right)^q, \left(\frac{\chi^2\left(\frac{\delta}{2}\right)}{2S_{(r)}}\right)^q\right).$$

The problem of obtaining the confidence interval for the reliability function $R(t) = (1-G^\beta(t))^\alpha$ can be solved by noting that $R(t_\circ;\alpha)$ is a decreasing function of α. Thus, $\Psi_1(x_1,x_2,...,x_n) \leq (1-G^\beta(t_\circ))^\alpha$ is equivalent to $\alpha \leq ln\Psi_1(x_1,x_2,...,x_n)/ln(1-G^\beta(t_\circ))$ and $\Psi_2(x_1,x_2,...,x_n) \geq (1-G^\beta(t_\circ))^\alpha$ is equivalent to $\alpha \geq ln\Psi_2(x_1,x_2,...,x_n)/ln(1-G^\beta(t_\circ))$. Therefore, the expression

$$P\left(\Psi_1(x_1,x_2,...,x_n) \leq (1-G^\beta(t_\circ))^\alpha \leq \Psi_2(x_1,x_2,...,x_n)\right) = 1-\delta$$

is equivalent to

$$P\left(\frac{ln\Psi_2(x_1,x_2,...,x_n)}{ln(1-G^\beta(t_\circ))} \leq \alpha \leq \frac{ln\Psi_1(x_1,x_2,...,x_n)}{ln(1-G^\beta(t_\circ))}\right) = 1-\delta. \tag{13}$$

Comparing (12) and (13), it immediately follows that $\chi^2\left(1-\frac{\delta}{2}\right)/2S_{(r)} = ln\Psi_2(x_1,x_2,...,x_n)/ln(1-G^\beta(t_\circ))$ and $\chi^2\left(\frac{\delta}{2}\right)/2S_{(r)} = ln\Psi_1(x_1,x_2,...,x_n)/ln(1-G^\beta(t_\circ))$.

Therefore,

$$\Psi_1 = exp\left[ln(1-G^{\beta}(t_\circ))\frac{\chi^2\left(\frac{\delta}{2}\right)}{2S_{(r)}}\right] and\ \Psi_2 = exp\left[ln(1-G^{\beta}(t_\circ))\frac{\chi^2\left(1-\frac{\delta}{2}\right)}{2S_{(r)}}\right].$$

Thus, for known β, $(1-\delta)100\%$ confidence interval for $R(t_\circ,\alpha)$ is given by

$$\left(exp\left[ln(1-G^{\beta}(t_\circ))\frac{\chi^2\left(\frac{\delta}{2}\right)}{2S_{(r)}}\right], exp\left[ln(1-G^{\beta}(t_\circ))\frac{\chi^2\left(1-\frac{\delta}{2}\right)}{2S_{(r)}}\right]\right).$$

In order to obtain the confidence interval for P, we utilize the fact that $\frac{2\alpha_1 S_{(r)}/2r}{2\alpha_2 T_{(r')}/2r'} \sim F_{2r,2r'}$. Thus, the confidence interval for P is given by

$$P\left[\left(\frac{rT_{(r')}F(\frac{\delta}{2})}{r'S_{(r)}}+1\right)^{-1} \leq \frac{\alpha_2}{\alpha_1+\alpha_2} \leq \left(\frac{rT_{(r')}F(1-\frac{\delta}{2})}{r'S_{(r)}}+1\right)^{-1}\right] = 1-\delta.$$

Therefore, for known β, $(1-\delta)100\%$ confidence interval for P is given by

$$\left[\left(\frac{rT_{(r')}F(\frac{\delta}{2})}{r'S_{(r)}}+1\right)^{-1}, \left(\frac{rT_{(r')}F(1-\frac{\delta}{2})}{r'S_{(r)}}+1\right)^{-1}\right].$$

3. Estimation based on the Sampling Scheme of Bartholomew

Throughout this section, we assume that n items are put on a test and we terminate life-testing experiment at a preassigned time $t_\circ$. Suppose we carry out time-censored test where the items that fail are immediately replaced. Let $X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$ be the failure times of n items under a test from (2). The test begins at time $X_{(0)} = 0$ and the system operates until $X_{(1)} = x_1$, when the first failure occurs. The failed item is replaced by a new one and the system operates until the second failure occurs at time $X_{(2)} = x_2$ and so on. The experiment is terminated at time $t_\circ$. Here, $X_{(i)}$ is the time until i^{th} failure measured from time 0.

3.1. UMVUEs and MLEs of α^q, $R(t)$ and P, based on the Sampling Scheme of Bartholomew

We, first provide an important lemma, which will be utilized in deducing UMVUE's and MLE's of α^q, $R(t)$ and P.

***Lemma 2** Let $N(t_\circ)$ be the number of failures during the interval $[0;t_\circ]$. Then,

---

*The proof of Lemma 2 is available from the corresponding author on request.

$$P[N(t_\circ) = r|t_\circ] = \frac{[-n\alpha\ ln\ (1-G^\beta(t_\circ))]^r}{r!} exp\left\{n\alpha\ ln\ (1-G^\beta(t_\circ))\right\}.$$

In the following theorems, we provide the UMVUEs of α^q, $R(t)$ and P, based on the sampling scheme of Bartholomew (1963).

Theorem 4 For positive integer q, the UMVUE of α^q is given by

$$\widetilde{\alpha}_I^q = \begin{cases} \frac{r!}{(r-q)!}[-n\ ln\left\{1-G^\beta(t_\circ)\right\}]^{-q}; & r-q>0 \\ 0; & \text{otherwise.} \end{cases}$$

Proof. It follows from Lemma 2 and the Fisher-Neyman factorization theorem [see Rohatgi and Saleh (2012), p. 341] that r is sufficient for α. Moreover, since the distribution of r belongs to an exponential family, it is also complete [see Rohatgi and Saleh (2012), p. 347]. The theorem now follows from the result that the q^{th} factorial moment of the distribution of r is given by

$$E[r(r-1)(r-2)...(r-q+1)] = \left[-n\alpha\ ln\left\{1-G^\beta(t_\circ)\right\}\right]^q.$$

Let us write the pdf (2) as follows:

$$f(x;\alpha,\beta) = \frac{\alpha\beta g(x)G^{\beta-1}(x)}{1-G^\beta(x)} \sum_{i=0}^{\infty} \frac{(-1)^i}{i!} \left\{-ln(1-G^\beta(x))\right\}^i \alpha^i.$$

Then, the Corollary 3 straight away follows from Theorem 4.

Corollary 3 *The UMVUE of $f(x;\alpha,\beta)$ at a specified point x is*

$$\widetilde{f}_I(x;\alpha,\beta) = \begin{cases} \frac{r\beta g(x)G^{\beta-1}(x)}{\left[-n\ ln(1-G^\beta(t_\circ))\right](1-G^\beta(x))} \left(1-\frac{ln(1-G^\beta(x))}{n\ ln(1-G^\beta(t_\circ))}\right)^{r-1}; \\ \qquad\qquad ln(1-G^\beta(x)) < n\ ln(1-G^\beta(t_\circ)) \\ 0; \qquad\qquad otherwise. \end{cases}$$

Theorem 5 The UMVUE of $R(t)$ at a specified point t is given by

$$\widetilde{R}(t) = \begin{cases} \left[1-\frac{ln(1-G^\beta(t))}{n\ ln(1-G^\beta(t_\circ))}\right]^r; & ln(1-G^\beta(t)) < n\ ln(1-G^\beta(t_\circ)) \\ 0; & \text{otherwise.} \end{cases}$$

Proof. Using Corollary (3),

$$\widetilde{R}(t) = \int_t^\infty \frac{r\beta g(x)G^{\beta-1}(x)}{\left[-n\ ln(1-G^\beta(t_\circ))\right](1-G^\beta(x))} \left(1-\frac{ln(1-G^\beta(x))}{n\ ln(1-G^\beta(t_\circ))}\right)^{r-1} dx,$$

and the result follows by substituting $\frac{ln(1-G^{\beta}(x))}{n\ ln(1-G^{\beta}(t_\circ))} = z$.

Let n items on X and m on Y be put on a life test, where X and Y are distributed as in (5) and (6). Let $t_\circ$ and $t_{\circ\circ}$ be the termination times for X and Y , respectively and r and r' be the number of failures before $t_\circ$ and $t_{\circ\circ}$, respectively. Obviously, using Corollary 3, the UMVUEs of $f_1(x;\alpha_1,\beta_1)$and $f_2(y;\alpha_2,\beta_2)$, based on the sampling scheme of Bartholomew is given by

$$\widetilde{f}_{1I}(x;\alpha_1,\beta_1) = \frac{r\beta_1 g(x)G^{\beta_1-1}(x)}{\left[-n\ ln(1-G^{\beta_1}(t_\circ))\right](1-G^{\beta_1}(x))}\left(1-\frac{ln(1-G^{\beta_1}(x))}{n\ ln(1-G^{\beta_1}(t_\circ))}\right)^{r-1}; \quad (14)$$

$$ln(1-G^{\beta_1}(x)) < n\ \ ln(1-G^{\beta_1}(t_\circ))$$

and

$$\widetilde{f}_{2I}(y;\alpha_2,\beta_2) = \frac{r'\beta_2 h(y)H^{\beta_2-1}(y)}{\left[-m\ ln(1-H^{\beta_2}(t_{\circ\circ}))\right](1-H^{\beta_2}(y))}\left(1-\frac{ln(1-H^{\beta_2}(y)}{m\ ln(1-H^{\beta_2}(t_{\circ\circ}))}\right)^{r'-1}; \quad (15)$$

$$ln(1-H^{\beta_2}(y)) < m\ ln(1-H^{\beta_2}(t_{\circ\circ})).$$

***Theorem 6** The UMVUE of P is given by

$$\widetilde{P}_I = \begin{cases} r'\int_{z=0}^{c}\left[1-\frac{ln\left\{1-G^{\beta_1}(H^{-1}(1-(1-H^{\beta_2}(t_{\circ\circ}))^{mz}))^{1/\beta_2}\right\}}{n\ ln\left\{1-G^{\beta_1}(t_\circ)\right\}}\right](1-z)^{r'-1}dz; \\ G^{-1}\left\{1-(1-G^{\beta_1}(t_\circ))^n\right\}^{\frac{1}{\beta_1}} \le H^{-1}\left\{1-(1-H^{\beta_2}(t_{\circ\circ}))^m\right\}^{\frac{1}{\beta_2}} \\ r'\int_{z=0}^{1}\left[1-\frac{ln\left\{1-G^{\beta_1}(H^{-1}(1-(1-H^{\beta_2}(t_{\circ\circ}))^{mz}))^{1/\beta_2}\right\}}{n\ ln\left\{1-G^{\beta_1}(t_\circ)\right\}}\right](1-z)^{r'-1}dz; \\ G^{-1}\left\{1-(1-G^{\beta_1}(t_\circ))^n\right\}^{\frac{1}{\beta_1}} > H^{-1}\left\{1-(1-H^{\beta_2}(t_{\circ\circ}))^m\right\}^{\frac{1}{\beta_2}}, \end{cases}$$

where $c = \frac{ln\left[1-H^{\beta_2}\left\{G^{-1}(1-(1-G^{\beta_1}(t_\circ))^n)^{1/\beta_1}\right\}\right]}{m\ ln\left\{1-G^{\beta_1}(t_\circ)\right\}}$.

Corollary 4 *The UMVUE of P, when X and Y belong to the same family of distributions, i.e., $G(\cdot) = H(\cdot)$ with $\beta_1 = \beta_2$ and $t_\circ = t_{\circ\circ}$ is given by*

$$\widetilde{P}_I = \begin{cases} r'\sum_{i=0}^{r'-1}(-1)^i\binom{r'-1}{i}\left(\frac{n}{m}\right)^{i+1}B(i+1,r+1); & n \le m \\ r'\sum_{j=0}^{r}(-1)^j\binom{r}{j}\left(\frac{m}{n}\right)^{j}B(j+1,r'); & n > m. \end{cases}$$

*The proof of Theorem 6 is available from the corresponding author on request.

It can be easily seen from Lemma 2 that the MLE of α^q based on the sampling scheme of Bartholomew (1963) is given by

$$\widehat{\alpha}_I^q = \left(\frac{-r}{n\ ln(1-G^\beta(t_\circ))}\right)^q. \tag{16}$$

Using (2), $R(t)$ at point t is given by

$$R(t) = (1-G^\beta(t))^\alpha. \tag{17}$$

From (17) and invariance property of MLEs, the MLE of $R(t)$ is given by

$$\widehat{R}(t)_I = [1-G^\beta(t)]^{\frac{-r}{n\ ln(1-G^\beta(t_\circ))}}. \tag{18}$$

Similarly, using the invariance property of MLE, the MLE of $f(x;\alpha,\beta)$ at a specified point x is

$$\widehat{f}_I(x;\alpha,\beta) = \frac{-r}{n\ ln\left\{1-G^\beta(t_\circ)\right\}} g(x)G^{\beta-1}(x)[1-G^\beta(x)]^{\frac{-r}{n\ ln\left\{1-G^\beta(t_\circ)\right\}}-1}.$$

The MLE of P, when X and Y belong to a different family of distributions, is given by

$$\widehat{P}_I = \int_{z=0}^{1} [1-G^{\beta_1}\left\{H^{-1}(z^{1/\beta_2})\right\}]^{\frac{-r}{n\ ln(1-G^{\beta_1}(t_\circ))}} \frac{-rr'}{m\ ln(1-H^{\beta_2}(t_{\circ\circ}))} \times (1-z)^{\frac{-r'}{m\ ln(1-H^{\beta_2}(t_{\circ\circ}))}} dz.$$

The MLE of P, when X and Y belongs to the same family of distributions, i.e-., $G(\cdot) = H(\cdot)$, $\beta_1 = \beta_2$ and $t_\circ = t_{\circ\circ}$, is given by

$$\widehat{P} = \frac{r'n}{r'n+rm}. \tag{19}$$

4. Simulation Study

In order to validate the results obtained in Sections 2 and 3, we first consider the Kum distribution as a particular case of the Kum-G distributions. The pdf and cdf of the Kum distribution are given by:

$$f(x;\alpha,\beta) = \alpha\beta x^{\beta-1}(1-x^\beta)^{\alpha-1};\ 0<x<1,\ \alpha,\beta>0. \tag{20}$$

$$F(x;\alpha,\beta) = 1-(1-x^\beta)^\alpha. \tag{21}$$

respectively.

4.1. Simulation Based on Type II Censoring

For comparing the performances of estimators of α^q based on Type II censoring scheme, we have generated 1000 random samples from (20) each of size $n = 50$ for (α,β)=(2,0.5),

(2,1), (2,2). For each sample we arranged the data in ascending order and considered a sample of first r ($\leq n$) observations. For different values of r =10, 20, 30 and 50, we have computed average values of $\widetilde{\alpha_{II}^q}$ and $\widehat{\alpha_{II}^q}$, their corresponding bias, MSE and approximate 95% confidence interval. For $q = 1$ and 2, results are reported in Table 1. It has been observed that MSE obtained corresponding to UMVUE is much lower than MSE obtained corresponding to MLE. Thus, the performance of UMVUE of α^q for $q = 1, 2$ based on Type II censoring is much better than the performance of MLE of α^q. From Table 1, we observe that as r increases, the performance improves in the sense that their MSE decreases. It is also interesting to note that, with increasing r, the two estimators come close to each other.

For comparing the performance of MLE and UMVUE of reliability function $R(t)$, the bias, MSE and 95% confidence intervals are presented in Table 2. Comparing the estimates on the basis of MSE, we observe that the MLE of $R(t)$ performs better than the UMVUE for all parametric settings. As r increases, the performance of both the estimators improve and both estimators come close to each other.

For investigating the performance of estimators of P , we have generated 1000 random samples from each of the populations X and Y with sizes (n,m) with $\beta_1 = \beta_2 = 2$ and (α_1, α_2) = (0.5,0.5),(0.5,1), (0.5, 1.5) and (1.5,2). Samples corresponding to both the populations are arranged in ascending order and first (r, r') observations are considered. For $(r, r') = (10, 10), (20, 20), (30, 25), (40, 40)$ and $(50, 50)$, we have computed average values of $\widetilde{P}$ and $\widehat{P}$, their corresponding bias, MSE and approximate 95% confidence interval and results are presented in Table 3. We observe that for all selected values of (r, r'), the MLE of P performs superior to the UMVUE of P in the sense that it has lower MSE.

4.2. Simulation Based on Sampling Scheme of Bartholomew

In order to obtain point estimates of $R(t)$ based on the sampling scheme of Bartholomew, we have generated 1000 random samples each of size 100 from (20) with $\alpha = 2$ and $\beta = 0.9$. By fixing the termination time at $t_\circ$, and replacing the failure by operating one, values of r (the number of failures before time $t_\circ$) is computed. For different termination time $t_\circ$ = 0.20, 0.50, 0.65, 0.80 and 0.90, we have computed average values of $\widetilde{R(t)}$ and $\widehat{R(t)}$, their corresponding bias, MSE and approximate 95% confidence interval. For different values of t results are presented in Table 4. It has been observed that for small values of t and small values of $t_\circ$, MLE is more efficient than UMVUE of $R(t)$. However, for large values of $t_\circ$, UMVUE becomes more efficient than MLE of $R(t)$. For large values of t and all values of $t_\circ$, both the estimators become equally efficient. The best results are obtained for $t_\circ = 0.65$ as bias and MSE are least for all values of t. This result shows the importance of termination time $t_\circ$ in the sampling scheme of Bartholomew.

Now, to investigate the performance of estimators of P based on the sampling scheme of Bartholomew, we have generated 1000 random samples from each of the population X and Y with sizes (n, m) with $\beta_1 = \beta_2 = 2$ and (α_1, α_2) = (0.5,0.75), (0.5,1), (0.5, 1.5) and (1.5,2.5). For each sample corresponding to both the population, fixing the termination time at $t_\circ = t_{\circ\circ}$ and replacing the failure by operating one, values of r (no. of failures before time $t_\circ$ in X) and values of r' (no. of failures before time $t_{\circ\circ}$ in Y) are computed. For $t_\circ = t_{\circ\circ}$= 0.50, 0.70 and 0.80, we have computed average values of $\widetilde{P_I}$ and $\widehat{P_I}$, their corresponding

bias, MSE and approximate 95% confidence interval for $n > m$ and $n < m$, and results are presented in Tables 5 and 6 respectively. From Table 5, for $n > m$, it is observed that for small m when $n = 50$, UMVUE of P performs superior than MLE of P. As m increases both the estimators are equally efficient. However, for $n < m$, the results given in Table 6 show that for all n with $m = 50$, the MLE of P is superior than the UMVUE and, as n increases, both the estimators become equally efficient.

5. Real Data Study

In this section, to illustrate the usefulness of our procedure, we present real data analysis. We consider the real data set used by Kumari *et al.* (2019), originally taken from Proschan (1963). The data represent the intervals between failures (in hours) of the air conditioning system of a fleet of 13 Boeing 720 jet airplanes. Canavos and Tsokos (1971) observed that the failure time distribution of the air conditioning system for each of the planes can be well approximated by exponential distributions. We have considered the planes '7913' and '7914' for our illustrative purposes. The data are presented below:

x_1 (Plane 7914): 3, 5, 5, 13, 14, 15, 22, 22, 23, 30, 36, 39, 44, 46, 50, 72, 79, 88, 97,102, 139, 188, 197, 210.

y_1(Plane 7913): 1, 4, 11, 16, 18, 18, 18, 24, 31, 39, 46, 51, 54, 63, 68, 77, 80, 82, 97, 106, 111, 141, 142, 163, 191, 206, 216.

Before applying the Kolmogorov–Smirnov (KS) test, we transform the above given two data sets in the range of unit interval by using the transformation $X_i = \frac{X_i}{max(X_i)+1}$ and $Y_i = \frac{Y_i}{max(Y_i)+1}$.The two transformed data sets are given below:

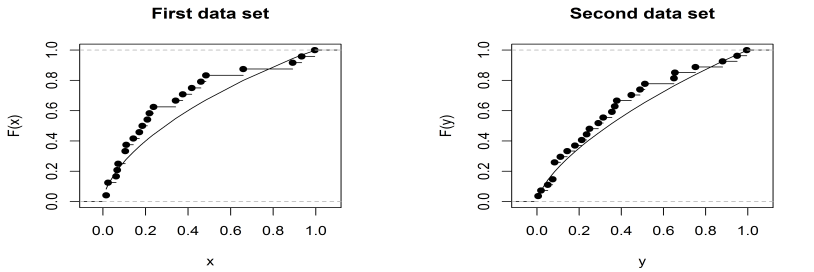
First data set x: (0.0142, 0.0237, 0.0616, 0.0664, 0.0711, 0.1043, 0.1090, 0.1422, 0.1706, 0.1848, 0.2085, 0.2180, 0.2370, 0.3412, 0.3744, 0.4171, 0.4597, 0.4834, 0.6588, 0.8910, 0.9336, 0.9953).

Second data set y: (0.0046, 0.0184, 0.0507, 0.0737, 0.0829, 0.1106, 0.1429, 0.1797, 0.2120, 0.2350, 0.2488, 0.2903, 0.3134, 0.3548, 0.3687, 0.3779, 0.4470, 0.4885, 0.5115, 0.6498, 0.6544, 0.7512, 0.8802, 0.9493, 0.9954).

We first apply the KS test to check whether the Kum distribution (20), fits the given X and Y populations. We obtain the following ML estimates of (α_1, β_1) and (α_2, β_2).

$$(\alpha_1, \beta_1)_{\text{complete data}} = (1.0728, 0.6022), \quad (\alpha_2, \beta_2)_{\text{complete data}} = (1.042, 0.6658).$$

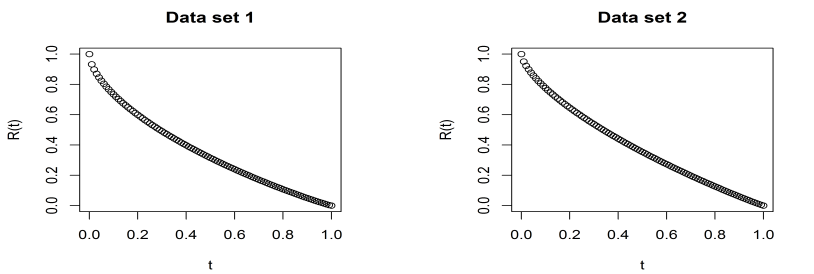
According to the KS test, we do not reject the null hypothesis that both the data observed for X $(KS = 0.18226; p = 0.4026)$ and the data observed for $Y(KS = 0.1289; p = 0.7604)$ are drawn from (20). Figure 1 confirms the good fit of (20), for these two data sets. In order to obtain the MLE of $R(t)$ and P based on Type II censoring, we first consider $r = 16$



(a) **Empirical and theoretical cdf of first data set**

(b) **Empirical and theoretical cdf of second data set**

Figure 1: **Plots of empirical and theoretical cdf**



(a) **Plot of ML Estimator of R(t) of first data set**

(b) **Plot of ML Estimator of R(t) of second data set**

Figure 2: **Plots of MLE of R(t)**

lifetimes from X population and the remaining 8 observations are considered as censored. Similarly, we consider first $r' = 20$ lifetimes from Y population and the remaining 7 observations are considered as censored. Considering the Kum distribution as a lifetime model for X-population, the MLEs of α_{1II} and β_{1II} are obtained as $\widehat{\alpha_{1II}} = 1.272$ and $\widehat{\beta_{1II}} = 0.6659$. Similarly, considering the Kum distribution as a lifetime model for Y-population, the MLEs of α_{2II} and β_{2II} are $\widehat{\alpha_{2II}} = 1.4677$ and $\widehat{\beta_{2II}} = 0.8128$. To evaluate MLE of P_{II} , we have considered the first data set as X-population and second data set as Y-population. We get $\widehat{P}_{II} = 0.5847$. For different values of t, we have evaluated MLE of $R(t)$ for X and Y populations, respectively. Results are plotted in Figure 2. In particular, for $t = 0.8$, $\widehat{R_{1II}}(t) = 0.1081$ and $\widehat{R_{2II}}(t) = 0.127$.

From Figure 2, it is clear that at initial time, the probability of survival is very high and as time increases the probability of survival decreases.

6. Conclusions

In this article, we have developed the estimation procedures for the Kum-G family of distributions based on Type II censoring and Bartholomew censoring schemes. Considera-

tions are given to both point and interval estimations. The finite sample performance of the UMVUE's and MLE's of reliability functions and other parameters are investigated using extensive Monte Carlo experiment. The comparisons are made on the basis of MSE of the estimators. The main conclusions of the simulation experiments are as follows.

For Type II censoring, for all values of n, the UMVUE of α^q performs better than MLE of α^q. On the contrary, the performance of MLE of $R(t)$ is better than the performance of UMVUE of $R(t)$ for all selected values of t. However, for large values of r, the performance of both the estimators is quite similar. Further, as r increases, MSE corresponding to both the estimator decreases. Similarly, for estimating P, the MLE performs superior than the UMVUE.

For the sampling scheme of Bartholomew, for small values of t and $t_\circ$, MLE is more efficient than UMVUE of $R(t)$. However, for large values of $t_\circ$, UMVUE becomes more efficient than MLE of $R(t)$. For large values of t and all values of $t_\circ$, both the estimators are almost equally efficient. The best results are obtained for $t_\circ = 0.65$ as the bias and MSE are least for all values of t. This result shows the importance of termination time $t_\circ$ in the sampling scheme of Bartholomew. For comparing the performance of MLE and UMVUE of P, we observe that, when $n = 50$ and $m < n$, UMVUE outperforms MLE. As m increases both the estimators become equally efficient. On the contrary, for $n < m$ and $m = 50$ for small n, MLE of P gives better performance than UMVUE. But as m increases both the estimators become almost equally efficient.

The paper focuses on developing classical estimators for different parameters and reliability functions of Kumaraswamy-G distributions under various sampling schemes and investigating their properties. However, an interesting alternative to MLE and UMVU estimators can be provided by the empirical Bayes approach or ML-II estimators based on the robust Bayesian approach of Shrivastava *et al.* (2019). We leave exploration of this area for future work.

Acknowledgement

The authors would like to thank the Editor and the anonymous reviewer(s) for their valuable comments and suggestions on an earlier version of the manuscript. Their suggestions have greatly improved the presentation of the manuscript.

References

Bartholomew, D. J., (1963). The sampling distribution of an estimate arising in life testing, *Technometrics*, 5, pp. 361–374.

Canavos, G., Tsokos, C. P., (1971). A Study of an Ordinary and Empirical Bayes Approach of Estimation of Reliability in the Gamma Life Testing Model. *Proceedings of IEEE Symposium on Reliability*, pp. 1–17.

Chaturvedi, A., Bhatnagar, A., (2020). Development of Preliminary Test Estimators and Preliminary Test Confidence Intervals for Measures of Reliability of Kumaraswamy-G

Distributions Based on Progressive Type-II Censoring. *Journal of Statistical Theory and Practice,* 14(3), pp. 1–28.

Chaturvedi, A., Tomer, S. K., (2003). UMVU estimation of the reliability function of the generalized life distributions. *Statistical Papers*, 44(3), pp. 301–313.

Cordeiro, G. M., De Castro, M., (2011). A new family of generalized distributions. Journal of Statistical Computation and Simulation, 81(7), pp. 883–898.

Dey, S., Mazucheli, J., Anis, M. S., (2017). Estimation of reliability of multicomponent stress–strength for a Kumaraswamy distribution. *Communications in Statistics - Theory and Methods*, 46(4), pp. 1560–1572, DOI: 10.1080/03610926.2015.1022457.

Dey, S., Mazucheli, J., Nadarajah, S., (2018). Kumaraswamy distribution: different methods of estimation. *Computational and Applied Mathematics*, 37(2), pp. 2094–2111. DOI: https://doi.org/10.1007/s40314-017-0441-1

Eldin, M. M., Khalil, N., Amein, M., (2014). Estimation of Parameters of the Kumaraswamy Distribution Based on General Progressive Type II Censoring. *American Journal of Theoretical and Applied Statistics*, 3(6), pp. 217–222, DOI: 10.11648/j.ajtas. 20140306.17

Fletcher, S.G., Ponnambalam, K., (1996). Estimation of reservoir yield and storage distribution using moments analysis. *Journal of Hydrology*, 182, pp. 259–275.

Ganji, A., Ponnambalam, K., Khalili, D.,Karamouz, M., (2006). Grain yield reliability analysis with crop water demand uncertainty. *Stochastic Environmental Research and Risk Assessment* , 20, pp. 259–277, DOI:https://doi.org/10.1007/s00477-005-0020-7.

Garg, M., (2009). On generalized order statistics from Kumaraswamy distribution. *Tamsui Oxford Journal of Mathematical Sciences*, 25(2), pp. 153–166.

Hassan, A. S., Sabry, M. A., Elsehetry, A. M., (2020). A New Family of Upper-Truncated Distributions: Properties and Estimation. *Thailand Statistician*, 18(2), pp. 196–214.

Johnson, N. L., & Kotz, S., (1970). Distributions in Statistics, vol. III: Continuous Univariate Distributions. New York: Houghton-Mifflin.

Jones, M. C., (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6(1), pp. 70–81, DOI: https://doi. org/10.1016/j.stamet.2008.04.001.

Kizilaslan, F., Nadar, M., (2016). Estimation and prediction of the Kumaraswamy distribution based on record values and inter-record times. *Journal of Statistical Computation*

and Simulation, 86(12), pp. 2471–2493, DOI: 10.1080/00949655.2015.111 9832.

Kumaraswamy, P., (1976). Sinepower probability density function. *Journal of Hydrology*, 31, pp. 181–184.

Kumaraswamy, P., (1978). Extended sinepower probability density function. *Journal of Hydrology*, 37, pp. 81–89.

Kumaraswamy, P., (1980). A generalized probability density function for double-bounded random process. *Journal of Hydrology*, 46, pp. 79–88.

Kundu, A., Chowdhury, S., (2018). Ordering properties of sample minimum from Kumaraswamy-G random variables. *Statistics*, 52(1), pp. 133–146, DOI: 10.1080/02 331888.2017.1353516.

Kumari T., Chaturvedi, A., Pathak, A., (2019). Estimation and Testing Procedures for the Reliability Functions of Kumaraswamy-G Distributions and a Characterization Based on Records. *Journal of Statistical Theory and Practice*, 13(1), DOI:10.1007/s42519-018-0014-7.

Mameli, V., Musio, M., (2013). A Generalization of the Skew-Normal Distribution: The Beta Skew-Normal. *Communications in Statistics - Theory and Methods*, 42(12), pp. 2229–2244, DOI: 10.1080/03610926.2011.607530.

Mameli, V., (2015). The Kumaraswamy skew-normal distribution. *Statistics & Probability Letters*, 104, pp. 75–81.

Nadar, M., Kizilaslan, F., Papadopoulos, A., (2014). Classical and Bayesian estimation of $P(Y < X)$ for Kumaraswamy's distribution. *Journal of Statistical Computation and Simulation*, 84:7, pp. 1505–1529, DOI: 10.1080/00949655.2012.750658.

Nadarajah, S., (2008). On the distribution of Kumaraswamy. *Journal of Hydrology*, 348(3-4), pp. 568–569, DOI: 10.1016/j.jhydrol.2007.09.008.

Nadarajah, S., Cordeiro, G. M., Ortega, E. M. M., (2012). General results for the Kumaraswamy-G distribution. *Journal of Statistical Computation and Simulation*, 82(7), pp. 951–979, DOI: 10.1080/00949655.2011.562504.

Patel, J. K., Kapadia, C. H., Owen D. B., (1976). Handbook of Statistical Distributions. *Marcel Dekker, New York.*

Ponnambalam, K., Seifi, A., Vlach, J., (2001). Probabilistic design of systems with general distributions of parameters. *International Journal of Circuit Theory and Applications*,

29, pp. 527–536, DOI: https://doi.org/10.1002/cta.173.

Proschan, F., (1963). Theoretical explanation of observed decreasing failure rate. *Technometrics*,15, pp. 375–383.

Rohatgi, V. K., Saleh, A. K. Md. E., (2012). An introduction to probability and statistics. Wiley, New York.

Seifi, A., Ponnambalam, K., Vlach, J., (2000). Maximization of Manufacturing Yield of Systems with Arbitrary Distributions of Component Values. *Annals of Operations Research* ,99, pp. 373–383, DOI: 10.1023/A:1019288220413.

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

Sindhu, T.N., Feroze, N., Aslam, M., (2013).Bayesian Analysis of the Kumaraswamy Distribution under Failure Censoring Sampling Scheme. *International Journal of Advanced Science and Technology*, 51, pp. 39–58.

Sundar, V., Subbiah, K., (1989). Application of double bounded probability density function for analysis of ocean waves. *Ocean Engineering*, 16(2), pp. 193–200.

Tamandi, M., Nadarajah, S., (2016). On the Estimation of Parameters of Kumaraswamy-G Distributions. *Communications in Statistics - Simulation and Computation*, 45(10), pp. 3811–3821, DOI: 10.1080/03610918.2014.957840.

Appendix

Table 1: **UMVUEs and MLEs of α^q for different values of β**

$r->$		**10**		**20**		**30**		**50**	
β	q	$\widetilde{\alpha^q}$	$\widehat{\alpha^q}$	$\widetilde{\alpha^q}$	$\widehat{\alpha^q}$	$\widetilde{\alpha^q}$	$\widehat{\alpha^q}$	$\widetilde{\alpha^q}$	$\widehat{\alpha^q}$
0.5	1	2.0374	2.2638	2.0153	2.1214	2.0136	2.083	1.998	2.0388
		0.0374	0.2638	0.0153	0.1214	0.0136	0.083	-0.002	0.0388
		0.5007	0.686	0.2287	0.2679	0.1554	0.173	0.0845	0.0895
		(1.994, 2.081)	(2.215, 2.312)	(1.986,2.045)	(2.09, 2.153)	(1.989, 2.038)	(2.058, 2.108)	(1.98,2.016)	(2.02, 2.057)
	2	4.0384	5.6089	4.0632	4.7523	4.0159	4.4511	4.0527	4.3077
		0.0384	1.6089	0.0632	0.7523	0.0159	0.4511	0.0527	0.3077
		11.1695	24.132	4.535	6.7641	2.6135	3.4138	1.4052	1.6791
		(3.831,4.246)	(5.321,5.897)	(3.931,4.195)	(4.598, 4.907)	(3.916, 4.116)	(4.34, 4.562)	(3.979,4.126)	(4.23,4.386)
1	1	2.0183	2.2425	2.0079	2.1136	2.0127	2.0821	1.9999	2.0407
		0.0183	0.2425	0.0079	0.1136	0.0127	0.0821	-1e-04	0.0407
		0.5733	0.7661	0.2212	0.2579	0.147	0.1639	0.0845	0.0897
		(1.971,2.065)	(2.19,2.295)	(1.979,2.037)	(2.083,2.144)	(1.989,2.036)	(2.058,2.107)	(1.982,2.018)	(2.022,2.059)
	2	4.1504	5.7644	4.0609	4.7496	4.0158	4.451	3.9988	4.2504
		0.1504	1.7644	0.0609	0.7496	0.0158	0.451	-0.0012	0.2504
		11.6885	25.6167	4.9774	7.3657	2.5661	3.3555	1.4462	1.6966
		(3.939, 4.362)	(5.47,6.058)	(3.923,4.199)	(4.588,4.911)	(3.917,4.115)	(4.341, 4.561)	(3.924, 4.073)	(4.171,4.33)
2	1	2.0028	2.2254	2.0186	2.1248	1.9962	2.065	2.0149	2.056
		0.0028	0.2254	0.0186	0.1248	-0.0038	0.065	0.0149	0.056
		0.4796	0.6429	0.2259	0.2655	0.1414	0.1555	0.0845	0.0909
		(1.96,2.046)	(2.178,2.273)	(1.989,2.048)	(2.094, 2.156)	(1.973, 2.019)	(2.041,2.089)	(1.997,2.033)	(2.038,2.074)
	2	3.9382	5.4697	3.9687	4.6417	3.9592	4.3883	3.9561	4.205
		-0.0618	1.4697	-0.0313	0.6417	-0.0408	0.3883	-0.0439	0.205
		10.4358	22.2834	3.726	5.5074	2.3075	2.9835	1.369	1.5866
		(3.738,4.138)	(5.192,5.748)	(3.849,4.088)	(4.502,4.782)	([1] 3.865,4.053)	(4.284,4.493)	(3.884,4.029)	(4.128,4.282)

Note: The 1st, 2nd, 3rd row represents the average estimates, average bias, MSE and the 4th row represents the confidence interval.

Table 2: **UMVUEs and MLEs of $R(t)$**

$r->$		**10**		**20**		**30**		**50**	
$t\downarrow$	$R(t)\downarrow$	$\widetilde{R(t)}$	$\widehat{R(t)}$	$\widetilde{R(t)}$	$\widehat{R(t)}$	$\widetilde{R(t)}$	$\widehat{R(t)}$	$\widetilde{R(t)}$	$\widehat{R(t)}$
0.15	0.3754	0.3773	0.3602	0.3762	0.3672	0.377	0.3709	0.3724	0.3688
		0.0019	-0.0152	8e-04	-0.0082	0.0016	-0.0045	-0.003	-0.0066
		0.0141	0.0128	0.0067	0.0064	0.0045	0.0044	0.0029	0.0029
		(0.37,0.385)	(0.353, 0.367)	(0.371,0.381)	(0.362,0.372)	(0.373,0.381)	(0.367,0.375)	(0.369, 0.376)	(0.365,0.372)
0.20	0.3056	0.3071	0.2941	0.3055	0.2986	0.3086	0.3038	0.3082	0.3053
		0.0016	-0.0115	0	-0.007	0.003	-0.0018	0.0027	-2e-04
		0.0134	0.0118	0.0061	0.0058	0.0045	0.0043	0.0024	0.0023
		(0.3,0.314)	(0.287,0.301)	(0.301,0.31)	(0.294,0.303)	(0.304,0.313)	(0.3,0.308)	(0.305, 0.311)	(0.302,0.308)
0.25	0.25	0.2512	0.2422	0.2536	0.2485	0.2489	0.2456	0.2477	0.2457
		0.0012	-0.0078	0.0036	-0.0015	-0.0011	-0.0044	-0.0023	-0.0043
		0.0124	0.0105	0.0058	0.0053	0.004	0.0038	0.0024	0.0023
		(0.244,0.258)	(0.236,0.249)	(0.249,0.258)	(0.244,0.253)	(0.245,0.253)	(0.242, 0.249)	(0.245,0.251)	(0.243,0.249)
0.30	0.2046	0.2058	0.2004	0.2031	0.2002	0.2062	0.2041	0.203	0.2018
		0.0012	-0.0041	-0.0015	-0.0044	0.0016	-5e-04	-0.0015	-0.0028
		0.0115	0.0097	0.0054	0.005	0.0034	0.0032	0.0022	0.0021
		(0.199,0.212)	(0.194,0.207)	(0.199,0.208)	(0.196, 0.205)	(0.203, 0.21)	(0.201,0.208)	(0.2,0.206)	(0.199,0.205)
0.4	0.1351	0.1371	0.1375	0.1344	0.1347	0.1344	0.1345	0.1366	0.1367
		0.002	0.0024	-6e-04	-4e-04	-7e-04	-6e-04	0.0016	0.0016
		0.0075	0.0062	0.0036	0.0033	0.0024	0.0023	0.0014	0.0013
		(0.132, 0.142)	(0.133,0.142)	(0.131,0.138)	(0.131,0.138)	(0.131,0.137)	(0.132,0.137)	(0.134,0.139)	(0.134,0.139)
0.6	0.0508	0.0501	0.0567	0.0505	0.054	0.0503	0.0527	0.0499	0.0513
		-7e-04	0.0059	-3e-04	0.0032	-5e-04	0.0019	-9e-04	5e-04
		0.0022	0.0021	0.0011	0.0011	8e-04	8e-04	4e-04	4e-04
		(0.047, 0.053)	(0.054,0.06)	(0.048,0.053)	(0.052,0.056)	(0.049,0.052)	(0.051,0.054)	(0.049, 0.051)	(0.05,0.053)

Note: The 1st, 2nd, 3rd row represents the average estimates, average bias, MSE and the 4th row represents the confidence interval.

Table 3: **UMVUEs and MLEs of P**

$\alpha_1->$	**0.5**		**0.5**		**0.5**		**1.5**	
$\alpha_2->$	**0.5**		**1**		**1.5**		**2**	
$P->$	**0.5**		**0.6666667**		**0.75**		**0.625**	
$(r,r')\downarrow$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$
(10,10)	0.4917	0.492	0.6673	0.6601	0.7521	0.7428	0.6231	0.6175
	-0.0083	-0.008	6e-04	-0.0066	0.0021	-0.0072	-0.0019	-0.0075
	0.0133	0.0121	0.0102	0.0095	0.0072	0.007	0.0118	0.0109
	(0.485,0.499)	(0.485, 0.499)	(0.661, 0.674)	(0.654,0.666)	(0.747,0.757)	(0.738,0.748)	(0.616,0.63)	(0.611, 0.624)
(20,20)	0.5004	0.5004	0.6643	0.6607	0.7485	0.7439	0.6252	0.6223
	4e-04	4e-04	-0.0024	-0.006	-0.0015	-0.0061	2e-04	-0.0027
	0.0062	0.0059	0.0053	0.0052	0.0037	0.0037	0.0055	0.0053
	(0.496,0.505)	(0.496,0.505)	(0.66,0.669)	(0.656,0.665)	(0.745,0.752)	(0.74,0.748)	(0.621, 0.63)	(0.618,0.627)
(30,25)	0.5	0.5008	0.666	0.6641	0.7492	0.7464	0.6249	0.6236
	0	8e-04	-6e-04	-0.0026	-8e-04	-0.0036	-1e-04	-0.0014
	0.0048	0.0046	0.0035	0.0034	0.0027	0.0027	0.0043	0.0041
	(0.496,0.504)	(0.497, 0.505)	(0.662, 0.67)	(0.66,0.668)	(0.746,0.752)	(0.743,0.75)	(0.621,0.629)	(0.62,0.628)
(40,40)	0.5028	0.5027	0.6695	0.6677	0.7499	0.7475	0.6243	0.6229
	0.0028	0.0027	0.0029	0.001	-1e-04	-0.0025	-7e-04	-0.0021
	0.003	0.0029	0.0025	0.0024	0.0017	0.0017	0.0027	0.0027
	(0.499,0.506)	(0.499,0.506)	(0.666,0.673)	(0.665,0.671)	(0.747, 0.752)	(0.745,0.75)	(0.621,0.628)	(0.62, 0.626)
(50,50)	0.4998	0.4998	0.6653	0.6638	0.7502	0.7483	0.6238	0.6227
	-2e-04	-2e-04	-0.0014	-0.0028	2e-04	-0.0017	-0.0012	-0.0023
	0.0024	0.0024	0.002	0.002	0.0014	0.0014	0.0022	0.0022
	(0.497,0.503)	(0.497, 0.503)	(0.663,0.668)	(0.661,0.667)	(0.748,0.753)	(0.746,0.751)	(0.621,0.627)	(0.62, 0.626)

Note: The 1st, 2nd, 3rd row represents the average estimates, average bias, MSE and the 4th row represents the confidence interval.

Table 4: **UMVUEs and MLEs of $R(t)$ based on the Sampling Scheme of Bartholomew**

$t_\circ->$		**0.20**		**0.50**		**0.65**		**0.80**		**0.90**	
$t\downarrow$	$R(t)\downarrow$	$\widetilde{R(t)}$	$\widehat{R(t)}$	$\widetilde{R(t)}$	$\widehat{R(t)}$	$\widetilde{R(t)}$	$\widehat{R(t)}$	$\widetilde{R(t)}$	$\widehat{R(t)}$	$\widetilde{R(t)}$	$\widehat{R(t)}$
0.15	0.3754	0.3545	0.3561	0.3555	0.3562	0.3661	0.3666	0.3994	0.3998	0.448	0.4483
		-0.0209	-0.0193	-0.0199	-0.0192	-0.0094	-0.0088	0.024	0.0244	0.0726	0.0729
		0.0019	0.0018	0.0015	0.0014	0.0011	0.0011	0.0017	0.0017	0.0063	0.0063
		(0.352,0.357)	(0.354,0.358)	(0.353,0.358)	(0.354,0.358)	(0.364,0.368)	(0.365,0.369)	(0.397,0.401)	(0.398,0.402)	(0.446,0.45)	(0.446,0.45)
0.20	0.3056	0.2839	0.2856	0.2836	0.2845	0.2974	0.298	0.3275	0.328	0.3794	0.3798
		-0.0217	-0.0199	-0.0219	-0.0211	-0.0082	-0.0075	0.0219	0.0224	0.0738	0.0742
		0.0019	0.0018	0.0014	0.0014	0.0011	0.0011	0.0016	0.0016	0.0066	0.0066
		(0.282,0.286)	(0.283,0.288)	(0.282,0.286)	(0.283,0.286)	(0.295,0.299)	(0.296,0.3)	(0.325, 0.33)	(0.326,0.33)	(0.377,0.381)	(0.378, 0.382)
0.25	0.25	0.2306	0.2325	0.2289	0.2298	0.2431	0.2438	0.2733	0.2738	0.3209	0.3213
		-0.0194	-0.0175	-0.0211	-0.0202	-0.0069	-0.0062	0.0233	0.0238	0.0709	0.0713
		0.0015	0.0015	0.0013	0.0012	9e-04	9e-04	0.0016	0.0016	0.0061	0.0062
		(0.228,0.233)	(0.23,0.235)	(0.227,0.231)	(0.228,0.232)	(0.241,0.245)	(0.242,0.246)	(0.271,0.275)	(0.272,0.276)	(0.319,0.323)	(0.319,0.323)
0.35	0.1668	0.1486	0.1507	0.1494	0.1505	0.1594	0.1602	0.1863	0.1869	0.2321	0.2326
		-0.0182	-0.0161	-0.0173	-0.0163	-0.0074	-0.0066	0.0195	0.0201	0.0653	0.0658
		0.0012	0.0012	0.001	9e-04	7e-04	7e-04	0.0011	0.0012	0.0051	0.0052
		(0.147, 0.15)	(0.149,0.153)	(0.148,0.151)	(0.149,0.152)	(0.158,0.161)	(0.159,0.162)	(0.185,0.188)	(0.185,0.189)	(0.23,0.234)	(0.231, 0.234)
0.45	0.1084	0.0951	0.0972	0.0962	0.0972	0.1044	0.1052	0.1245	0.1251	0.1619	0.1625
		-0.0133	-0.0112	-0.0121	-0.0111	-0.004	-0.0032	0.0161	0.0167	0.0536	0.0541
		7e-04	7e-04	6e-04	6e-04	5e-04	5e-04	8e-04	8e-04	0.0036	0.0036
		(0.094,0.097)	(0.096,0.099)	(0.095,0.098)	(0.096,0.099)	(0.103,0.106)	(0.104,0.106)	(0.123,0.126)	(0.124,0.126)	(0.16, 0.164)	(0.161,0.164)
0.55	0.0668	0.0572	0.059	0.0573	0.0582	0.0638	0.0646	0.0797	0.0803	0.1103	0.1109
		-0.0096	-0.0077	-0.0095	-0.0086	-0.0029	-0.0022	0.0129	0.0135	0.0436	0.0441
		4e-04	4e-04	3e-04	3e-04	2e-04	2e-04	5e-04	5e-04	0.0024	0.0024
		(0.056,0.058)	(0.058,0.06)	(0.056,0.058)	(0.057,0.059)	(0.063,0.065)	(0.064,0.066)	(0.079, 0.081)	(0.079, 0.081)	(0.109,0.112)	(0.11,0.112)
0.7	0.0267	0.0217	0.023	0.0217	0.0224	0.0249	0.0254	0.0347	0.0352	0.0526	0.0531
		-0.005	-0.0037	-0.0049	-0.0043	-0.0018	-0.0013	0.008	0.0085	0.0259	0.0264
		1e-04	1e-04	1e-04	1e-04	1e-04	1e-04	2e-04	2e-04	9e-04	9e-04
		(0.021,0.022)	(0.022,0.024)	(0.021,0.022)	(0.022,0.023)	(0.024,0.025)	(0.025,0.026)	(0.034,0.035)	(0.035,0.036)	(0.052,0.053)	(0.052,0.054)

Note: The 1st, 2nd, 3rd row represents the average estimates, average bias, MSE and the 4th row represents the confidence interval.

Table 5: **UMVUEs and MLEs of P based on the Sampling Scheme of Bartholomew**

$\alpha_1->$	**0.5**		**0.5**		**0.5**		**1.5**	
$\alpha_2->$	**0.75**		**1**		**1.5**		**2.5**	
$P->$	**0.6**		**0.6666667**		**0.75**		**0.625**	
$t_\circ = t_{\circ\circ}\downarrow$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$
$(n=50)>(m=35)$								
0.50	0.6011	0.5944	0.6535	0.6482	0.7314	0.7278	0.6001	0.5978
	0.0011	-0.0056	-0.0132	-0.0185	-0.0186	-0.0222	-0.0249	-0.0272
	0.0166	0.0169	0.0123	0.0128	0.0077	0.0079	0.0048	0.0049
	(0.593,0.609)	(0.586,0.602)	(0.647, 0.66)	(0.641,0.655)	(0.726,0.737)	(0.722,0.733)	(0.596,0.604)	(0.594,0.602)
0.70	0.5915	0.5885	0.6408	0.6383	0.7028	0.701	0.5763	0.575
	-0.0085	-0.0115	-0.0258	-0.0283	-0.0472	-0.049	-0.0487	-0.05
	0.0058	0.0059	0.0054	0.0055	0.0054	0.0056	0.0038	0.004
	(0.587,0.596)	(0.584,0.593)	(0.637,0.645)	(0.634,0.643)	(0.699,0.706)	(0.697, 0.704)	(0.574,0.579)	(0.573,0.577)
0.80	0.5781	0.5759	0.6235	0.6217	0.6784	0.6769	0.5602	0.5591
	-0.0219	-0.0241	-0.0431	-0.045	-0.0716	-0.0731	-0.0648	-0.0659
	0.0039	0.0041	0.0046	0.0048	0.0072	0.0074	0.0051	0.0053
	(0.574,0.582)	(0.572,0.58)	(0.62, 0.627)	(0.618,0.625)	(0.676,0.681)	(0.674, 0.68)	(0.558,0.562)	(0.557,0.561)
$(n=50)>(m=45)$								
0.50	0.5962	0.5945	0.6564	0.655	0.7261	0.7252	0.6012	0.6006
	-0.0038	-0.0055	-0.0103	-0.0116	-0.0239	-0.0248	-0.0238	-0.0244
	0.013	0.0131	0.012	0.012	0.0079	0.008	0.0044	0.0044
	(0.589,0.603)	(0.587, 0.602)	(0.65,0.663)	(0.648,0.662)	(0.721,0.731)	(0.72,0.731)	(0.597,0.605)	(0.597,0.604)
0.70	0.5893	0.5886	0.639	0.6383	0.7029	0.7024	0.5757	0.5753
	-0.0107	-0.0114	-0.0277	-0.0283	-0.0471	-0.0476	-0.0493	-0.0497
	0.0056	0.0056	0.0053	0.0053	0.0052	0.0052	0.0038	0.0038
	(0.585, 0.594)	(0.584, 0.593)	(0.635,0.643)	(0.634,0.642)	(0.699,0.706)	(0.699,0.706)	(0.573,0.578)	(0.573,0.578)
0.80	0.5781	0.5776	0.626	0.6255	0.6801	0.6797	0.5598	0.5596
	-0.0219	-0.0224	-0.0407	-0.0411	-0.0699	-0.0703	-0.0652	-0.0654
	0.0038	0.0038	0.0042	0.0042	0.0068	0.0068	0.0051	0.0052
	(0.575,0.582)	(0.574,0.581)	(0.623, 0.629)	(0.622, 0.629)	(0.677,0.683)	(0.677,0.682)	(0.558,0.562)	(0.558,0.561)

Note: The 1st, 2nd, 3rd row represents the average estimates, average bias, MSE and the 4th row represents the confidence interval.

Table 6: **UMVUEs and MLEs of P based on the Sampling Scheme of Bartholomew**

$\alpha_1->$	**0.5**		**0.5**		**0.5**		**1.5**	
$\alpha_2->$	**0.75**		**1**		**1.5**		**2.5**	
$P->$	**0.6**		**0.6666667**		**0.75**		**0.625**	
$t_\circ = t_{\circ\circ}\downarrow$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$	$\widetilde{P}$	$\widehat{P}$
$(n=35)<(m=50)$								
0.50	0.5958	0.6024	0.6516	0.6568	0.734	0.7375	0.6031	0.6055
	-0.0042	0.0024	-0.0151	-0.0099	-0.016	-0.0125	-0.0219	-0.0195
	0.0165	0.0164	0.0137	0.0135	0.009	0.0088	0.0054	0.0053
	(0.588,0.604)	(0.594,0.61)	(0.644,0.659)	(0.65,0.664)	(0.728,0.74)	(0.732,0.743)	(0.599,0.607)	(0.601, 0.61)
0.70	0.5859	0.5889	0.6374	0.6399	0.7034	0.7052	0.576	0.5773
	-0.0141	-0.0111	-0.0293	-0.0268	-0.0466	-0.0448	-0.049	-0.0477
	0.007	0.0069	0.0064	0.0062	0.0062	0.006	0.004	0.0039
	(0.581,0.591)	(0.584,0.594)	(0.633,0.642)	(0.635,0.644)	(0.699,0.707)	(0.701,0.709)	(0.573,0.578)	(0.575, 0.58)
0.80	0.5758	0.578	0.6241	0.626	0.678	0.6795	0.5567	0.5578
	-0.0242	-0.022	-0.0426	-0.0407	-0.072	-0.0705	-0.0683	-0.0672
	0.0043	0.0041	0.0054	0.0053	0.0078	0.0076	0.0056	0.0055
	(0.572,0.58)	(0.574,0.582)	(0.62,0.628)	(0.622,0.63)	(0.675,0.681)	(0.676,0.683)	(0.555, 0.559)	(0.556, 0.56)
$(n=45)<(m=50)$								
0.50	0.5949	0.5966	0.6594	0.6607	0.7375	0.7384	0.601	0.6016
	-0.0051	-0.0034	-0.0073	-0.0059	-0.0125	-0.0116	-0.024	-0.0234
	0.0153	0.0153	0.0117	0.0116	0.0082	0.0081	0.0043	0.0043
	(0.587,0.603)	(0.589,0.604)	(0.653,0.666)	(0.654,0.667)	(0.732,0.743)	(0.733,0.744)	(0.597,0.605)	(0.598,0.605)
0.70	0.5761	0.5764	0.6374	0.638	0.7013	0.7018	0.5769	0.5773
	-0.0489	-0.0486	-0.0293	-0.0286	-0.0487	-0.0482	-0.0481	-0.0477
	0.0039	0.0039	0.0056	0.0056	0.0055	0.0055	0.0036	0.0036
	(0.574, 0.579)	(0.574, 0.579)	(0.633, 0.642)	(0.634, 0.642)	(0.698,0.705)	(0.698,0.705)	(0.575,0.579)	(0.575,0.58)
0.80	0.577	0.5775	0.6254	0.6258	0.6791	0.6795	0.559	0.5593
	-0.023	-0.0225	-0.0413	-0.0408	-0.0709	-0.0705	-0.066	-0.0657
	0.004	0.004	0.0044	0.0044	0.0072	0.0071	0.0053	0.0053
	(0.573,0.581)	(0.574,0.581)	(0.622,0.629)	(0.623,0.629)	(0.676,0.682)	(0.677,0.682)	(0.557,0.561)	(0.557,0.561)

Note: The 1st, 2nd, 3rd row represents the average estimates, average bias, MSE and the 4th row represents the confidence interval.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 153–171, DOI 10.21307/stattrans-2022-009
Received – 14.12.2020; accepted – 13.10.2021

Long-term sovereign interest rates in Czechia, Hungary and Poland: a comparative assessment with an affine term structure model

Jakub Janus[1]

ABSTRACT

This paper provides a comparative evaluation of the behaviour of long-term sovereign yields in Czechia, Hungary and Poland from 2001 to 2019. An affine term structure model developed by Adrian, Crump and Moench (2013) is used as an empirical framework for the decomposition of the bond yields into term premium and risk-neutral components. We document a substantial compression in term premia which started in Central European economies around 2013 and played a decisive role in the changes that occurred in 10-year sovereign yields. This pattern, however, was more prevalent in Czechia and Poland than in Hungary. We show that long-term rates in all three economies remained higher than in Germany due to relatively large risk-neutral components. Nevertheless, cross-country correlations became increasingly dependent on term premium dynamics, both among Central European economies and between each of them and Germany. These results are robust to bias-correction in the baseline models and interpreted in the light of the general interest rates decline in the global economy. Potential policy implications are also discussed.

Key words: long-term interest rates, affine term structure model, term premium, risk-neutral rates, Central Europe.

1. Introduction

This paper investigates sovereign long-term interest rates in Czechia, Hungary, and Poland in an attempt to provide a detailed account of their behaviour from 2001 to 2019. Are the 10-year government rates mostly driven by a term premium component that investors demand for holding such securities, or by an expected path of short-term interest rates (i.e. risk-neutral rates)? Why did they decline by so much from the early 2000s, and so sharply after 2013? Is there a strong international comovement in long-term rates among the three economies, and how do they depend on foreign bond yields? The answers to these questions pose an empirical challenge but have weighty implications for both the broader economy and policy-making. Long-term yields and their components convey rich information about current and future states of an economy, for instance about expected trends in investment and consumption. They are, at the same time, a key variable for governments and central banks, with consequences for public debt management and monetary transmission mechanism (Gürkaynak and Wright, 2012). Central European (CE) economies constitute an interesting case study of long-term rates for two additional reasons. First, long-term rates are an essential channel of cross-border propagation of shocks, and the case of CE

[1]Cracow University of Economics. Department of Macroeconomics, Poland.
E-mail: jakub.janus@uek.krakow.pl. ORCID: https://orcid.org/0000-0002-2131-6077.

economies provides further evidence on the nature of comovement in interest rates and the impact of the international financial factors on a small open emerging economy. Second, 10-year sovereign rates are among the benchmarks for convergence between CE economies and the euro area. Hence, they remain relevant in the discussion on international economic and financial integration of these countries.

The empirical strategy of this paper builds upon an affine term structure model developed by Adrian, Crump and Moench (2013) (henceforth: ACM), a tractable method to perform a yield curve decomposition using a sequence of linear regressions. The estimation of this model makes use of a set of sovereign bond yields of various maturities to disentangle Czech, Hungarian, and Polish 10-year yields into risk-neutral components (expectations of the short-term rate) and term premia. Next, we employ these estimates to assess changes in long-term interest rates in each of the countries. Our study is among the first to provide a comparative analysis of 10-year treasury yields in all three CE economies using structural yield curve modelling and to present a comprehensive explanation of shifts in long-term rates and their cross-border linkages. As an extension to the main part of the analysis, we examine the small sample bias correction of the baseline ACM model.

We arrive at three main conclusions. First, we find a substantial term premium compression that started around 2013 in all three economies and brought those components to very low (Hungary) and possibly negative levels (Czechia and Poland). Second, we show that shifts in term premia components played a decisive role in the behaviour of 10-year sovereign yields. This pattern was more prevalent in Czechia and Poland than in Hungary but post-2009 the relative importance of term premium components increased in all three countries. Third, we demonstrate that long-term rates in CE economies were higher than in Germany, their leading economic and financial partner, due to relatively larger risk-neutral components. At the same time, cross-country correlations in the long-term yields were increasingly more dependent on the term premium dynamics, both among the three CE economies and between each of them and Germany.

The paper is organised as follows. The next section briefly discusses the research background and recent studies in the area. Section 3 presents the dataset of sovereign bond yields. Section 4 provides an overview of the affine term structure model used in the paper. The empirical outcomes are laid out in Section 5, while Section 6 discusses the results. The final section concludes.

2. Related literature

After the period of "benign neglect" (Turner, 2013), the long-term interest rates entered the centre of macroeconomic policy debate. Starting from the 2000s, sudden or unexpected shifts in those rates, exemplified by Greenspan's "conundrum" of 2004-2006 (Rudebusch, Sack and Swanson, 2007), drove more of the attention to macroeconomic effects of changes at the longer end of the yield curve. This concern intensified once new tendencies appeared in the global economy: persistently low inflation rates and inflation expectations, looming secular stagnation (Eggertsson, Mehrotra and Summers, 2016), prolonged quantitative easing and its subsequent tapering by the Federal Reserve (Kuttner, 2018), global shortages of safe assets (Caballero, Farhi and Gourinchas, 2017), or constraints to fiscal policy (Blan-

chard, 2019). As of more recently, markets got troubled with the dynamics of bond markets during the COVID-19 pandemic crisis.

From the point of view of small and open CE economies, the importance of long-term interest rates is further magnified by international linkages in bond yields (Obstfeld, 2015). The global financial cycle, whereby shocks in "central" economies, i.e. the US or the euro area, are transmitted to other countries, is consequential for capital flows, exchange rates, or monetary spillovers to such countries (Rey, 2016). This may also produce externalities for local bond markets and have an impact on domestic economies (Kolasa and Wesołowski, 2020). There is now accumulating evidence that a substantial part of spillovers from the ECB's monetary policy, both standard and non-standard one, is transmitted to CE economies via sovereign bond yields (Grabowski and Stawasz-Grabowska, 2020; Janus, 2020). Hence, the detailed evidence on the long-term interest rates is of sheer importance to policymakers in the CE countries, because exogenous movements in these yields may hamper the ability to influence interest rates on the longer end of the yield curve. As such, long-term rates are also worth investigating in the context of Czech, Hungarian, and Polish monetary policy effectiveness under financial globalization.

The empirical research on long-term bonds decomposition focuses mostly on the US and other advanced economies. The studies in this area often use dynamic term structure models and demonstrate that advanced economies exhibit similar patterns of changes in long-term interest rates, mostly due to comovement in term premium components that are subject to inflation shocks and international liquidity conditions (Wright, 2011). It is also shown that changes in long-term yields are amplified by global risk factors and international effects of monetary easing (Abbritti et al., 2018). Using international panel datasets that cover the CE economies, Mehrotra, Moessner and Shu (2019) and Albagli et al. (2019) both find that risk-neutral and term premium components of 10-year yields played a role in an intensification of monetary policy spillovers from the US to the rest of the world post-2009. However, their conclusions on the relative importance of those components for spillovers to advanced and emerging economies are mixed.

To the best of our knowledge, there is a shortage of comparative analyses on the long-term interest rates in Czechia, Hungary, and Poland that would employ empirical term structure modelling. Several studies deal with long-term yields through the lens of such a framework in individual CE economies. In the Czech case, Dvorák, Komárková and Kucera (2019) use an affine model along with the CDS quotations to retrieve components of long-term yields in Czechia. They attribute the post-crisis decline in these rates mostly to the portfolio effects, such as flight to quality. Interest rate expectations and term premium components in Hungary are analysed by Horváth et al. (2014), who show that these factors very often follow trends observed in emerging economies and international bond markets. Jabłecki, Raczko and Wesołowski (2016) employ the ACM term structure model for Poland and provide supporting evidence for a substantial decline in the term premium on 10-year bonds, which dips into negative territory after 2014. They explain these changes by the search-for-yield behaviour of foreign investors. Using an estimated DSGE model, Wesołowski (2018) presents the outcomes of a term structure decomposition and shows that term premium shocks have a significant impact on GDP in Poland, but their role in driving inflation is limited. In a recent contribution, Dec (2021) provides a detailed investigation

into the Polish sovereign bond market taking into account its limited liquidity and discusses specific properties of this market.

3. Government bonds yield curve data

In this paper, we focus on sovereign bonds, securities that are issued by respective central governments and remain comparable in terms of their risk properties. In an optimal setting, we would prefer to use a yield-curve dataset containing zero-coupon government bond price with as many maturities as possible, corresponding, for instance, to the acclaimed US database developed by Gürkaynak, Sack and Wright (2007) and maintained by the New York Fed. However, given the limited accessibility of such series, this study uses a dataset that consists of benchmark bond yields, following the practice of Albagli et al. (2019). The time series for each CE economy include securities with up to six maturities. For Czechia, they consist of 2, 3, 4, 5, and 10-year bonds, since data on 1-year bonds for this economy are not available until January 2009. Hungarian government bonds comprise those of 6 months, 1, 3, 5, and 10-year maturities. In this case, the time-series on 2-year bonds were not fully available and were swapped for the 6-month rate. Interest-rate series that we use for Poland have maturities of 1, 2, 3, 4, 5, and 10 years. Additionally, we include Germany as a benchmark case throughout the analysis, given the importance of this economy for CE countries. The dataset of German sovereign bonds ranges from 1 to 10 years. Data are obtained from Refinitiv Datastream.

Throughout the study, we use monthly frequency data, defined as the observation on the last trading day of the month, ranging from June 2001 to December 2019. All of the interest rates were firstly converted into continuously compounded rates. For each economy, the yield curve was fitted using the Nelson-Siegel-Svensson (NSS) framework (Svensson, 1994). Next, the estimated NSS parameters were used to retrieve yield curves for maturities from 1 to 120 months, which subsequently served as an input to the ACM term structure model. The yield curves that we obtain for each economy (not reported here) are generally upward-sloping. It must be noted, however, that their slope is inverted for the initial years of the analysis, especially in Hungary and to some extent in Poland, most likely due to the ongoing disinflation that still took place in these economies during that period, along with high and variable short-term rates. Toward the end of the sample, there are clear signs of the yield flattening, similar to the phenomena observed in the advanced economies (Joslin, Priebsch and Singleton, 2014).

4. An outline of the ACM affine term structure model

There are several leading approaches to separate various components of bond yields (see a comprehensive review by Rebonato, 2018). The essential decomposition breaks them down into two chunks, the expected short-term interest rate and a time-varying term premium, requested by an investor to compensate for an additional risk of holding a security for an extended period of time. However, given that bond components are not observable and depend on the relationship between yields on assets of different maturities, we need

a term structure model to capture the underlying factors driving bond yields. The modern workhorse of the macro-financial literature that serves this purpose is a (Gaussian) affine term structure model. This class of models is based on the assumption that observed yields and their expectation components can be expressed as affine functions of several risk factors, which summarise the term structure of interest rates and possibly other economic or financial variables. Interestingly enough, even though there is no uniform methodology to conceptualize and estimate such models, competing approaches largely concur on trends and dynamics of long-term interest rate components in the US and the euro area, as shown by Cohen, Hördahl and Xia (2018).

An important example of the dynamic term structure model was introduced by ACM. It is a computationally efficient and parsimonious technique to build a series of linear regressions and estimate pricing factors, which, in turn, are used to retrieve ex-ante neutral rates and term premia for various maturities of bonds. The ACM model is based on the notion that arbitrage opportunities are absent in the bond market. Under no-arbitrage, the pricing kernel M_t is defined through risk-adjusted future pay-offs generated by an asset:

$$P_t^{(n)} = E_t \left[M_{t+1} P_{t+1}^{(n-1)} \right] \tag{1}$$

where $P_t^{(n)}$ is a price of a security (bond) with maturity of n. In order to approximate the prices of risk, $K = 5$ principal factors (i.e. risk factors) are extracted from demeaned yields. This practice goes back to the work of Litterman and Scheinkman (1991), where the first three factors are intuitively interpreted as level, slope, and curvature of the yield curve. The factors are assumed to follow a VAR(1) process:

$$X_{t+1} = \mu + \Phi X_t + v_{t+1} \tag{2}$$

with normally distributed shocks to the state variables, $v_{t+1} \sim N(0, \Sigma)$. The stochastic discount factor, on the other hand, is expressed using r_t, the continuously compounded risk-free rate, and the price of risk, given by λ_t:

$$M_{t+1} = \exp(-r_t - \frac{1}{2}\lambda_t'\lambda_t - \lambda_t'\Sigma^{-1/2}v_{t+1}) \tag{3}$$

What is important in this type of model is that the price of risk is assumed to have an affine form, as in an influential work by Duffee (2002):

$$\lambda_t = \Sigma^{-1/2}(\lambda_0 + \lambda_1 X_t), \tag{4}$$

while ex-ante excess returns for a given yield of maturity n are given by:

$$rx_{t+1}^{(n-1)} = \ln P_{t+1}^{(n-1)} - \ln P_t^{(n)} - r_t, \qquad n = 2, \ldots, N, \tag{5}$$

with r_t also defined as an affine function of risk factors, $r_t = \delta_0 + \delta_1 X_t$.

The major contribution of ACM comes with a conceptual approach to the estimation procedure of λ_t based on factors in X_t and innovations to the VAR process. This procedure

consists of three steps. In the first step, the VAR system from Equation (2) is estimated by OLS. Residuals are collected in matrix $\hat{V}$, and the variance-covariance matrix is calculated as $\hat{\Sigma} = \hat{V}\hat{V}'/T$.

The second step starts by stacking the excess returns for all maturities and periods in an $N \times T$ matrix rx. Here, based on the fitted NSS curve, the $n = 1$ month yield is taken as a risk-free rate (r_t) of the model, and excess returns are calculated for $N = 12$ maturities of $n = 6, 12, 18, \ldots, 60, 84, 120$ months. In this stage, the ACM approach makes use of the basic assumption that eventually allows us to retrieve the neutral rate and term premium. Namely, it states that the excess returns may be decomposed into four parts: (a) expected return, (b) convexity adjustment, (c) priced return innovation, and (d) return pricing error. Hence, using a form stacked across maturities and time, it is indicated to run the following regression:

$$rx = \alpha + \beta' V + cX_{-} + E, \tag{6}$$

where α is a constant, V contains contemporaneous innovations to the VAR, and X_{-} denotes lagged factors. Next, the error variance is calculated as $\hat{\sigma}^2 = tr(\hat{E}\hat{E}'/NT)$, based on an $N \times T$ matrix of errors, $\hat{E}$.

In the third step of the procedure, we use the outcome we obtained so far and calculate the market prices of risk from Equation (4). The first of them is given as:

$$\hat{\lambda}_0 = (\hat{\beta}\hat{\beta}')^{-1}\hat{\beta}\left(\hat{a} + \frac{1}{2}\left(\hat{B}^* vec(\hat{\Sigma}) + \hat{\sigma}^2\right)\right), \tag{7}$$

where $\hat{B}^* = [vec(\beta^{(1)}\beta^{(1')}) \ldots vec(\beta^{(N)}\beta^{(N')})]$ is an $N \times K^2$ matrix. Each of its elements $\beta^{(n)}$ come from a $K \times N$ matrix $\hat{\beta}$ (the matrix contains coefficients on $\hat{V}$ in Equation (6) for a given maturity n). The second price of risk is derived as:

$$\hat{\lambda}_1 = (\hat{\beta}\hat{\beta}')^{-1}\hat{\beta}\hat{c}. \tag{8}$$

Finally, once $\hat{\lambda}_0$ and $\hat{\lambda}_1$ are estimated, we retrieve the sovereign yield curve in a recursive way. This part of the modelling strategy follows a general rule applied in affine term structure models to perform yield decomposition (Rebonato, 2018). Each bond yield is expressed again as an affine process, $\hat{y}_t(n) = -\frac{1}{n}(A_n + B_n' X_t)$, and we denote factor loadings for excess returns as $\hat{\beta}^{(n)} = B_n'$. Altogether the system takes the form of two recursions and two initial conditions:

$$A_n = A_{n-1}' + B_{n-1}'(\mu - \lambda_0) + \frac{1}{2}(B_{n-1}'\Sigma B_{n-1}') - \lambda_0, \tag{9a}$$

$$B_n' = B_{n-1}'(\Phi - \lambda_1) - \delta_1', \tag{9b}$$

$$A_0 = 0, \quad B_0 = 0. \tag{9c}$$

The risk-neutral component, i.e. the expected short-term interest rate, is extracted by setting both market prices of risk λ_t to zero, because it is best described as a yield that does not include any compensation for risk. The term premium component, in turn, is defined as a

difference between the bond yield implied by the model and the neutral rate:

$$y_t^P(n) = \hat{y}_t(n) - y_t^N(n). \tag{10}$$

As we aim at studying in detail only the decomposition of 10-year bond yields, we denote $y_t^P(120) = y_t^P$ and $y_t^N(120) = y_t^N$ in the remainder of the paper. It must be highlighted here that y_t^P may be treated in a sense broader than mere term premium (duration or interest-rate risk), as it gathers deviation from the neutral rate that may originate from various financial and economic shocks, such as shifts in consumption growth or inflation. Additionally, we calculate a long- and short-term interest rate spread that approximates the yield curve slope, y_t^S, based on the difference between the observed 10-year and 2-year bond yields (1-year rate for Hungary). We use the term spread as a simple reference point for our term premium estimates.

5. Empirical results

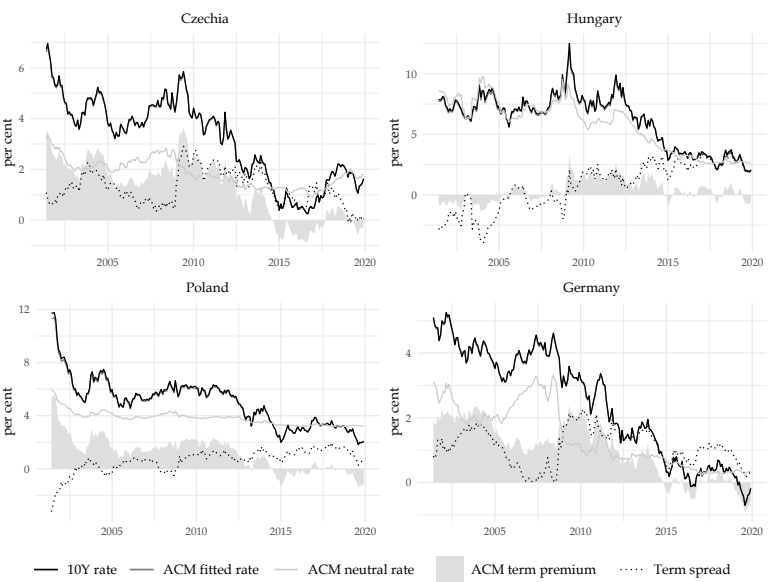
This section presents empirical results of the study. First, we explore the outcome of the estimated pricing models and examine changes in sovereign bond yields using the resulting decompositions. Next, we investigate international comovement in long-term rate components, both among the CE economies and between each of them and Germany. Finally, as a robustness check, we consider the bias-corrected outcomes of the ACM model.

5.1. Term premium and risk-neutral rate measures

The actual and fitted 10-year treasury bond yields, along with their components in all three CE economies and Germany, are depicted in Figure 1. It must first be noted that the ACM models produce a high-quality fit to 10-year yields. The series of actual and fitted yields are almost identical, which indicates small errors relative to the values of the modelled interest rates, just as in the ACM study on the US data. Starting with the estimates of the risk-natural rates, we observe notable differences in these components among the three CE economies. In Hungary, the neutral rate was considerably higher, especially in the 2000s, and more variable than in Czechia and Poland. A significant downward movement that began in 2009-2010 brought this rate in Hungary from around 9% in 2009 to 3% in 2015. Conversely, the initial neutral rate in Czechia was relatively low throughout the entire timespan. The country experienced a noticeable decline in this component between June 2008 and September 2013 (from 2.77% to 1.20%), but the neutral rate increased again post-2017. In Poland, the most prominent feature is the flattening of the neutral rate series in the latter part of the sample. After a decline from around 6% in the early 2000s, it quickly hit a plateau of around 4% in 2005, and by a slow downward movement it reached 3.22% in December 2019. Finally, the German neutral rate moved between 2% and 3% before the crisis. It approached 0.55% in March 2013, going down from the highest value of 3.31% in June 2008.

Term premia, the differences between actual and neutral rates, exhibit similar patterns in Czechia and Poland. First, the average value of this component throughout 2001-2019

was equal to 1.84% in the former economy, and 1.71% in the latter. The term premium in Poland, however, generally made up to a much smaller fraction of the 10-year yield. For example, in January 2006 the premium was equal to around 24% of the actual rate in this economy, while in Czechia this fraction stood at ca. 39%. Second, around 2008-2009 there was a clear upswing in premia in both economies (although a more pronounced one in Czechia), which may be understood as an increase in compensation for risk of holding Czech and Polish 10-year bonds during the onset of the financial and economic crisis. The same is true for Germany in this period. The term premium remained relatively high and positive in Czechia and Poland until 2012 when it started declining rapidly. In Poland, it hit zero in July 2014, and in Czechia a month later. From 2015 to 2019, its mean values were negative. Spreads between 10-year and 2-year bonds were generally lower than the term premium before 2013 and higher afterwards. There are some periods, such as 2008-2009 and 2014-2015, when the term spread moved in the same direction as the model-implied term premium.



Notes: risk-neutral rates and term premia estimated in the ACM term structure model; term spread calculated as a difference between 10-year and 2-year bond yields.

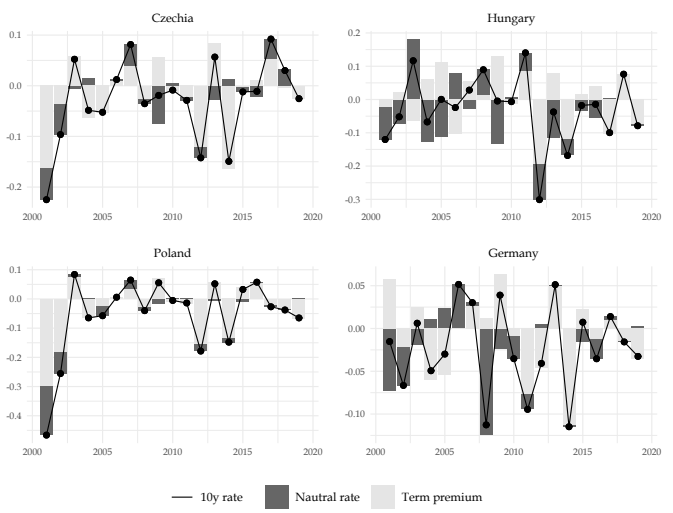
Figure 1. 10-year sovereign yields and their risk-neutral and term premium components

Due to generally high natural rates, the picture is different for Hungary. Up to 2010, the behaviour of the ACM term premium and the term spread was quite unlike in Poland or Czechia. The expected path of short term rates in Hungary was higher than the 10-year yield for most of the period from 2001 to 2009. Therefore, the term premium in this economy remained predominantly negative, going as low as -1.5% in 2003, which must be considered an anomaly. It may be a consequence of the fact that the sovereign yield curve in this economy was inverted for a relativity long period, compared to Czechia and Poland. In Czechia, this phenomenon did not occur, while in Poland, it was only detected

at the very beginning of the sample. In models re-estimated using yield series covering the post-2003 period, when inflation rates in CE economies further declined, the behaviour of term premia is highly comparable to the baseline estimates for all economies (those results are available upon request). The connections between term premium and observed rates in Hungary became more similar to that in Czechia and Poland from 2009 onwards. The term premium hiked to roughly 2.5% in 2008-2009, and then, around 2014-2015, it significantly declined, reaching very low and negative values in the last five years of the analysis.

5.2. Changes in long-term yields and their components

To get a closer look into long-term yields dynamics in CE economies, we first obtain month-to-month differences in long-term interest rates and their components. We next graph average annualised changes in 10-year yields, decomposed into risk-neutral rates and term premia (Figure 2). This decomposition indicates that in Czechia and Poland, changes in actual yields revealed a stronger connection to term premium dynamics. Most of the time, term premia were also decisive for the direction in which the 10-year rates moved in a given year. In fact, large swings in those yields in Czechia and Poland, such as ones observed between 2012 and 2014, were driven specifically by term premia. On the other hand, the role of the neutral rates was more pronounced for Hungary and it became less evident only post-2014. What seems distinct for Hungary among the CE economies, but also for Germany before 2010, is that there were numerous periods when changes in both components had inverse signs. For example, throughout 2002-2007 the risk-neutral rate and term premium dragged Hungarian 10-year rate in opposite directions, largely offsetting their individual contributions to the actual bond yield.



Notes: annualised dynamics calculated as average monthly changes in respective rates and their components during a given year.

Figure 2. Decomposition of annual changes in the 10-year sovereign yields

The variance shares of respective 10-year yield components, calculated as in Moench (2018), supplement our previous observations (Table 1). In Czechia and Poland, the ratio of changes in neutral rates to 10-year rates $(\Delta y_t^N, \Delta y_t)$ in the entire sample equalled 12.82% and 15.16%, respectively. This ratio decreased in the second subsample (2009-2019), and for Czechia it went as low as 4.78%. At the same time, the variance shares of term premium Δy_t^P in Δy_t increased, and this component became a visibly stronger driver of interest rates. In Czechia, the term premium variance shares were larger than the corresponding value for neutral rate by the factor of 19.5, and in Poland by 11.5. However, the most evident shift is observed for Hungary, where the variance ratio for term premium boosted from 21.86% to 77.60%. The variance shares of the term spread in sovereign bond yields $(\Delta y_t^S, \Delta y_t)$ were comparatively lower than their term premium counterparts. This ratio was highest for Hungary in the second subperiod but generally low in the Polish bond market, which speaks to the disconnect of those two series, especially before 2009. Post-2009, when the term spread shares categorically increased, their behaviour confirms a stronger impact of risk premia on bond yields.

Table 1. Changes in 10-year yields and their components: variance shares and correlations

		Variance shares			Correlations		
		$\Delta y_t^N, \Delta y_t$	$\Delta y_t^P, \Delta y_t$	$\Delta y_t^S, \Delta y_t$	$\Delta y_t^N, \Delta y_t^P$	$\Delta y_t^N, \Delta y_t^S$	$\Delta y_t^P, \Delta y_t^S$
2001-2019	Czechia	0.1282	0.8527	0.4739	-0.0850	-0.5501***	0.7937***
	Hungary	0.4112	0.5784	0.3611	-0.0159	-0.4177***	0.8846***
	Poland	0.1516	0.8208	0.1651	0.5069***	-0.2878***	0.3826***
	Germany	0.3168	0.6754	0.3074	-0.0684	-0.5103***	0.8149***
2001-2008	Czechia	0.2843	0.6914	0.1457	0.2684**	-0.5063***	0.5574***
	Hungary	0.7599	0.2186	-0.2027	-0.2347**	-0.7011***	0.8231***
	Poland	0.1959	0.7658	-0.0368	0.6769***	-0.3877***	0.0613
	Germany	0.5404	0.4461	-0.0140	-0.1329	-0.6932***	0.7245***
2009-2019	Czechia	0.0478	0.9362	0.6421	-0.2217**	-0.5483***	0.8638***
	Hungary	0.2220	0.7760	0.6735	0.1572*	-0.0459	0.9563***
	Poland	0.0792	0.9126	0.4894	0.2639***	-0.0771	0.7241***
	Germany	0.1675	0.8263	0.5150	0.0414	-0.2156**	0.8862***

Notes: 10-year yields variance shares of the ACM natural rate, term premium, and 10-year over 2-year spread are given as $\frac{Cov(\Delta y_t, \Delta y_t^N)}{Var(\Delta y_t)}$, $\frac{Cov(\Delta y_t, \Delta TP)}{Var(\Delta y_t)}$, and $\frac{Cov(\Delta y_t, \Delta y_t^S)}{Var(\Delta y_t)}$, respectively; ***, **, and * indicate significance of Pearson linear correlation coefficient estimates at the 0.1, 0.05, and 0.01 levels, respectively.

As we turn to correlations of long-term interest rates components, it must first be noted that in the entire sample coefficients calculated between changes in estimated risk-neutral rates and term premia $(\Delta y_t^N, \Delta y_t^P)$ are statistically significant only for Poland. In this case, the term premium was typically decreasing while the neutral rate was also going down. However, the opposite regularity of negative correlation between $(\Delta y_t^N$ and $\Delta y_t^P)$ entails in Hungary in the first subsample and in Czechia post-2009. Neutral rates and yield spreads $(\Delta y_t^N, \Delta y_t^S)$ were generally inversely correlated, even though this correlation is not strong. Negative correlations indicate in this case that downward shifts in the neutral rate were connected to an increase in the slope of sovereign yield curves. Term premia and term spreads $(\Delta y_t^P, \Delta y_t^S)$ mostly changed in the same direction. This relationship, however, was weaker pre-2009, for all economies but most notably for Poland.

5.3. International comovements in long-term interest rates

In order to investigate international linkages in long-term yields, we first calculate average cross-country differences in 10-year yields and their components pre- and post-2009

(Table 2). The two-sided t-tests for differences in means indicate an interesting relationship. Distances in actual bond yields between the CE economies remain almost the same in two subsamples, with statistically insignificant differences in mean values. In Czechia, the 10-year rate was on average ca. 3 points lower than in Hungary and ca. 1.9 points lower than in Poland, while Hungarian yields were around 1.1 point higher than Polish ones. At the same time, Hungary reduced its distance to Czechia and Poland both when it comes to the neutral rate and the term premium. On the other hand, for Czechia and Poland, the differences in term premium components vis-à-vis Germany decreased, but the distance between the neutral rates significantly widened. For the Hungary-Germany pair, the opposite was true.

Table 2. Mean values of differences in 10-year bond yields and their components

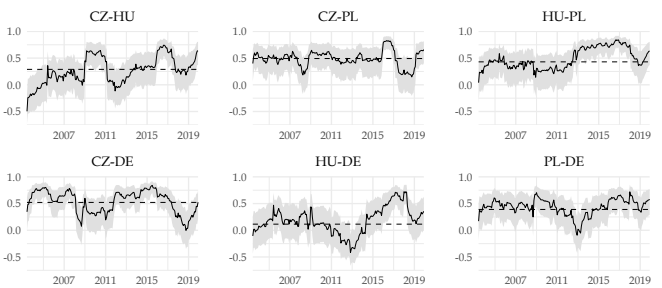
		y_t	y_t^N	y_t^P	y_t^S
Czechia - Hungary	2001-2008	-3.0078	-5.1967	2.2842	2.3090
	2009-2019	-3.0995	-3.0048	-0.0303	-0.7084
	t-test	0.5333	0.0000	0.0000	0.0000
Czechia - Poland	2001-2008	-1.9189	-1.8369	-0.0072	1.1753
	2009-2019	-1.8600	-2.0258	0.1944	0.2499
	t-test	0.5859	0.0000	0.0076	0.0000
Czechia - Germany	2001-2008	0.3463	-0.0738	0.4015	0.1710
	2009-2019	0.9730	0.7956	0.1618	0.1421
	t-test	0.0000	0.0000	0.0000	0.5506
Hungary - Poland	2001-2008	1.0889	3.3598	-2.2913	-1.1337
	2009-2019	1.2395	0.9791	0.2247	0.9583
	t-test	0.4289	0.0000	0.0000	0.0000
Hungary - Germany	2001-2008	3.3541	5.1228	-1.8827	-2.1380
	2009-2019	4.0725	3.8004	0.1922	0.8505
	t-test	0.0000	0.0000	0.0000	0.0000
Poland - Germany	2001-2008	2.2652	1.7630	0.4086	-1.0043
	2009-2019	2.8330	2.8214	-0.0326	-0.1078
	t-test	0.0000	0.0000	0.0000	0.0000

Notes: y_t, y_t^N, y_t^P, and y_t^S denote country differences in average values of actual 10-year sovereign rates, neutral rate, term premium, and term spread, respectively; p-values of the two-sided t-test; null hypothesis: mean values are equal between subsamples.

Next, we analyse correlations in bond yields and their components among countries. The rolling correlations that we obtain for changes in the term premia differ substantially, both across pairs of countries and in time (Figure 3). The correlation coefficient between Czechia and Poland was positive, fairly stable (in the whole sample: 0.49), and predominantly significant in the rolling window of 24 months. In pairs Czechia-Hungary and Hungary-Poland, we observe an upward trend in point estimates of the coefficient. In the former case, however, its values were on average smaller and increased in two periods: 2009-2011 and 2016-2017. The whole-sample correlation in term premia was comparably high and significant for Czechia-Germany (0.52) and Poland-Germany (0.39). In both cases, the coefficient declined around the global financial crisis (GFC), although visibly earlier for Czechia (around 2009) than for Poland (around 2011). Above all, developments in those correlations are consistent with comovements in 10-year yields. For Hungary-Germany, the correlation was hardly significant (in the whole sample, p-value of 0.089), and the confidence band included zero right up to 2013-2015 when it moved into positive territory where is stayed until 2018.

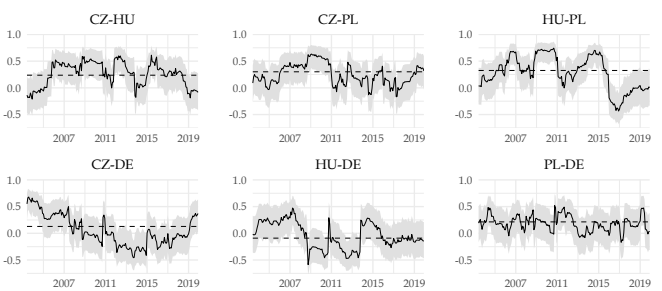
Figure 4 summarizes correlations of monthly changes in risk-neutral rates. In general, when it comes to correlations among the CE economies, the estimates are notably lower

than the previously reported values and range from 0.24 (Czechia-Hungary) to 0.33 (Poland-Hungary; both p-values of 0.00). The rolling correlations became larger from 2005 to 2010 (Czechia-Hungary and Czechia-Poland), and from 2013 to 2015, with a brief period of significantly negative values around 2015-2016 (Hungary-Poland). Corresponding correlation coefficients with Germany were considerably smaller. In the entire sample, the point estimate of the coefficient was highest for Poland (0.21, p-value 0.00), lower for Czechia (0.13, p-value 0.06), and close to zero for Hungary. Additionally, there was a clear negative trend in the rolling correlations for Czechia-Germany and Hungary-Germany pairs, at least until 2013-2014.



Notes: correlation calculated using month-to-month differences in term premia estimated in the ACM term structure model; solid line: point estimates of rolling correlation coefficient (window equal to 24 months); shaded area: 90-percent confidence interval; dotted line: correlation coefficient calculated for the entire timespan.

Figure 3. Cross-country correlations: changes in term premia



Notes: correlation calculated using month-to-month differences in risk-neutral rates estimated in the ACM term structure model; see Figure 3.

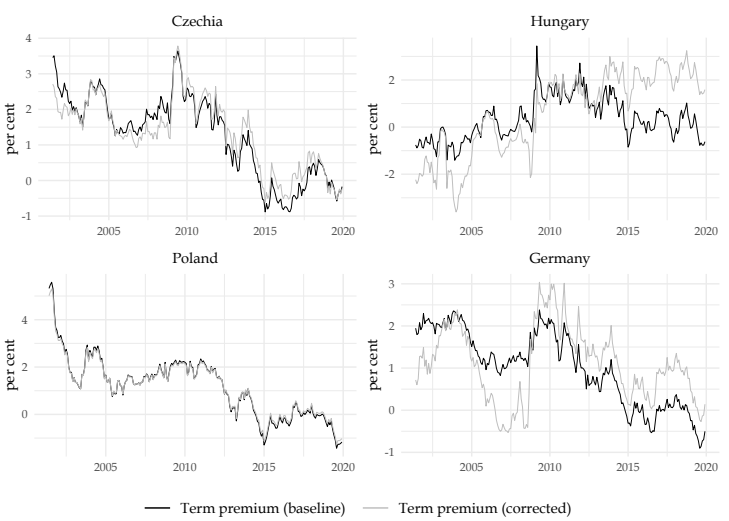
Figure 4. Cross-country correlations: changes in risk-neutral rates

5.4. Bias-corrected estimates

In the baseline ACM term structure model, the variables used to retrieve risk-neutral rates are derived from principal components of bond yield series. These components, in turn, are assumed to jointly follow an auto-regressive process given in Equation (2). However, as documented by Bauer and Hamilton (2012), such a VAR model estimated on a set of

interest rates may be subject to small-sample bias. This bias is mainly related to high persistence – in our case, a downward trend – in interest rate series used to extract risk factors. In consequence, the baseline procedure may overestimate the term premium components and, equivalently, underestimate changes in neutral rates in the underlying bond yields. To ease this problem, we employ the stationarity adjustment correction for VAR models developed by Kilian (1998). In a nutshell, this procedure aims at eliminating unit or explosive roots in auto-regressive models by discarding the estimated models that reveal such characteristics. Based on this approach, we estimate the bias-corrected ACM model using 10000 bootstrap replications. Resulting term premium estimates are depicted in Figure 5.

Noticeably, in the Polish case, the term premium series (and risk-neutral rates) obtained from the bias-corrected model turn out to be almost identical to the one from the baseline model. The bias-correction, however, has an impact on the decomposition of long-term rates for all remaining economies. There are two main differences with respect to the initial models. First, risk-neutral rates become somewhat lower and more variable for Czechia, Hungary, and Germany, especially after 2009. On average, they decrease by 0.3 points for Czechia, and by 1.1 points for Hungary. Interestingly, the negative values that we obtain for Germany from 2015 are now more consistent with the evidence on the natural rate of interest in the euro area presented by Holston, Laubach and Williams (2017). Second, the term premium estimates are generally lower pre-2009 and higher afterwards. Even though they are not necessarily negative after 2013, they still reach comparably low levels in Czechia and Germany, and their general tendencies are unchanged relative to the baseline estimates. In Hungary post-2013, the term premium fluctuates as much as 1.5 points above the baseline outcomes.



Notes: term premia estimated in the ACM term structure model; bias correction based on Kilian (1998) with bootstrap using 10000 replications.

Figure 5. Model-implied term premium series: baseline and bias-corrected estimates

Next, we inspect whether the correction of the ACM model substantially influences the results on the international comovement in long-term interest rates (Table 3). As expected, the role of term premia diminishes in the corrected model at the expense of neutral rates. However, correlations in term premia remain visibly more important than comovements in neutral rates for Czechia-Poland, Czechia-Germany, and Poland-Germany pairs. Corresponding correlations between Hungary and Germany are even weaker than in the baseline model. Correlations in pairs Hungary-Poland and Czechia-Hungary stay on similar levels when it comes to the neutral rate, but are now lower for the term premium, again speaking for a distinctive character of the Hungarian sovereign bond market. In sum, however, these results suggest that our baseline estimates are subject to uncertainty, but the common patterns in their comovements identified before remain uninterrupted.

Table 3. Cross-country correlations between the risk-neutral and term premia components: baseline vs. bias-corrected ACM model

	Baseline		Bias-corrected	
	Δy_t^N	Δy_t^P	Δy_t^N	Δy_t^P
Czechia - Hungary	0.2368***	0.2892***	0.2320***	0.1606**
Czechia - Poland	0.3006***	0.4920***	0.2821***	0.4156***
Czechia - Germany	0.1288*	0.5215***	0.1332**	0.3493***
Hungary - Poland	0.3255***	0.4297***	0.3412***	0.2341***
Hungary - Germany	-0.0904	0.1145*	-0.1059	0.0400
Poland - Germany	0.2138***	0.3898***	0.1520**	0.2658***

Notes: see Table 1.

6. Discussion

Based on the empirical results reported in the previous section, this part of the paper discusses our major findings. In the first place, our results signify the role played by term premium components in the cross-border transmission of changes in long-term sovereign rates. For all three CE economies, we show that a higher dependence of bond yields on changes in term premia led the way to a more substantial influence of foreign factors on domestic bond markets. This pattern was evident for Czechia and Poland, which supports Kolasa and Wesołowski (2020), who demonstrate that in the Polish case, quantitative easing policies in advanced economies generated large effects for domestic sovereign bond markets precisely via term premia. In the case of Hungary, the situation seems to be more complex, as the country differs from Czechia and Poland when it comes to international interdependencies in the long-term interest rates. A plausible explanation of such differences comes from this country's higher idiosyncratic risk. Orłowski and Tsibulina (2014), for example, attribute Hungary's relative financial disintegration with the euro area to its weaker macroeconomic fundamentals. There is, however, an indication of convergence in term premium behaviour between Hungary and two remaining CE economies post-2012.

The role of the neutral rate as a driver of 10-year yields dynamics between the CE economies and Germany proved to be of secondary importance in the period of 2001-2019. It is worth noting that from around 2013, the decline of term premium in Germany coincided with a negative trend in the risk-neutral rate in this economy. On the contrary, in Poland, the neutral rate remained flat, and in Czechia it even increased in the last three years of the analysis. Consequently, as of 2019, the actual 10-year rates in Czechia, Hungary, and

Poland still stood 1.82, 2.27, and 2.25 points higher than German ones, respectively. The differences in risk-neutral rates equalled 1.46, 2.23, and 2.89. If we agree that the yields on German *bunds* are the lower-end reference point of bond yields in the European Union, the way for the 10-year bonds in the CE economies to converge to the core euro area levels must involve a reduction in the expected path of short-term interest rates. In this regard, the monetary response of the CE central banks to the COVID-19-induced crisis may contribute to a decrease in the risk-neutral component of bond yields in these economies, but this warrants further research in the area.

The finding that the behaviour of long-term sovereign rates in the CE countries became more reliant on factors other than the expected path of short term rates has important policy implications. On the one hand, it indicates that between 2009 and 2019 monetary policies in the CE economies were generally tighter and more passive than in the EMU, which accounts for sluggish changes in the neutral rates. On the other hand, it may also imply a limited impact of domestic monetary policy in these economies on the longer end of the yield curve through changes in term premia. What is more, it has been shown that as term premium components decline, they tend to become more sensitive to sudden decompressions, especially when they take unusually low values or fall below zero (Kopp and Williams, 2018). An exogenous negative shock may boost risk premia by altering inflation expectations, an outlook on future growth, or investors' sentiments, and have an eventual impact on monetary and macroeconomic conditions in CE economies.

In a broader sense, our results highlight the importance of international financial factors for the CE economies, and related susceptibility to shocks that originate in larger economies (Rey, 2016). It is not the same as saying that central banks in CE economies are entirely powerless in influencing long-term interest rates. However, in small economies with open bond markets, this impact seems to be limited, especially following the GFC. Even though Czechia, Hungary, and Poland avoided large spikes in 10-year yields during the taper tantrum episode of 2013, the interdependencies we identify tend to increase their exposure to sudden term premium shifts. For example, inflationary pressures or high levels of public debt may force investors to demand higher premium on long-term bonds. Conversely, we do not find much evidence that would lend support to the widely discussed secular stagnation hypothesis (Eggertsson, Mehrotra and Summers, 2016). As far as the 10-year sovereign yields go, the role of neutral rate in spreading a decline in interest rates to CE economies was relatively small, to begin with, and it diminished after the GFC.

7. Conclusions

Long-term interest rates rank among the most important financial and macroeconomic benchmarks, both from a domestic and international point of view. This paper aimed to examine the sovereign 10-year bond yields in three Central European economies, Czechia, Hungary, and Poland, from 2001 to 2019. We employed the ACM term structure to extract time-varying risk-neutral and term premium components of bond yields. We discussed the evolution of these components, along with their relative role in driving the actual interest rates. In the next steps, we studied international comovements of 10-year yields between

the CE economies and Germany. As an extension, we corrected the baseline term structure model for a small sample bias.

In summary, three main points stand out. First, we find that Czechia and Poland, on the one hand, and Hungary, on the other, exhibited different patterns of sovereign long-term interest rates decomposition before 2009. Hungary experienced elevated neutral rates (the expected path of short-term interest rates), while Poland and Czechia relatively high term premia. Post-2009, both risk-neutral and term premium components declined substantially in all three CE economies. In particular, term premia were considerably compressed and approached zero (Hungary) and negative values (Czechia and Poland) around 2013, driving 10-year yields to historically low levels.

Second, we demonstrate that term premia played a larger role in the dynamics of 10-year interest rates in all three CE economies throughout 2001-2019. After 2009, their contribution was even more pronounced. Shifts in term premia explained around 90% of changes in long-term rates in Czechia and Poland, and over 80% in Hungary. At the same time, the role of the risk-neutral rates widely diminished. This phenomenon resembles tendencies observed in the last decade in the major advanced economies.

Third, we show that the 10-year rates in CE economies were higher than in Germany due to relatively larger values of risk-neutral rates, rather than term premium components. Cross-country correlations in 10-year yields were driven mostly by changes in term premia, and Czechia and Poland exhibited stronger ties with each other and with Germany. Hungary's connection to other economies was generally feeble but increased post-2012. Additionally, we demonstrate that the bias-corrected term structure models often produce higher estimates of term premia and lower neutral rates, especially in the second part of the sample.

A major limitation of this study comes from the fact that it relies solely on information embedded in the term structure of benchmark sovereign yields. In subsequent research, the results may be enhanced by using different interest-rate datasets and extending the study to include the COVID-19 period, e.g. by investigating the impact of non-standard monetary policies implemented in CE economies on long-term bond yields (Rebucci, Hartley and Jimenez, 2021). Also, it may be the case that macroeconomic risk factors, which could help in obtaining more precise estimates of risk premia, are "unspanned" by the yield curve (Joslin, Priebsch and Singleton, 2014). Some of those predictors may be directly included in yield curve modelling to expand our understanding of long-term interest rates in CE economies. As such, the results provided in this paper may serve as a useful yardstick for future work in this area.

Acknowledgements

Financial support from a research subsidy granted to Department of Macroeconomics, Cracow University of Economics is acknowledged, no. 7/EEM/2020/POT. The author thanks the Editor and two anonymous reviewers for insightful comments on the paper. The author is also grateful to participants of the 4^{th} Workshop on Political Macroeconomics held in Kraków, September 2019 for helpful discussions.

References

Abbritti, M., Dell'Erba, S., Moreno, A., Sola, S., (2018). Global factors in the term structure of interest rates. *International Journal of Central Banking*, 14(2), pp. 301–339.

Adrian, T., Crump, R. K., Moench, E., (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics*, 110(1), pp. 110–138.

Albagli, E., Ceballos, L., Claro, S., Romero, D., (2019). Channels of US monetary policy spillovers to international bond markets. *Journal of Financial Economics*, 134(2), pp. 447–473.

Bauer, M. D., Hamilton, J. D., (2018). Robust bond risk premia. *Review of Financial Studies*, 31(2), pp. 399–448.

Blanchard, O., (2019). Public debt and low interest rates. *American Economic Review*, 109(4), pp. 1197–1229.

Caballero, R. J., Farhi, E., Gourinchas, P. O., (2017). The safe assets shortage conundrum. *Journal of Economic Perspectives*, 31(3), pp. 29–46.

Cohen, B. H., Hördahl, P., Xia, D., (2018). Term premia: models and some stylised facts. *BIS Quarterly Review*, September, pp. 79–91.

Dec, M., (2021). Parsimonious yield curve modeling in less liquid markets, *GRAPE Working Papers*, 52.

Duffee, G. R., (2002). Term premia and interest rate forecasts in affine models. *Journal of Finance*, 57(1), pp. 405–443.

Dvorák, M., Komárková, Z., Kucera, A., (2019). The Czech Government Yield Curve Decomposition at the Lower Bound. *Czech Journal of Economics and Finance*, 69(1), pp. 2–36.

Eggertsson, G. B., Mehrotra, N. R., Summers, L. H., (2016). Secular Stagnation in the Open Economy. *American Economic Review*, 106(5), pp. 503–507.

Grabowski, W., Stawasz-Grabowska, E., (2020). How have the European central bank's monetary policies been affecting financial markets in CEE-3 countries? *Eurasian Economic Review*, 11, pp. 43–83.

Gürkaynak, R. S., Sack, B., Wright, J. H., (2007). The U.S. Treasury yield curve: 1961 to the present. *Journal of Monetary Economics*, 54(8), pp. 2291–2304.

Gürkaynak, R. S., Wright, J. H., (2012). Macroeconomics and the term structure. *Journal of Economic Literature*, 50(2), pp. 331–367.

Holston, K., Laubach, T., Williams, J. C., (2017). Measuring the natural rate of interest: International trends and determinants. *Journal of International Economics*, 108, pp. S59–S75.

Horváth, D., Kálmán, P., Kocsis, Z., Ligeti, I., (2014). What factors influence the yield curve? *MNB Bulletin*, 9(1), pp. 28–39.

Jabłecki, J., Raczko, A., Wesołowski, G., (2016). Negative bond term premia - a new challenge for Polish conventional monetary policy. *BIS Papers*, 89, pp. 303–315.

Janus, J., (2020). Is ECB Rocking the Boat? Unconventional Monetary Policy in the EMU and Volatility Spillovers to Poland. *Eastern European Economics*, 58(1), pp. 50–67.

Joslin, S., Priebsch, M., Singleton, K. J., (2014). Risk premiums in dynamic term structure models with unspanned macro risks. *Journal of Finance*, 69(3), pp. 1197–1233.

Kilian, L., (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics*, 80(2), pp. 218–230.

Kolasa, M., Wesołowski, G., (2020). International spillovers of quantitative easing. *Journal of International Economics*, 126, pp. 103–330.

Kopp, E., Williams, P., (2018). A Macroeconomic Approach to the Term Premium. *IMF Working Papers*, 18(140).

Kuttner, K. N., (2018). Outside the box: Unconventional monetary policy in the great recession and beyond. *Journal of Economic Perspectives*, 32(4), pp. 121–146.

Litterman, R.B., Scheinkman, J., (1991). Common Factors Affecting Bond Returns. *The Journal of Fixed Income*, 1(1), pp. 54-61.

Mehrotra, A., Moessner, R., Shu, C., (2019). Interest rate spillovers from the United States : expectations, term premia and macro-financial vulnerabilities. *CESifo Working Paper Series*, 814.

Moench, E., (2018). The term structures of global yields. *BIS Papers*, 102.

Obstfeld, M., (2015). Trilemmas and Tradeoffs: Living with Financial Globalization. In *Global Liquidity, Spillovers to Emerging Markets and Policy Responses*. C. Raddatz, D. Saravia and J. Ventura (eds.) Santiago, Chile: Central Bank of Chile, pp. 13–78.

Orłowski, L. T., Tsibulina, A., (2014). Integration of central and eastern European and the euro-area financial markets: Repercussions from the global financial crisis. *Comparative Economic Studies*, 56, pp. 376–395.

Rebonato, R., (2018). *Bond Pricing and Yield Curve Modeling. A Structural Approach*. Cambridge: Cambridge University Press.

Rebucci, A., Hartley, J., Jimenez, D., (2021). An Event Study of COVID-19 Central Bank Quantitative Easing in Advanced and Emerging Economies. *NBER Working Papers*, 2733.

Rey, H., (2016). International channels of transmission of monetary policy and the Mundellian trilemma. *IMF Economic Review*, 64(1), pp. 6–35.

Rudebusch, G. D., Sack, B. P., Swanson, E. T., (2007). Macroeconomic implications of changes in the term premium. *Federal Reserve Bank of St. Louis Review*, 89, pp. 241–270.

Svensson, L. E. O., (1994). Estimating and Interpreting Forward Interest Rates: Sweden 1992-1994. *NBER Working Papers*, 4871.

Turner, P., (2013). Benign neglect of the long-term interest rate. *BIS Working Paper*, 403.

Wesołowski, G., (2018). Do long-term interest rates drive GDP and inflation in small open economies? Evidence from Poland. *Applied Economics*, 50(57), pp. 6174–6192.

Wright, J. H., (2011). Term Premia and Inflation Uncertainty: Empirical Evidence from an International Panel Dataset. *American Economic Review*, 101, pp. 1514–1534.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23 No. 1, pp. 173–184, DOI 10.21307/stattrans-2022-010
Received – 13.10.2020; accepted – 30.09.2021

Variance estimation in stratified adaptive cluster sampling

Uzma Yasmeen[1], Muhammad Noor-ul-Amin[2], Muhammad Hanif[3]

ABSTRACT

In many sampling surveys, the use of auxiliary information at either the design or estimation stage, or at both these stages is usual practice. Auxiliary information is commonly used to obtain improved designs and to achieve a high level of precision in the estimation of population density. Adaptive cluster sampling (ACS) was proposed to observe rare units with the purpose of obtaining highly precise estimations of rare and specially clustered populations in terms of least variances of the estimators. This sampling design proved to be more precise than its more conventional counterparts, including simple random sampling (SRS), stratified sampling, etc. In this paper, a generalised estimator is anticipated for a finite population variance with the use of information of an auxiliary variable under stratified adaptive cluster sampling (SACS). The bias and mean square error expressions of the recommended estimators are derived up to the first degree of approximation. A simulation study showed that the proposed estimators have the least estimated mean square error under the SACS technique in comparison to variance estimators in stratified sampling.

Key words: variance estimator, stratified sampling, stratified adaptive cluster sampling (SACS).

Significance statement

The stratified adaptive cluster sampling technique is an efficient technique used for the population of plants or animals in biological and ecological surveys, which are effective in estimating the population variance when the measurement of the variability of observations is difficult. The main task of this study is to develop a sampling technique and efficient estimators using auxiliary information for estimation of finite population variance. On the basis of a simulation study, the performance of the proposed variance estimators using the stratified adaptive cluster sampling technique is better than the competing variance estimators using stratified random sampling for clustered, hidden and patchy populations.

[1] University of Waterloo, Canada. The University of Lahore, Pakistan. E-mail: uzmayasmeen15@gmail.com.
[2] COMSATS Institute of Information Technology, Lahore. Pakistan. E-mail: nooramin.stats@gmail.com. ORCID: https://orcid.org/0000-0003-2882-221X
[3] National College of Business Administration and Economics, Pakistan.

1. Introduction

In the theory of survey sampling, it is well recognized at the estimation stage known information of auxiliary variables to increase the precision of the estimators of unknown population parameter(s) in different sampling techniques. Several authors have used information of auxiliary variates like population mean, population variance, kurtosis, skewness, etc., to estimate population mean and variance of the study variable. In our daily life, variation is present everywhere. According to the law of nature, no two things or individuals are the same. For instance, a manufacturer wants stable information about the level of variation in consumer's response to his product to be able to know whether to increase or reduce his price, or improve the value of his product. For these reasons, several authors have done important work in this field such as Das and Tripathi [1], Isaki [2], Shabbir and Gupta [3], Tailor et al. [4] and Subramani and Kumarapandiyan [5] and Yasmeen et al. [8].

In ecological and environmental surveys, conventional sampling techniques are not frequently applicable as the consequences obtained from such sampling designs may not be reliable due to lack of information on characteristics of the population. Different types of population are needed in different sampling procedures according to their characteristics. Many populations of plants or animals have different types of cluster tendencies, often the size or the shape of the cluster cannot be recognized before the survey. Applications of sampling methods in real environmental sciences were discussed by Cormack [6]. He mentioned special designs and these designs are needed for environmental sciences. ACS is an efficient method for population which have cluster rare tendencies suggested by Thompson [7]. He suggested an unbiased estimator for the population mean by modifying the Hansen-Hurwitz estimator.

The greatest challenge is the estimation of COVID-19 virus cases in all over the world. Many people do not like to come to laboratories for COVID-19 test or they do not have any symptoms. In such situations for the selection of a sample using simple random sampling or many other techniques for testing to estimate the actual number of COVID-19 cases in the different countries. But precise estimates are still a challenge. Chandra et al. (2021) recommended adaptive cluster sampling technique to provide precise estimates of the number of infected persons. So in real life, the adaptive cluster sampling, stratified adaptive cluster sampling and systematic adaptive cluster sampling technique might be a useful source in finding or tracing the distribution of COVID-19-affected people, for developing the mathematical models for prediction and to get the efficient estimates of the number of COVID-19-affected cases.

In the geographical survey area for the application of the stratified adaptive cluster sampling technique, the survey region can be divided into smaller but same regions

according to a recognized variable like plant species, habitat or soil type. While a large region seems to be homogeneous, the stratified sampling approach based on geographical locations can be used to produce a sample gathered over the whole area. The overall variability in the estimation can be reduced when stratification leads to reduced variability within a stratum with larger variation across strata. Moreover, the most precise estimator is obtained when the units in each stratum are as similar as possible Thompson [12].

By the motivation of many authors, we anticipated variance estimator utilizing midrange, kurtosis, the median, the tri-mean, the coefficient of correlation, the coefficient of variation, the coefficient of skewness and the quartile deviation with the help of single auxiliary variate information under stratified adaptive cluster sampling approach to estimate the finite population variance respectively.

In this paper, variance estimators have been suggested under the stratified adaptive cluster sampling (SACS) design. Section 2 presents the sampling procedure to follow the SACS design. In Section 3, the proposed estimators for the estimation of population variance have been presented. The derivation of the bias and minimum mean square error (MSE) of each estimator are obtained. The simulation study of the proposed estimators and the existing ratio estimator is accomplished in Section 4. Section 5 concludes this paper.

2. Stratified adaptive cluster sampling design for variance estimation

Suppose a finite population Q of size N divided into L non-overlapping strata of size N_h. Let y_{hi} be the value of response varíate (y) and x_{hi} be the value of auxiliary varíate (x), for ith ($i=1,2,\ldots..,N_h$) population unit Q in the hth stratum ($h=1,2,\ldots,L$). Suppose a sample of size n_h from hth stratum is drawn by SRSWOR, where $\left(\sum_{h=1}^{L} n_h = n\right)$. When we ignore the finite population correction factor, the usual variance of $\bar{y}_{st}$ is given by $\text{var}(\bar{y}_{st}) = \sum W_h^2 \frac{S_{y(h)}^2}{n_h} = s_{y(h)}^2$ where $s_{y(h)}^2$ is the variance of population for stratum h. Suppose

$$e_{yh} = \frac{\left(s_{yh}^2 - S_{yh}^2\right)}{S_{yh}^2} \text{ and } e_{xh} = \frac{\left(s_{xh}^2 - S_{xh}^2\right)}{S_{xh}^2} \text{ so that } E\left(e_{yh}^2\right) = \frac{\left(\beta_{2y(h)} - 1\right)}{n_h};$$

$$E\left(e_{xh}^2\right) = \frac{\left(\beta_{2x(h)} - 1\right)}{n_h};$$

$$E\left(e_{yh}e_{xh}\right)=\frac{\left(\vartheta_{22(h)}-1\right)}{n_h};E\left(e_{xh}^2\right)=\frac{\left(\beta_{2x(h)}-1\right)}{n_h};\text{ where }\beta_{2y(h)}=\frac{u_{40(h)}}{u_{20(h)}^2},\beta_{2x(h)}=\frac{u_{04(h)}}{u_{02(h)}^2}$$

are the coefficients of kurtosis of y and x respectively, in the hth stratum and $\varphi_{22(h)}=\frac{u_{22(h)}}{u_{02(h)}^2u_{20(h)}}$, where $u_{ab(h)}=\sum_{i=1}^{N_h}\left(y_{hi}-\bar{y}_h\right)^a\left(x_{hi}-\bar{x}_h\right)^b$.

For the SACS design, let N be denoted as the total number of units in the population. Associated with unit T_{hi}, the ith unit of stratum h is a variable of interest y_{hi}. For any unit T_h of the population, the neighbourhood of unit T_{hi} is defined as a collection of units which includes T_{hi} and with the property that ,if unit $T_{h'i'}$ is in the neighbourhood of unit T_{hi}s, then unit T_{hi} is in the neighbourhood of unit $T_{h'i'}$. A unit T_{hi} is said to fulfil pre-defined condition of interest if the y value associated with that unit is in a specified set C.

The first sample of units is selected from a population using stratified random sampling technique, that is within stratum h, a simple random sample on n_h units designated without replacement, the selection for separate strata being made independently. When a selected unit satisfies the condition, all units in its neighbourhood not already in the sample are added to the sample. Still further units may be added to the sample when any extra-added units satisfies the pre-defined condition, so that the final sample include every unit in the neighbourhood of some sample unit satisfying the condition.

For the unit T_{hi}, the new variate w_{hi} is the total of the y variate of the network to which belongs to $T_{hi.}$ Weighted by the stratum sampling fraction and divided by a weighted sum of the network-stratum intersection.

$w_{hi}=\frac{\frac{n_h}{N_h}\sum_{k=1}^{L}\xi_{khi}}{\sum_{k=1}^{L}\frac{n_k}{N_k}m_{khi}}$, where, suppose, m_{khi} denotes the number of units in the intersection of stratum h with the network which contains unit T_{hi}. ξ_{khi} is the total value of the y-values in the intersection of stratum h with the network that consist of T_{hi} unit and m_{khi} is the number of units in this intersection. The usual variance of $\bar{y}_{wst}$ is given by $\text{var}(\bar{y}_{wst})=\sum W_h^2\frac{s_{wy(h)}^2}{n_h}=s_{y(wst)}^2$ where $s_{wy(h)}^2$ is the population variance for stratum h.

Suppose $e_{wyh}=\frac{\left(s_{wyh}^2-S_{wyh}^2\right)}{S_{wyh}^2}$ and $e_{wxh}=\frac{\left(s_{wxh}^2-S_{wxh}^2\right)}{S_{wxh}^2}$ so, $E\left(e_{wyh}\right)=E\left(e_{wxh}\right)=0;$

$$E\left(e_{wyh}^2\right)=\frac{\left(\beta_{2wy(h)}-1\right)}{n_h};\ E\left(e_{wxh}^2\right)=\frac{\left(\beta_{2wx(h)}-1\right)}{n_h};\ E\left(e_{wyh}e_{wxh}\right)=\frac{\left(\vartheta_{22(wh)}-1\right)}{n_h},$$

$\beta_{2wy(h)}=\frac{u_{40(h)}}{u_{20(h)}^2},\beta_{2wx(h)}=\frac{u_{04(h)}}{u_{02(h)}^2}$ are the coefficients of kurtosis of *y* and *x*, respectively, in the *hth* stratum, and $\varphi_{22w(h)}=\frac{u_{22(h)}}{u_{02(h)}^2u_{20(h)}}$, where

$$u_{abw(h)}=\frac{\sum_{i=1}^{N_h}\left(wy_{hi}-w\bar{y}_h\right)^a\left(wx_{hi}-w\bar{x}_h\right)^b}{N_h}.$$

The variance estimator for stratified sampling given in Shabbir and Gupta [3] is

$$\delta=\sum_{h=1}^{L}\frac{W_h^2}{n_h}\left(s_{y(h)}^2\frac{S_{x(h)}^2}{s_{x(h)}^2}\right).$$

3. Proposed estimator

The following proposed estimator is suggested of finite population variance utilizing the coefficient of skewness, kurtosis, tri-mean, median, quartile deviation under stratified adaptive cluster sampling given by

$$t_G=\sum_{h=1}^{L}\frac{w_h^2}{n_h}\left(s_{wy(h)}^2\frac{\gamma S_{wx(h)}^2+\chi}{\gamma s_{wx(h)}^2+\chi}\right). \tag{3.1}$$

where $\gamma=1,\rho,\beta_1,\beta_2;\chi=0,1,TM,M_d,QD$

By using the notations given in Section 2, the proposed variance estimator (3.1) can be written as

$$t_G=\sum_{h=1}^{L}\frac{w_h^2}{n_h}\left(S_{wy(h)}^2(1+\bar{e}_{wy})\left(1-\frac{\gamma S_{wx(h)}^2e_{wx(h)}}{\gamma s_{wx(h)}^2+\chi}\right)\right). \tag{3.2}$$

After simplifications of this estimator, applying expectations on both sides of (3.3), using the notation given in Section 2, the bias of the recommended variance estimator (3.1) is given by

$$Bias(t_G)=\sum_{h=1}^{L}\frac{w_h^2}{n_h^2}\left(-\omega_iS_{wy(h)}^2\left(Q_{22(h)}-1\right)\right). \tag{3.3}$$

In order to derive mean square error of t_G, again using (3.3), ignoring the higher order power terms, we attain

$$MSE(t_G)=\sum_{h=1}^{L}\frac{w_h^4}{n_h^3}\left(S_{wy(h)}^4\left(B_{2y(h)}-1\right)+\omega_{1(h)}^2\left(B_{2x(h)}-1\right)-2\omega_{1(h)}\left(Q_{22(h)}-1\right)\right). \tag{3.4}$$

$$\omega_{i(h)}=\frac{\gamma S_{wx(h)}^2}{\gamma S_{wx(h)}^2+\chi_{(h)}}, \qquad \text{i=1, 2……..17}$$

where

$$\omega_{1(h)}=1,\ \omega_{2(h)}=\frac{S_{wx(h)}^2}{S_{wx(h)}^2+1},\ \omega_{3(h)}=\frac{S_{wx(h)}^2}{S_{wx(h)}^2+TM_{(h)}},\ \omega_{4(h)}=\frac{S_{wx(h)}^2}{S_{wx(h)}^2+MD_{(h)}},$$

$$\omega_{5(h)}=\frac{S_{wx(h)}^2}{S_{wx(h)}^2+QD_{(h)}},\ \omega_{6(h)}=\frac{\rho_{(h)}S_{wx(h)}^2}{\rho_{(h)}S_{wx(h)}^2+1},\ \omega_{7(h)}=\frac{\rho_{(h)}S_{wx(h)}^2}{\rho_{(h)}S_{wx(h)}^2+TM_{(h)}},$$

$$\omega_{8(h)}=\frac{\rho_{(h)}S_{wx(h)}^2}{\rho_{(h)}S_{wx(h)}^2+MD_{(h)}}\ \omega_{9(h)}=\frac{\rho_{(h)}S_{wx(h)}^2}{\rho_{(h)}S_{wx(h)}^2+QD_{(h)}},\ \omega_{10(h)}=\frac{\beta_{1(h)}S_{wx(h)}^2}{\beta_{1(h)}S_{wx(h)}^2+1},$$

$$\omega_{11(h)}=\frac{\beta_{1(h)}S_{wx(h)}^2}{\beta_{1(h)}S_{wx(h)}^2+TM_{(h)}},\ \omega_{12(h)}=\frac{\beta_{1(h)}S_{wx(h)}^2}{\beta_{1(h)}S_{wx(h)}^2+MD_{(h)}},\ \omega_{13(h)}=\frac{\beta_{1(h)}S_{wx(h)}^2}{\beta_{1(h)}S_{wx(h)}^2+QD_{(h)}},$$

$$\omega_{14(h)}=\frac{\beta_{2(h)}S_{wx(h)}^2}{\beta_{2(h)}S_{wx(h)}^2+1},\ \omega_{15(h)}=\frac{\beta_{2(h)}S_{wx(h)}^2}{\beta_{2(h)}S_{wx(h)}^2+TM_{(h)}},\ \omega_{16(h)}=\frac{\beta_{2(h)}S_{wx(h)}^2}{\beta_{2(h)}S_{wx(h)}^2+MD_{(h)}}$$

$$\omega_{17(h)}=\frac{\beta_{2(h)}S_{wx(h)}^2}{\beta_{2(h)}S_{wx(h)}^2+QD_{(h)}}.$$

We try these estimators but in this article we discussed three estimators which are most efficient among them.

4. Simulation study

The stratified adaptive cluster sampling design is an efficient procedure and flexible technique for capturing more information for hidden and patchy clustered populations. It is a well recognized procedure and suitable a different choice to collect a sample from population. In the section of this study, we consider the performance of SACS design having single auxiliary variable with conventional and non-conventional measures for the variance estimation of population variance. Dryver and Chao [10] and Chao et al. [11] have generated the values for the study variable utilizing the models

(4.1) and (4.2) in such a procedure that there is a strong positive correlation between study variate and simulated variate at unit level.

$$y_{hi} = 4x_{hi} + \varepsilon_{hi} ; \quad \varepsilon_j \sim N(0, x_{hi}) \tag{4.1}$$

$$y_{hi} = 4w_{xhi} + \varepsilon_{xhi} ; \quad \varepsilon_j \sim N(0, w_{xhi}) \tag{4.2}$$

In this system the population is simulated using the model (4.1) has recognized a clearly mentioned sign for strong correlation between variable of interest and auxiliary variable at unit level as well as network level. The model (4.2) has a clear sign for a strong correlation between auxiliary variate and response variable at the network level for simulated population. The population used for simulation is taken from Thompson [8]. The study area is divided into 2 strata and the stratum size is 20*10 = 200, units formed from 20 rows and 10 columns. For each iteration, an initial sample is selected by simple random sampling without replacement in every stratum. A neighbouring unit is distinct as the spatially adjacent units (left, right, top, and bottom) of that unit. 10,000 iterations were performed for every estimator to attain an efficient estimate. The total of *x* values is 223 and the average value is 0.5575. The total of *y* values is 690 and the average value of *y* values is 1.725. The values of y are generated by model 5.1. The condition of interest $c=\{y: y>0\}$ for added units in the sample is used. In each stratum, initial simple random sample sizes were varied n_h = 2, 3, 4, 5, 10, 20, 25, 30, 40, 50, 100 and initial samples n = 4, 6, 8, 10, 20, 40, 50, 60, 80, 100, 200 were used for all strata. The expected final sample of size *n* differs from sample to sample in SACS.

Let $E(v)$ = 7.89, 11.69, 15.41, 19.04, 36.08, 65.59, 78.64, 90.84, 113.31, 133.99, 226.06 denote the final expected sample size in stratified adaptive cluster sampling.

The expected sample size is the expected number of distinct units in the final sample, is the sum of the *N* inclusion probabilities of all the quadrants in the population. It is generally larger than the initial sample size and increases with the increase of the initial sample size.

The comparison of the proposed variance estimator has been made with the variance estimator given in Shabbir and Gupta [3]. The expected final sample size varies from sample to sample in ACS.

The estimated variances of all the suggested estimators considered in this paper have been computed and presented in Table 5.1. and Table 5.2. for population generated by model 4.1 and 4.2 respectively. The results obtained from Table 5.1. and Table 5.2. show that the variances of the proposed estimators are very low as compared to the variance estimator given in Shabbir and Gupta [3] estimator and the results of variances of all the suggested estimators are decreasing on different primary and expected final sample size.

Ten thousand iterations were used to generate samples from population to the achievement of the efficient estimates. In this paper, the variances of the proposed estimators better by increasing the length of the sample according to the results of the simulation study, which are shown in Table 5.1. and Table 5.2. The suggested stratified variance estimators have been developed using known population co-efficient of variation, population co-efficient of correlation, population tri-mean, median, quartile deviation, coefficient of skewness and coefficient of kurtosis with the information of interest variable and auxiliary variable. It is observed that the variances of the developed estimators are smaller than the variance estimator given in Shabbir and Gupta [3] estimator under the stratified adaptive cluster sampling procedure and as Figure 5.1. designates, the developed variance estimators perform better than the variance estimator given in Shabbir and Gupta [3] when the initial sample size is greater than 2 or more.

Hence, the class of stratified variance estimators developed in this paper can be used for enhanced results and preferred for hidden and patchy populations in practical location. Moreover, simulation studies reveal that the variances for the suggested estimators are lower than the variance estimator given Shabbir and Gupta [3] in conditions of population.

5. Concluding discussion

The consequences of the study simulation are given in Table 5.1. with model 5.1 and Table (5.2.) with model 5.2 for three proposed estimators from SACS procedure. The results of the simulation study are obtained from population provides efficient form of variances from SACS technique than the stratified simple random sampling technique for the highly clumped, patchy or hidden population. In this paper, three estimators have been proposed by using a number of conventional and non-conventional measures for interest variable with single auxiliary variable. The variances have been shown in Tables 5.1. and Table 5.2. The developed estimators are found to be more efficient as compared to variance estimator given by Shabbir and Gupta [3] for estimation of finite population variance. The last conclusion obtained from the simulation study confirmed that the proposed estimators are most efficient than the estimator given by Shabbir and Gupta [3] in SACS design on population for different sizes of the sample. The results of variances have shown in Figure 5.1. indicate that the proposed estimators have been minimum variance compare than the existing estimator. The results of Table 5.2. have presented poorer performance of the estimator given in Shabbir and Gupta [3] paper using stratified simple random sampling because of the week correlation between interest variate and auxiliary variate at the network level. According to this study of simulation the 0/0 quantity is assumed to be zero

because Dryver and Chao [2] assumed 0/0 as 0 for the ratio estimator so that the suggested variance estimators using SACS technique returned the values for small sample sizes in population. The amount of estimated variance on population was found to decrease as well as the sample size increases.

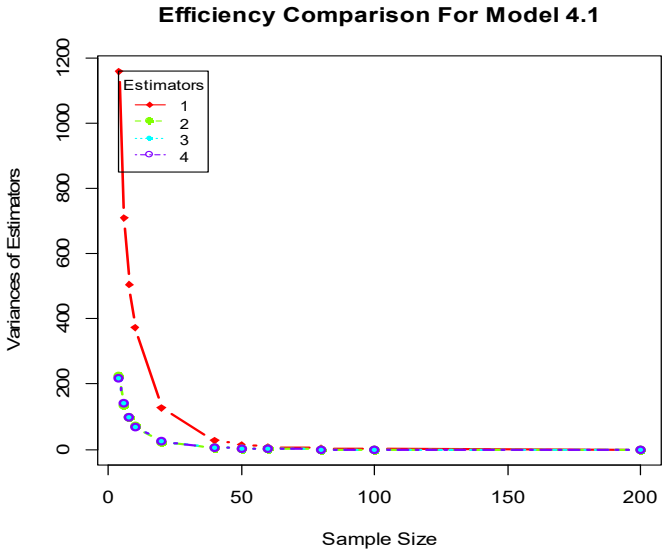
Finally, on the basis of the simulation study, the performance of the proposed variance estimators using SACS technique is better than the competing variance estimator given in Shabbir and Gupta [3] estimator using stratified random sampling for clustered, hidden and patchy populations. The results of simulation study supported that stratified adaptive cluster sampling scheme can be a cost effective and time-saving sampling scheme, and it is suitable for sampling in a spatially accumulated population.

Table 5.1. Variances for different suggested estimators under stratified adaptive cluster design under simulated model 5.1 using population

Sample size s	$E(v)$	$\mathrm{var}(\delta)$	$\mathrm{var}(t_1)$	$\mathrm{var}(t_2)$	$\mathrm{var}(t_3)$
4	7.89	1157.959	226.426	217.049	219.604
6	11.69	708.404	138.573	140.596	141.114
8	15.41	504.876	97.2011	95.733	98.784
10	19.04	373.676	70.459	70.898	69.946
20	36.08	127.508	22.929	23.543	23.844
40	65.59	28.111	5.131	5.159	5.119
50	78.64	14.823	2.681	2.672	2.643
60	90.84	8.303	1.468	1.464569	1.463
80	113.31	2.732	0.468	0.448	0.459
100	133.99	0.946	0.140	0.151	0.146
200	226.06	0.016	0.000	0.000	0.000

Table 5.2. Variances for different suggested estimators under stratified adaptive cluster design under simulated model 5.2 using population

Sample size s	$E(v)$	$\mathrm{var}(\delta)$	$\mathrm{var}(t_1)$	$\mathrm{var}(t_2)$	$\mathrm{var}(t_3)$
4	7.89	13217511	240.646	218.97	216.392
6	11.69	7393951	151.805	138.892	138.683
8	15.41	5064571	102.362	94.193	98.191
10	19.04	3959354	74.795	69.168	70.735
20	36.08	1363876	24.810	23.598	23.677
40	65.59	432186.7	5.202	5.145	5.131
50	78.64	277516.3	2.706	2.686	2.694
60	90.84	161725.8	1.463	1.465	1.454
80	113.31	65089.37	0.462	0.466	0.452
100	133.99	29852.82	0.143	0.142	0.144
200	226.06	194.552	0.000	0.000	0.000



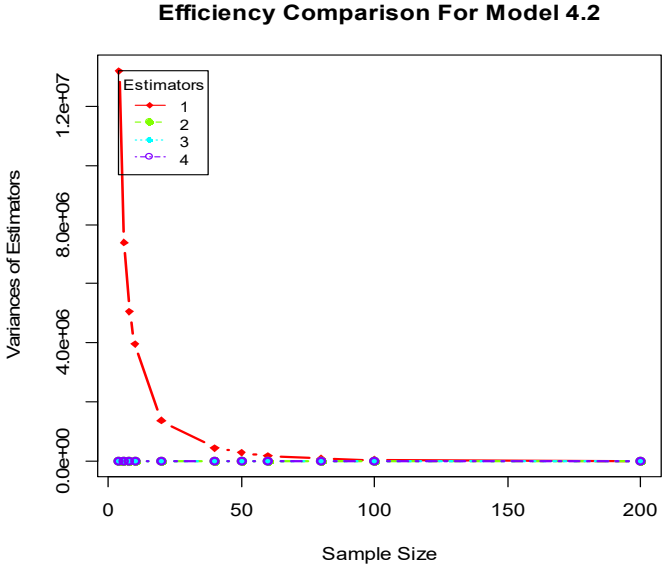


Figure 5.1. Variances of developed estimators under SACS design to the variance estimator given in Shabbir and Gupta [10] under stratified sampling with comparable sample sizes under models.

REFRENCES

Boomer, K., Werner, C., Brantley, S., (2000). CO2 emissions related to the Yellowstone volcanic System, J. Geophys. Res., 105(B5), pp. 10817-10830.

Chandra, G., Tiwari, N., & Nautiyal, R., (2021). Adaptive cluster sampling-based design for estimating COVID-19 cases with random samples, Current Science (00113891), 120(7).

Chao, C. T., Dryver, A. L., Chiang, T. C., (2011). Leveraging the Rao-Blackwell Theorem to improve ratio estimators in adaptive cluster sampling, Environmental and Ecological Statistics 18, pp. 543-568.

Cormack, R. M, (1988). Statistical challenges in the environmental sciences, A Personal view, Journal of the Royal Statistical Society: Series A, 151, pp. 201-210.

Das, A. K. , Tripathi, T. P., (1978). Use of auxiliary information in estimating the finite population variance, Sankhya, 40, pp. 139-148.

Dryver, A. L.,Chao, C. T., (2007). Ratio estimators in adaptive cluster sampling, Environmetrics, 18, pp. 607–620.

Isaki, C. T., (1983). Variance estimation using auxiliary information, Journal of the American Statistical Association, 78, pp. 117-123.

Shabbir, J., Gupta, S., (2010). Some estimators of finite population variance of stratified sample mean, Communications in Statistics–Theory and Methods, 16, pp. 3001-3008.

Subramani, J, Kumarapandiyan, G., (2013). Estimation of variance using known coefficient of variation and median of an auxiliary variable, Journal of Modern Applied Statistical Methods, 12, pp. 58-64.

Tailor, R., Tailor, R., Parmar, R., Kumar, M., (2012). Dual to ratio cum product estimator using known parameters of auxiliary variables, Journal of Reliability and Statistical Studies, 5, pp. 65-71.

Thompson, S. K., (1990). Adaptive Cluster Sampling, Journal of the American Statistical Association, 85, pp. 1050–1059.

Thompson S., (1991b). Stratified adaptive cluster sampling, Biometrika, 78(2), pp. 389-397.

Thompson, S. K., (2002). Sampling, New York: Wiley.

Yasmeen, U., Noor ul Amin, M., Hanif M., (2018). Exponential Estimators of Finite PopulationVariance Using Transformed Auxiliary Variables, Proceedings of the National Academy of Sciences, India Section A: Physical Sciences.

APPENDIX

Table C.1. x-population

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	14	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	36	34	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	3	63	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	16	3	0	0	0	2	12	12	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	2	57	65	17	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	5	14	5	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

STATISTICS IN TRANSITION new series, March 2022
Vol. 23 No. 1, pp. 185–200, DOI 10.21307/stattrans-2022-011
Received – 13.06.2020; accepted – 26.03.2021

Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart

Akeem Ajibola Adepoju[1], **Sauta S. Abdulkadir**[2], **Danjuma Jibasen**[3], **Haruna Chiroma**[4]

ABSTRACT

Some industrial data often come with uncertainty, which in some cases depends on the decision of those responsible for taking the measurement in the production process. While the fuzzy approach helps to tackle the ambiguity that arises in the measurement, an interval type-2 fuzzy set deals with such uncertainty better due to its flexibility over the control limits of its control chart. This paper aims to develop an Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart (IT2FEWMA) under the fuzzy type-2 condition. This development will facilitate monitoring small and moderate shifts in the production process in conditions of uncertainty.

Key words: Exponentially weighted moving average control chart, Fuzzy control chart, Fuzzy sets, Interval Type-2 fuzzy sets, Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart, Statistical process control.

1. Introduction

The control chart scheme is made up of variable and attribute control charts, the former of which constitutes observations with continuous random variable while the later constitutes observations of a discrete random variable. The exponentially weighted moving average (EWMA) control chart was introduced by Roberts, (1959). Many studies have then been conducted to extend this methodology. The EWMA chart is known for its sensitivity on a small shift in the process mean. In classical statistical process control chart, data are defined in crisp value. However, gage,

[1] Department of Statistics, Kano University of Science and Technology, Wudil. Kano State. Nigeria. E-mail: akeebola@gmail.com. ORCID: https://orcid.org/0000-0003-1376-7369.

[2] Department of Statistics and Operation Research, Modibb Adama University. Adamawa State, Nigeria. E-mail: ssabdulkadir@mautech.edu.ng. ORCID: https://orcid.org/0000-0002-9104-8207.

[3] Department of Statistics and Operation Research, Modibb Adama University. Adamawa State, Nigeria. E-mail: djibasan2001@gmail.com. ORCID: https://orcid.org/0000-0003-3190-235X.

[4] Ins Department of Computer and Engineering, University of Hafr Al Batin, Saudi Arabia. E-mail: freedonchi@yahoo.com. ORCID: https://orcid.org/0000-0003-3446-4316.

environment conditions, operator's discretion, which are the measurement determinants, can collect the measurement with "vagueness" or "uncertain". These uncertainties about the measurements lead to difficult challenges while attempting to obtain a crisp result from the process. Human discretion is one of the significant factors used in defining the quality characteristics. However, data from such discretion are bound with uncertainty and direct application of the classical control charts on such data may require further information and transformation of the classical control chart limits and the data points. In such a situation, evaluation of fuzzy data is best achieved using fuzzy control charts as a tool. The fuzzy set theory was introduced by Zadeh (1965). Observations with uncertainty are handled mathematically with the fuzzy logic. A number of publications with many applications in Statistical Process Control (SPC) integrated with the fuzzy system has gained more efforts from many authors. The amalgamation of different SPC tools with the type-1 fuzzy theory, intuitionistic fuzzy theory, hesitant fuzzy theory, neurosophic fuzzy theory, type-2 fuzzy theory, and recently, the interval type-2 fuzzy theories have come into existence in recent publications. However, no publication exists in Interval Type-2 Exponentially Weighted Moving Average (IT2FEWMA) control chart. This paper is designed in a fuzzy environment for three-dimensional data, that is, the upper membership value, lower membership value and their respective representative value. The principal contribution of this research is to develop the theoretical foundation of the Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Charts (IT2FEWMA) and their application.

2. Literature review

In a real life situation, vagueness and uncertainty occur as a result of human error usually due to judgmental decision based on qualitative measures, such as the weather is hot or too hot or cold is based on qualitative measurement which might not be presented with exact value. In most cases, one tends to ask what degree of the hotness or coldness of the weather is. Zadeh (1965) proposed the conceptual idea of the fuzzy set theory. He proposed the type-1 fuzzy sets, with a degree of membership called crisp membership value, whose values are over the range 0 to 1. He also proposed the type-2 fuzzy sets, which is an extension of the type-1 fuzzy sets. This type-2 fuzzy set is three dimensional, that is, it comprises two membership functions, thus, the upper membership function and the lower membership function and the representative values. The conception of the idea of the fuzzy control chart and application was firstly documented by Raz and Wang (1990), and Wang and Raz (1990). A considerable number of authors made various contributions to the extensional development in this area of research, which include but are not limited to Kanagawa

et al. (1993), El-shal and Moris (2000), Rowlands and Wang (2000), Gulbay *et al.* (2004), Karnik and Mendel (2001), Mendel and John (2002), Cheng (2005), Gulbay and Kahraman (2006a), Mendel *et al.* (2006), Erginel (2008), Senturk and Erginel (2009), Senturk (2010), Senturk *et al.* (2010), Kaya and Kahraman (2011), Erginel *et al.* (2011), Senturk *et al.* (2014), Erginel (2014), Poongodi and Muthulakshmi (2015), Cervantes and Castillo (2015), Wang and Hyrniewicz (2015), Edmundas *et al.* (2015), Castillo *et al.* (2016), Chen and Huang (2016), Castillo *et al.* (2016), Hou *et al.* (2016), Kaya *et al.* (2017), Senturk and Antucheviciene (2017), Erginel *et al.* (2018), Ontiveros-Robles *et al.* (2018), Adepoju (2018), Ercan and Anagun (2018), Adepoju *et al.* (2019a), Adepoju *et al.* (2019c), and Adepoju *et al.* (2019b).

3. Methodology

3.1. Type-2 fuzzy sets and Interval type-2 fuzzy sets

Definition 1. A type-2 fuzzy set (T2 FS) denoted by $\tilde{\tilde{A}}$ in a universe of discourse is characterized by a type-2 membership function given as $\mu_{\tilde{\tilde{A}}}(x,u)$, where $x \in X$ and $u \in J_x \subseteq [0,1]$. Mathematically, this can be expressed as

$$\tilde{\tilde{A}} = \left\{ \left((x,u), \mu_{\tilde{\tilde{A}}}(x,u) \right) \middle| \forall x \in X, \forall u \in J_x [0,1], 0 \le \mu_{\tilde{\tilde{A}}}(x,u) \le 1 \right\}$$

where J_x denotes an interval $[0,1]$. This type-2 fuzzy set can be expressed as

$$\tilde{\tilde{A}} = \int_{x \in X} \int_{u \in J_x} \mu_{\tilde{\tilde{A}}}(x,u) / (x,u),$$

where $\int\int$ denotes union over all admissible x and u as given by Mendel *et al.* (2006) as well as Kahraman (2014).

3.2. Interval type-2 fuzzy sets

An interval type-2 fuzzy sets (IT2 FS) also known as closed interval type-2 fuzzy set (CIT2 FS) can be defined as a special case of type-2 fuzzy set $\tilde{\tilde{A}}$ represented by the type-2 membership function $\mu \approx (x,u)$. If all $\mu \approx (x,u) = 1$. It follows that the interval type-2 fuzzy set is expressed as

$$\tilde{\tilde{A}} = \int_{x \in X} \int_{u \in J_x} 1 / (x,u)$$

where $J_x \subseteq [0,1]$, Ghorabaee *et al.* (2016),

4. The proposed Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart.

4.1. Exponentially Weighted Moving Average Control Chart (EWMA)

Exponentially Weighted Moving Average Control Chart (EWMA) was introduced by Roberts (1959). It is a better option to Shewhart control chart based on its sensitivity to a small shift in the process mean. It uses both current and historical data to monitor any slight shift in the mean of the process and its statistic is expressed as

$$z_i = \lambda \bar{X}_i + (1-\lambda) z_{i-1}$$

where z_i and $\bar{X}_i$ denote the ith exponentially weighted moving average and ith sample average respectively, and $i = 1, 2, 3, ..., k$, λ is the smoothing constant and it is given as $0 < \lambda < 1$. The starting value being the first sample mean is the process target such that $z_0 = \bar{\bar{X}}$, where $\bar{\bar{X}}$ is the grand mean.

The classical EWMA control chart limits are established as follows.

$$\begin{aligned} UCL &= \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]} \\ CL &= \mu_0 \\ LCL &= \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}\left[1-(1-\lambda)^{2i}\right]} \end{aligned} \tag{1}$$

where UCL, CL and LCL are upper control limit, centre line and lower control limit respectively, $\boldsymbol{L}$ is the width of the control limits, λ is the smoothing constant and σ is the standard deviation.

If σ^2/n is the variance of $\bar{X}_i$ independent random variables drawn from a population with known standard deviation σ, with sample size n, then
for a small sample number *i*

$$\sigma_{zi}^2 = \sigma^2/n\left(\frac{\lambda}{2-\lambda}\right)\left[1-(1-\lambda)^{2i}\right],$$

for a moderate and large sample number *i*

$$\sigma_{zi}^2 = \sigma^2/n\left(\frac{\lambda}{2-\lambda}\right).$$

For a moderate and large sample size n, the statistic below can be used to obtaine the estimates of the upper control limit, center line and the lower control limit of the classical EWMA control chart.

$$UCL_{ewma} = \bar{\bar{\bar{X}}} + 3\sigma \Big/ \sqrt{n} \sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$
$$CL_{ewma} = \bar{\bar{\bar{X}}}$$
$$LCL_{ewma} = \bar{\bar{\bar{X}}} - 3\sigma \Big/ \sqrt{n} \sqrt{\left(\frac{\lambda}{2-\lambda}\right)} \tag{2}$$

4.2. Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart (IT2FEWMA)

For a large sample number when the σ is known, the IT2FEWMA control limits of the control chart is obtained by fuzzification and expressed as

$$UCL_{IT2FEWMA} = \left(\bar{X}_a^U, \bar{X}_b^U, \bar{X}_c^U, \bar{X}_d^U\right) + \frac{3}{\sqrt{n}}\left(\sigma_a^U, \sigma_b^U, \sigma_c^U, \sigma_d^U\right)\sqrt{\left(\frac{\lambda}{2-\lambda}\right)},$$
$$\left(\bar{X}_a^L, \bar{X}_b^L, \bar{X}_c^L, \bar{X}_d^L\right) + \frac{3}{\sqrt{n}}\left(\sigma_a^L, \sigma_b^L, \sigma_c^L, \sigma_d^L\right)\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$
$$= \left[\begin{array}{l} \bar{X}_a^U + \frac{3\sigma_a^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_b^U + \frac{3\sigma_b^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_c^U + \frac{3\sigma_c^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \\ \bar{X}_d^U + \frac{3\sigma_d^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}; \min\left(H_1\left(A_i^U\right), H_2\left(A_i^U\right)\right), \\ \bar{X}_a^L + \frac{3\sigma_a^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_b^L + \frac{3\sigma_b^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_c^L + \frac{3\sigma_c^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \\ \bar{X}_d^L + \frac{3\sigma_d^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}; \min\left(H_1\left(A_i^L\right), H_2\left(A_i^L\right)\right) \end{array}\right]$$

$$CL_{IT2FEWMA} = \left(\begin{array}{l} \left(\bar{X}_a^U, \bar{X}_b^U, \bar{X}_c^U, \bar{X}_d^U\right); \min\left(H_1\left(A_i^U\right), H_2\left(A_i^U\right)\right), \\ \left(\bar{X}_a^L, \bar{X}_b^L, \bar{X}_c^L, \bar{X}_d^L\right); \min\left(H_1\left(A_i^L\right), H_2\left(A_i^L\right)\right) \end{array}\right)$$

(3)

$$UCL_{IT2FEWMA} = \left(\bar{X}_a^U, \bar{X}_b^U, \bar{X}_c^U, \bar{X}_d^U\right) - \frac{3}{\sqrt{n}}\left(\sigma_a^U, \sigma_b^U, \sigma_c^U, \sigma_d^U\right)\sqrt{\left(\frac{\lambda}{2-\lambda}\right)},$$

$$\left(\bar{X}_a^L, \bar{X}_b^L, \bar{X}_c^L, \bar{X}_d^L\right) - \frac{3}{\sqrt{n}}\left(\sigma_a^L, \sigma_b^L, \sigma_c^L, \sigma_d^L\right)\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}$$

$$= \left[\begin{array}{l} \bar{X}_a^U - \frac{3\sigma_a^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_b^U - \frac{3\sigma_b^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_c^U - \frac{3\sigma_c^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \\ \bar{X}_d^U - \frac{3\sigma_d^U}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}; \min\left(H_1\left(A_i^U\right), H_2\left(A_i^U\right)\right), \\ \bar{X}_a^L - \frac{3\sigma_a^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_b^L - \frac{3\sigma_b^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \bar{X}_c^L - \frac{3\sigma_c^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}, \\ \bar{X}_d^L - \frac{3\sigma_d^L}{\sqrt{n}}\sqrt{\left(\frac{\lambda}{2-\lambda}\right)}; \min\left(H_1\left(A_i^L\right), H_2\left(A_i^L\right)\right) \end{array}\right]$$

4.3. Defuzzication of Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart (IT2FEWMA)

$$DIT2_{trap(i)}^U = \frac{\left(a_{i4}^U - a_{i1}^U\right) + \left(H_2\left(\tilde{A}_1^U\right)a_{i2}^U - a_{i1}^U\right) + \left(H_1\left(\tilde{A}_1^U\right)a_{i3}^U - a_{i1}^U\right)}{4} + a_{i1}^U$$

$$DIT2_{trap(i)}^L = \frac{\left(a_{i4}^L - a_{i1}^L\right) + \left(H_2\left(\tilde{A}_1^L\right)a_{i2}^L - a_{i1}^L\right) + \left(H_1\left(\tilde{A}_1^L\right)a_{i3}^L - a_{i1}^L\right)}{4} + a_{i1}^L \quad (4)$$

$$DIT2_{trap(i)} = \frac{DIT2_{trap(i)}^U + DIT2_{trap(i)}^L}{2}$$

$i = 1, 2, 3, ..., n$.

Similarly, from the equation (4) the modified BNP technique is being transformed to the interval type-2 fuzzy exponentially moving average (IT2FEWMA) control chart limits as expressed below.

$$DIT2FEWMA_{trap(i)}^U = \frac{\left(\bar{X}a_{.4}^U - \bar{X}a_{.1}^U\right) + \left(H_2\left(\tilde{A}_1^U\right)\bar{X}a_{.2}^U - \bar{X}a_{.1}^U\right) + \left(H_1\left(\tilde{A}_1^U\right)\bar{X}a_{.3}^U - \bar{X}\,a_{.1}^U\right)}{4} + \bar{X}a_{.1}^U$$

$$DIT2FEWMA^{L}_{trap(i)} = \frac{\left(\bar{X}a^{L}_{.4} - \bar{X}a^{L}_{.1}\right) + \left(H_2\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.2} - \bar{X}a^{L}_{.1}\right) + \left(H_1\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.3} - \bar{X}a^{L}_{.1}\right)}{4} + \bar{X}a^{L}_{.1} \tag{5}$$

$$DIT2FEWMA_{trap(i)} = \frac{DIT2^{U}_{trap(i)} + DIT2^{L}_{trap(i)}}{2}. \tag{6}$$

$i = 1, 2, 3, ..., n$

For the upper control limit:

$$UCL(DIT2FEWMA^{U}_{trap(i)}) = \frac{\left(\bar{X}a^{U}_{.4} - \bar{X}a^{U}_{.1}\right) + \left(H_2\left(\tilde{A}^{U}_{1}\right)\bar{X}a^{U}_{.2} - \bar{X}a^{U}_{.1}\right) + \left(H_1\left(\tilde{A}^{U}_{1}\right)\bar{X}a^{U}_{.3} - \bar{X}\,a^{U}_{.1}\right)}{4} + \bar{X}a^{U}_{.1}$$

$$UCL(DIT2FEWMA^{L}_{trap(i)}) = \frac{\left(\bar{X}a^{L}_{.4} - \bar{X}a^{L}_{.1}\right) + \left(H_2\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.2} - \bar{X}a^{L}_{.1}\right) + \left(H_1\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.3} - \bar{X}a^{L}_{.1}\right)}{4} + \bar{X}a^{L}_{.1}$$

$$UCL(DIT2FEWMA_{trap(i)}) = \frac{UCL(DIT2FEWMA^{U}_{trap(i)}) + UCL(uDIT2^{L}_{trap(i)})}{2}.$$

(7)

For the lower control limit:

$$LCL(DIT2FEWMA^{U}_{trap(i)}) = \frac{\left(\bar{X}a^{U}_{.4} - \bar{X}a^{U}_{.1}\right) + \left(H_2\left(\tilde{A}^{U}_{1}\right)\bar{X}a^{U}_{.2} - \bar{X}a^{U}_{.1}\right) + \left(H_1\left(\tilde{A}^{U}_{1}\right)\bar{X}a^{U}_{.3} - \bar{X}\,a^{U}_{.1}\right)}{4} + \bar{X}a^{U}_{.1}$$

$$LCL(DIT2FEWMA^{L}_{trap(i)}) = \frac{\left(\bar{X}a^{L}_{.4} - \bar{X}a^{L}_{.1}\right) + \left(H_2\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.2} - \bar{X}a^{L}_{.1}\right) + \left(H_1\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.3} - \bar{X}a^{L}_{.1}\right)}{4} + \bar{X}a^{L}_{.1}$$

$$LCL(DIT2FEWMA_{trap(i)}) = \frac{LCL(DIT2FEWMA^{U}_{trap(i)}) + LCL(DIT2FEWMA^{L}_{trap(i)})}{2}$$

(8)

For the centre line:

$$CL(DIT2FEWMA^{U}_{trap(i)}) = \frac{\left(\bar{X}a^{U}_{.4} - \bar{X}a^{U}_{.1}\right) + \left(H_2\left(\tilde{A}^{U}_{1}\right)\bar{X}a^{U}_{.2} - \bar{X}a^{U}_{.1}\right) + \left(H_1\left(\tilde{A}^{U}_{1}\right)\bar{X}a^{U}_{.3} - \bar{X}\,a^{U}_{.1}\right)}{4} + \bar{X}a^{U}_{.1}$$

$$CL(DIT2FEWMA^{L}_{trap(i)}) = \frac{\left(\bar{X}a^{L}_{.4} - \bar{X}a^{L}_{.1}\right) + \left(H_2\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.2} - \bar{X}a^{L}_{.1}\right) + \left(H_1\left(\tilde{A}^{L}_{1}\right)\bar{X}a^{L}_{.3} - \bar{X}a^{L}_{.1}\right)}{4} + \bar{X}a^{L}_{.1}$$

$$CL(DIT2FEWMA_{trap(i)}) = \frac{CL(DIT2FEWMA^{U}_{trap(i)}) + CL(DIT2FEWMA^{L}_{trap(i)})}{2}$$

(9)

This transformation technique above can be used to defuzzify each sample, such that every defuzzified sample point is monitored with the defuzzified control limits. And the criterion for the in-control criteria expressed as

$$LCL(DIT2FEWMA) < DIT2FEWMA_{(i)} < UCL(DIT2FEWMA)\,. \tag{10}$$

The equation (10) above indicates the in-control process, otherwise the process is out of control.

5. Application of the Interval Type-2 fuzzy Exponentially Weighted Moving Average Control Chart

The application of the IT2FEWMA was conducted by simulating the data. Below is the data with 60 sample points with trapezium fuzzy number of measurement and the upper and lower membership values.

Table 1. Interval type-2 trapezoidal fuzzy number of measurement simulated from the parameters of the existing process data

Sn	X_a^u	X_b^u	X_c^u	X_d^u	$H_1(A_i^u)$	$H_2(A_i^u)$	X_a^l	X_b^l	X_c^l	X_d^l	$H_1(A_i^l)$	$H_2(A_i^l)$
1	3.62	3.81	3.99	4.21	1	1	3.50	3.69	3.90	4.12	0.7	0.5
2	3.61	3.77	3.99	4.18	1	1	3.51	3.69	3.87	4.11	0.8	0.6
3	3.62	3.81	4.00	4.20	1	1	3.50	3.69	3.88	4.11	0.7	0.6
4	3.59	3.81	3.98	4.20	1	1	3.52	3.71	3.89	4.08	0.8	0.6
5	3.60	3.81	4.00	4.20	1	1	3.52	3.70	3.91	4.09	0.6	0.5
6	3.62	3.80	3.99	4.18	1	1	3.49	3.69	3.91	4.08	0.8	0.6
7	3.60	3.78	4.00	4.21	1	1	3.51	3.72	3.90	4.11	0.9	0.8
8	3.59	3.80	3.98	4.19	1	1	3.50	3.71	3.90	4.11	0.7	0.5
9	3.59	3.80	4.00	4.21	1	1	3.49	3.70	3.90	4.09	0.8	0.6
10	3.60	3.80	4.00	4.20	1	1	3.49	3.73	3.92	4.10	0.8	0.7
11	3.61	3.81	3.97	4.21	1	1	3.52	3.69	3.91	4.10	0.6	0.5
12	3.59	3.80	4.01	4.19	1	1	3.49	3.70	3.90	4.09	0.7	0.6
13	3.61	3.79	4.00	4.18	1	1	3.50	3.70	3.91	4.10	0.7	0.5
14	3.60	3.80	3.99	4.22	1	1	3.48	3.69	3.90	4.10	0.8	0.7
15	3.58	3.81	3.99	4.20	1	1	3.50	3.71	3.92	4.11	0.9	0.8
16	3.61	3.80	4.01	4.21	1	1	3.50	3.70	3.89	4.09	0.8	0.7
17	3.62	3.80	4.01	4.20	1	1	3.51	3.72	3.91	4.10	0.9	0.8
18	3.59	3.82	3.99	4.19	1	1	3.50	3.72	3.91	4.10	0.7	0.5
19	3.60	3.80	4.00	4.21	1	1	3.50	3.70	3.91	4.08	0.8	0.6
20	3.60	3.79	4.00	4.21	1	1	3.51	3.70	3.90	4.11	0.7	0.6
21	3.61	3.80	3.98	4.21	1	1	3.50	3.71	3.91	4.12	0.8	0.6
22	3.60	3.80	3.99	4.19	1	1	3.50	3.71	3.90	4.09	0.6	0.5
23	3.60	3.80	4.00	4.21	1	1	3.51	3.71	3.91	4.09	0.8	0.6
24	3.60	3.80	4.00	4.21	1	1	3.51	3.71	3.90	4.11	0.9	0.8
25	3.61	3.79	4.00	4.19	1	1	3.49	3.71	3.93	4.11	0.7	0.5

Table 1. Interval type-2 trapezoidal fuzzy number of measurement simulated from the parameters of the existing process data (cont.)

Sn	X_a^u	X_b^u	X_c^u	X_d^u	$H_1(A_i^u)$	$H_2(A_i^u)$	X_a^l	X_b^l	X_c^l	X_d^l	$H_1(A_i^l)$	$H_2(A_i^l)$
26	3.61	3.81	4.00	4.22	1	1	3.51	3.69	3.90	4.10	0.8	0.6
27	3.59	3.81	4.01	4.19	1	1	3.52	3.70	3.90	4.10	0.8	0.7
28	3.60	3.80	4.01	4.21	1	1	3.51	3.71	3.89	4.09	0.6	0.5
29	3.61	3.81	3.98	4.19	1	1	3.49	3.71	3.90	4.09	0.7	0.6
30	3.61	3.81	4.00	4.20	1	1	3.50	3.69	3.91	4.08	0.7	0.5
31	3.59	3.79	4.00	4.20	1	1	3.50	3.69	3.91	4.11	0.8	0.7
32	3.59	3.80	3.97	4.20	1	1	3.50	3.70	3.91	4.10	0.9	0.8
33	3.58	3.79	4.00	4.20	1	1	3.51	3.70	3.88	4.09	0.7	0.6
34	3.61	3.81	4.00	4.19	1	1	3.50	3.71	3.92	4.09	0.7	0.5
35	3.60	3.81	3.99	4.21	1	1	3.50	3.72	3.90	4.09	0.7	0.5
36	3.61	3.79	4.01	4.20	1	1	3.49	3.70	3.91	4.09	0.8	0.6
37	3.61	3.79	4.01	4.21	1	1	3.49	3.69	3.89	4.09	0.7	0.6
38	3.61	3.80	4.01	4.17	1	1	3.51	3.72	3.90	4.10	0.8	0.6
39	3.60	3.81	3.98	4.21	1	1	3.51	3.70	3.89	4.12	0.6	0.5
40	3.61	3.81	4.00	4.20	1	1	3.50	3.70	3.89	4.10	0.8	0.6
41	3.61	3.81	4.00	4.19	1	1	3.51	3.70	3.91	4.11	0.9	0.8
42	3.60	3.79	4.00	4.20	1	1	3.48	3.70	3.90	4.10	0.7	0.5
43	3.62	3.80	4.00	4.19	1	1	3.51	3.69	3.89	4.10	0.8	0.6
44	3.60	3.80	4.01	4.19	1	1	3.50	3.71	3.91	4.09	0.8	0.7
45	3.61	3.80	3.99	4.20	1	1	3.50	3.71	3.91	4.09	0.6	0.5
46	3.58	3.79	4.00	4.20	1	1	3.49	3.71	3.89	4.08	0.7	0.6
47	3.59	3.78	4.01	4.19	1	1	3.51	3.70	3.88	4.11	0.7	0.5
48	3.59	3.81	3.99	4.19	1	1	3.49	3.69	3.88	4.10	0.8	0.7
49	3.59	3.78	4.00	4.21	1	1	3.51	3.71	3.89	4.11	0.9	0.8
50	3.59	3.79	4.01	4.19	1	1	3.50	3.70	3.90	4.11	0.7	0.5
51	3.60	3.80	4.00	4.20	1	1	3.48	3.70	3.90	4.09	0.8	0.6
52	3.61	3.80	4.00	4.18	1	1	3.50	3.71	3.89	4.10	0.8	0.7
53	3.61	3.81	3.98	4.21	1	1	3.51	3.72	3.91	4.08	0.6	0.5
54	3.61	3.80	4.00	4.20	1	1	3.49	3.68	3.90	4.11	0.7	0.5
55	3.60	3.81	4.00	4.20	1	1	3.49	3.72	3.87	4.11	0.8	0.6
56	3.61	3.81	4.00	4.20	1	1	3.50	3.72	3.89	4.10	0.7	0.6
57	3.60	3.80	3.99	4.20	1	1	3.50	3.72	3.90	4.11	0.8	0.6
58	3.59	3.80	4.00	4.20	1	1	3.50	3.70	3.89	4.08	0.6	0.5
59	3.61	3.79	4.00	4.20	1	1	3.50	3.68	3.89	4.11	0.8	0.6
60	3.59	3.79	3.99	4.20	1	1	3.51	3.71	3.89	4.09	0.9	0.8

5.1. Result and discussion

The fuzzy numbers are modelled with the interval type-2 FEWMA control chart using trapezoidal membership functions.

Table 2. The representation of the result of the interval type-2 FEWMA control chart

UCL_{FEWMA}	3.606533	3.806300	4.006047	4.206225	1	1	3.506508	3.706719	3.906252	4.106251	0.6	0.5
CL_{FEWMA}	3.600000	3.800000	4.000000	4.200000	1	1	3.500161	3.700372	3.899905	4.099904	0.6	0.5
LCL_{FEWMA}	3.593838	3.793660	3.993353	4.193530	1	1	3.493813	3.694024	3.893558	4.093557	0.6	0.5

Table 2 above shows the limits of the FEWMA Control Chart and this include the upper control limit, centre line and lower control limit. This is obtained from equation 3.

Table 3. The representation of the defuzzification values of the IT2FEWMA control chart

$DIT2FEWMA^{U}_{(i)}$	3.906291	$UCL - DIT2FEWMA_{(i)}$	3.823862	$DIT2FEWMA^{L}_{(i)}$	3.741432
$DIT2FEWMA^{U}_{(i)}$	3.899944	$CL - DIT2FEWMA_{(i)}$	3.817514	$DIT2FEWMA^{L}_{(i)}$	3.735085
$DIT2FEWMA^{U}_{(i)}$	3.893597	$LCL - DIT2FEWMA_{(i)}$	3.811167	$DIT2FEWMA^{L}_{(i)}$	3.728737

Table 3 above shows the limits of the diffuzified IT2FEWMA Control Chart and this include the upper control limit, centre line and the lower control limit. It is obtained from equation 7, 8 and 9.

Table 4. The values of the defuzzification of each and every sample and its corresponding decision criteria

$UCL(DIT2FEWMA)$	$DIT2FEWMA_{(i)}$	$LCL(DIT2FEWMA)$	$3.811167 < DIT2FEWMA_{(i)} < 3.823862$
3.900753	3.818206	3.735658	In-control
3.898874	3.816990	3.735105	In-control
3.900749	3.817966	3.735183	In-control
3.899352	3.817340	3.735328	In-control
3.900304	3.818219	3.736135	In-control
3.899481	3.816832	3.734183	In-control
3.900007	3.817989	3.735971	In-control
3.899131	3.817155	3.735178	In-control
3.900071	3.817340	3.734610	In-control
3.899979	3.817546	3.735113	In-control
3.899754	3.817950	3.736145	In-control
3.899773	3.817041	3.734309	In-control
3.899344	3.817188	3.735031	In-control
3.900413	3.817038	3.733664	In-control
3.899769	3.817680	3.735590	In-control
3.900462	3.817564	3.734666	In-control

Table 4. The values of the defuzzification of each and every sample and its corresponding decision criteria (cont.)

$UCL(DIT2FEWMA)$	$DIT2FEWMA_{(i)}$	$LCL(DIT2FEWMA)$	$3.811167 < DIT2FEWMA_{(i)} < 3.823862$
3.900647	3.818358	3.736069	In-control
3.900034	3.817611	3.735188	In-control
3.900183	3.817433	3.734683	In-control
3.899973	3.817875	3.735776	In-control
3.899969	3.817745	3.735522	In-control
3.899457	3.817105	3.734754	In-control
3.900235	3.817901	3.735567	In-control
3.900266	3.818138	3.736009	In-control
3.899793	3.817553	3.735314	In-control
3.900891	3.818070	3.735249	In-control
3.899879	3.817843	3.735806	In-control
3.900303	3.817711	3.735119	In-control
3.899988	3.817285	3.734583	In-control
3.900272	3.817237	3.734202	In-control
3.899423	3.817221	3.735018	In-control
3.899181	3.817042	3.734902	In-control
3.899181	3.817022	3.734862	In-control
3.900298	3.817679	3.735061	In-control
3.900225	3.817659	3.735093	In-control
3.900152	3.817387	3.734623	In-control
3.900463	3.817348	3.734233	In-control
3.899939	3.817813	3.735687	In-control
3.899884	3.817897	3.735911	In-control
3.900545	3.817807	3.735069	In-control
3.90005	3.818053	3.736056	In-control
3.899764	3.817109	3.734455	In-control
3.900386	3.817791	3.735196	In-control
3.899959	3.817511	3.735062	In-control
3.899905	3.817429	3.734952	In-control
3.899195	3.816681	3.734167	In-control
3.899326	3.817301	3.735277	In-control
3.899648	3.816942	3.734237	In-control
3.899549	3.817576	3.735603	In-control
3.899507	3.81727	3.735033	In-control
3.900055	3.816914	3.733773	In-control
3.899708	3.817343	3.734979	In-control
3.900194	3.817774	3.735353	In-control
3.900445	3.817423	3.734401	In-control
3.900095	3.817479	3.734863	In-control
3.900643	3.818006	3.735369	In-control
3.899956	3.817829	3.735702	In-control
3.899821	3.817094	3.734366	In-control
3.900089	3.817403	3.734717	In-control
3.898954	3.817156	3.735358	In-control

Table 4 above show the results obtained from equation (10), which is the difuzzified values set against the difuzzified control limits. The results indicate that the process is in control, since all the sample points are within the control limits.

6. Conclusion

Several publications with great application surfaces in the literature recently on statistical process control incorporated with fuzzy system. Amalgamation of different SPC tools with type-1 fuzzy, intuitionistic fuzzy, hesitant fuzzy, type-2 fuzzy, and recently the interval type-2 fuzzy control chart has come into existence in recent publications. However, no publication exists in Interval Type-2 Exponentially Weighted Moving Average (IT2FEWMA) control chart.

This paper extends the control limits of the classical control chart of the exponentially weighted moving average (EWMA). The IT2FEWMA is advantageous over the classical EWMA due to its flexibility over the control limits, but it is not capable of detecting a big shift in the process due to the fact that classical EWMA does not have such capacity too. This paper is a new addition to the existing Statistical Process Control Tools. It is useful when the process engineer needs to monitor a process whose measurement is obtained in fuzzy environment and a small shift needs to be detected.

Future research: Multivariate IT2FEWMA control chart, interval type-2 intuitionistic FEWMA control chart and interval type-2 Hesitant FEWMA control chart can be developed, also comparative studies can be established between the IT2FEWMA and FEWMA control chart.

References

Adepoju A. A., (2018). Performance of Fuzzy Control Chart over the Traditional Control Chart. Benin Journal of Statistics, Vol 1, pp. 101–112.

Adepoju A. A, Isah A. M., Ahmed S., Wasiu Y. A., Samuel A. N., Ibrahim A., (2019). Trapezoid Fuzzy-Shewhart Control Chart Based on α-Level Mid-range Transformation and its Sensitivity Measures. Professional Statisticians Society of Nigeria. Edited Proceedings of 3rd International Conference, Vol. 3, pp. 488–492

Adepoju A. A. Mohammed U., Adamu K., Agog N. S., Isah A. M. Yekini W. A. (2019). Interval Type-2 Fuzzy Control Chart for Non-Conformity Per Unit.

Professional Statisticians Society of Nigeria. Edited Proceedings of 3rd International Conference, Vol. 3, pp. 493–498.

Adepoju A. A., Mohammed U., Sani S. S., Adamu K., Tukur K, Ishaq A. I., (2019). Statistical Properties of Negative Binomial distribution under Imprecise Observation. Journal of the Nigerian Statistical Association, Vol. 31.

Castillo, O., Cervantes, L., Soria, J., Sanchez, M., Castro, J. R. (2016). A generalized type-2 fuzzy granular approach with applications to aerospace. Inf. Sci. 354, pp. 165–177

Castillo O., Leticia, A., Castro, J. R., Mario Garcia-Valdez A., (2016). Comparative study of type-1 fuzzy logic systems, interval type-2 fuzzy logic systems and generalized type-2 fuzzy logic systems in control problems. Inf. Sci. 354, pp. 257–274

Cervantes, L., Castillo, O. (2015). Type-2 fuzzy logic aggregation of multiple fuzzy controllers for airplane flight control. Inf. Sci. 324, pp. 247–256

Chen, L.-H., Huang, C.-H., (2016). Design of a fuzzy zone control chart for improving the process variation monitoring capability. Journal of Applied Sciences, 16, pp. 201–208.

Cheng, C. B., (2005). Fuzzy process control: construction of control charts with fuzzy numbers. Fuzzy sets System 154, pp. 287–303.

El-Shal, S. M., Morris, A. S., (2000). A fuzzy rule-based algorithm to improve the performance of statistical process control in quality systems. *Journal of Intelligent & Fuzzy Systems*, 9, pp. 207–223.

Ercan H., Anagun A., (2018). Different methods to fuzzy $\overline{X}$ -R control charts used in production: Interval type-2 fuzzy set example, *Journal of Enterprise Information Management*, vol.31, no. 6, pp. 848-866.

Erginel, N., (2008). Fuzzy individual and moving range control charts with α -cuts. *Journal of Intelligent & Fuzzy Systems*, 19, pp. 373–383.

Erginel, N., (2014). Fuzzy rule based p-np control charts. *Journal of Intelligent & Fuzzy Systems*, 27, 159–171. Monitoring capability. *Journal of Applied Sciences*, 16, pp. 201–208.

Erginel N. Senturk S. Yildiz G., (2018). Modeling Attribute Control Charts by Interval type-2 Fuzzy sets. *Soft computing,* 22, p. 5033

Erginel N, Senturk S, Kahraman C, Kaya I., (2011). Evaluating the packing process in food industry using Fuzzy $\tilde{X}$ and S control charts. Int JComput Intell Syst 4(4), pp. 509–520

Gulbay, M., Kahraman, C., (2006a). Development of fuzzy process control charts and fuzzy unnatural pattern analyses. *Computational Statistics and Data Analysis*, 51, pp. 434–451.

Gulbay, M., Kahraman, C., Ruan, D., (2004). α-cut fuzzy control charts for linguistic data. *International Journal of Intelligent Systems*, 19, pp. 1173–1196.

Hou, S.,Wang, H., Feng, S., (2016). Attribute control chart construction based on fuzzy score number. *Symmetry*, 8, pp. 3–13.

Kahraman, C., Oztayşi, B., Sarı, İ. U., Turano˘glu, E., (2014). Fuzzy analytic hierarchy process with interval type-2 fuzzy sets. *Knowledge – Based Systems,* 59, pp. 48–57.

Kanagawa, A., Tamaki, F., Ohta, H., (1993). Control charts for process average and variability based on linguistic data. *Intelligent Journal of Production Research*, 31(4), pp. 913–922.

Karnik, N. N., Mendel, J. M., (2001). Operations on type-2 fuzzy sets. *Fuzzy Sets and Systems*, 122, pp. 327–348.

Kaya, İ., Kahraman, C., (2011). Process capability analyses based on fuzzy measurement and fuzzy control charts. *Expert Systems with Applications*, 38, pp. 3172–3184.

Kaya, İ., Erdoğan, M., Yıldız, C., (2017). Analysis and control of variability by using fuzzy individual control charts. *Applied Soft Computing*, 51, pp. 370–381.

Keshavarz G. M., Zavadskas, E. K., Amiri, M., Antucheviciene, J., (2016). A new method of Assessment based on fuzzy ranking and aggregated weights (AFRAW) for MCDM problems under type-2 fuzzy environment. *Economic Computation and Economic Cybernetics Studies and Research,* 50(1), pp. 39–68.

Mehdi K. G., Maghsoud A., Jamshid S. S., Edmundas K. Z. (2015). Multi-Criteria Project Selection Using an Extended VIKOR Method with Interval Type-2 Fussy

Sets. *International Journal of Information Technology and Decision Making*, 14(5), pp. 993–1016

Mendel J. M., John, R. I (2002). Type-2 fuzzy sets made simple. IEE transactions on Fuzzy System, 10(2), 117-127.

Mendel, J.M., John, R.I., Liu, F. (2006). Interval type-2 fuzzy logic systems made simple. *IEE Transactions on Fuzzy Systems*, 14(6), pp. 808–821.

Montgomery D. C., (2009). Introduction to Statistical Quality Control, Sixth Edition. John Wiley & Sons, Inc.

Ontiveros-Robles, E., Melin, P., Castillo, O. (2018). Comparative analysis of noise robustness of type-2 fuzzy logic controllers. Kybernetika, 54(1), pp. 175–201

Poongodi, T., Muthulakshmi, S., (2015). Fuzzy control chart for number of customer of Ek /M/1 queueing model. *International Journal of Advanced Scientific and Technical Research*, 3(5), pp. 9–22.

Raz, T., Wang J. H. (1990) Probabilistic and memberships approaches in the construction of control chart for linguistic data, Production. Planning and Control, 1(3), 147-157.

Rowlands, H., Wang, L. R. (2000). An approach of fuzzy logic evaluation and control in SPC. *Quality Reliability Engineering Intelligent*, 16, 91–98.

Şenturk, S. (2010). Fuzzy regression control chart based on _-cut approximation. *International Journal of Computational Intelligence Systems*, 3(1), pp. 123–140.

Senturk S., Erginel N., Kaya I., Kahraman C., (2014). Fuzzy exponentially weighted moving average control chart for univariate data with a real case application, Applied Soft Computing, Vol 22, pp. 1–10

Şenturk, S., Erginel, N. (2009). Development of fuzzy $\tilde{\bar{X}} - \tilde{R}$ and $\tilde{\bar{X}} - \tilde{S}$ control charts using _ cuts. *Information Sciences*, 179, 1542–1551.

Sevil Senturk, Jurgita Antucheviciene. (2017). Interval Type-2 Fuzzy c-Control Charts: An Application in a Food Company. *Informatica*, Vol. 28, No. 2, pp. 269–283.

Wang, D., Hyrnıewıcz, O. (2015). A fuzzy nonparametric shewhart charts based on the bootstrap approach. *International Journal of Applied Mathematics and Computer Science*, 25, pp. 389–401.

Wang, J. H., Raz, T. (1990). On the construction of control charts using linguistic variables. *Intelligent Journal of Production Research*, 28, pp. 477–487.

Zadeh L. A., (1965) Fuzzy sets, Information and Control, 8(3), pp. 338–353.

Zadeh L. A., (1975). The concept of a linguistic variable and its application to approximate reasoning_I, Information Sciences 8(3) 199–249.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23 No. 1, pp. 201–212, DOI 10.21307/stattrans-2022-012
Received – 04.02.2021; accepted – 23.08.2021

Estimating the confidence interval of the regression coefficient of the blood sugar model through a multivariable linear spline with known variance

Anna Islamiyati[1], **Raupong**[2], **Anisa Kalondeng**[3], **Ummi Sari**[4]

ABSTRACT

Estimates from confidence intervals are more powerful than point estimates, because there are intervals for parameter values used to estimate populations. In relation to global conditions, involving issues such as type 2 diabetes mellitus, it is very difficult to make estimations limited to one point only. Therefore, in this article, we estimate confidence intervals in a truncated spline model for type 2 diabetes data. We use a non-parametric regression model through a multi-variable spline linear estimator. The use of the model results from the irregularity of the data, so it does not form a parametric pattern. Subsequently, we obtained the interval from beta parameter values for each predictor. Body mass index, HDL cholesterol, LDL cholesterol and triglycerides all have two regression coefficients at different intervals as the number of the found optimal knot points is one. This value is the interval for multivariable spline regression coefficients that can occur in a population of type 2 diabetes patients.

Key words: confidence interval, diabetes, known variance, spline.

1. Introduction

There are two commonly known regression coefficient estimates, namely point and interval estimation. Both are valid for all regression approaches, including nonparametric regression. Non-parametric regression is used when the assumption of the relationship between the predictor and the response is unknown, so we must estimate its function. Some estimators that have been developed include spline truncated (Aprilia, Islamiyati and Anisa, 2019), spline smoothing (Lestari, Budiantara

[1] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. E-mail: annaislamiyati701@gmail.com (correspondence author). ORCID: https://orcid.org/0000-0001-6441-0306.

[2] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. E-mail: raupong.stat.uh@gmail.com..

[3] Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. E-mail: nkalondeng@gmail.com.

[4] Hasanuddin University Teaching Hospital, Indonesia, E-mail: ummisari.rsunhas@gmail.com.

and Chamidah, 2019), spline penalized (Islamiyati, Fatmawati and Chamidah, 2019), local polynomial (Chamidah, Gusti, Tjahjono and Lestari, 2019), Kernel (Chamidah and Saifuddin, 2013), Fourier series (Mardianto, Tjahjono and Rifada, 2020), and Gaussian process (Saegusa, 2020) , and spline principal component analysis (Islamiyati, Kalondeng, Sunusi, Zakir and Amir, 2022). For this article, we use spline truncated to estimate multi-variable non-parametric regression functions. The flexibility of the estimator by involving the knot point causes easy visual interpretation. This has become one of the main advantages of spline truncated in real applications. In addition to the knot point, the spline also considers the optimal order, which works simultaneously in the estimation model.

Confidence interval estimates are performed in cases of known or unknown variance. It depends on the available information related to population variance. We often find there is a case that was investigated by several researchers with different methods. These studies can provide information about the condition of the population variance. For example, diabetes data is a global disease that has been widely studied in various fields of study. Some studies consider the results of measurements of random blood sugar levels of patients using a penalized spline estimator of one smoothing parameter (Islamiyati, Fatmawati and Chamidah, 2020) and two smoothing parameters (Islamiyati, Sunusi, Kalondeng, Fatmawati and Chamidah, 2020). Also, some consider the patient's fasting blood sugar data (Islamiyati, Raupong and Anisa, 2019), the patient's calorie diet (Islamiyati, Fatmawati and Chamidah, 2020), and the detection of lifestyle of diabetic patients (Islamiyati, 2022). In the country of India, a Genome-Wide Association Scan study identified more than 65 common genetic variants associated with type 2 diabetes (Singh, 2015). In Indonesia, especially in Makassar City, blood sugar after meals has a variance based on the measurement results through weighted penalized spline (Islamiyati, Fatmawati and Chamidah, 2018).

However, the studies still consider a single point estimate in the non-parametric regression approach. That can cause a large difference from the estimated value of the regression coefficients obtained for research at different locations and times. Some studies on estimating confidence intervals that provide wider tolerance values in non-parametric regression coefficients, for example, a comparison of results of confidence interval estimates from spline smoothing with Bayesian (Wang and Wahba, 2003), estimation of confidence intervals with uniform distribution on non-parametric regression curves (David, Tom and Douglas, 2001) and the use of B spline estimators in polynomial spline (Mao and Zhao, 2003). Therefore, we reviewed the estimated confidence interval in diabetes data using a spline truncated. For application consideration, we use a linear spline on the dimensions of many predictor variables. The results of this article are expected to provide lower and upper limits of an interval of spline regression coefficient values from diabetes data. Diabetes data were obtained

from the Hasanuddin University Teaching Hospital in Makassar, Indonesia by taking factors in body mass index, HDL and LDL cholesterol, and triglycerides.

Furthermore, the core content of this article discusses the theoretical form of the model and proceeds to the application of diabetes data using the linear spline. The knot point being tried is limited to the use of three knots with the optimal knot point selection method through Generalized Cross-Validation (GCV) values. Based on the theory that has been widely used by non-parametric regression that a regression model that provides a minimum GCV value means that the model is the optimal model used in interpreting data conditions.

2. Methodology

2.1. Data source

Data on patients with type 2 diabetes mellitus were obtained from the Hasanuddin University Teaching Hospital, Makassar, Indonesia. The data were recorded from the medical records of diabetic patients who were hospitalized from 2014-2018. We selected 84 people as samples in this study because they had complete media records according to the factors studied (Appendix 1).

2.2. Multi-variable linear spline models in the non-parametric regression approach

The non-parametric regression function that contains one response and several predictor variables is estimated with a linear spline so it is called a multi-variable linear spline model. Multi-variable non-parametric regression models can be stated as follows:

$$y_i = f\left(t_{1i}, t_{2i}, \ldots, t_{pi}\right) + \varepsilon_i \tag{1}$$

where y_i is the response in the *i*-samples, and $f\left(t_{1i}, t_{2i}, \ldots, t_{pi}\right)$ is a function in the predictor $t_1, t_2, \ldots, t_p$ and ε_i is an error in the *i*-samples.

The function $f\left(t_{1i}, t_{2i}, \ldots, t_{pi}\right)$ in equation (1) is estimated with a spline truncated estimator. Each function in each predictor can be stated as follows:

$$\left.\begin{aligned} f\left(t_{1i}\right) &= \alpha_{10} + \alpha_{11} t_{1i} + \sum_{u_1=1}^{s_1} \alpha_{1(1+u_1)} \left(t_{1i} - k_{1u_1}\right)_+ \\ f\left(t_{2i}\right) &= \alpha_{20} + \alpha_{21} t_{2i} + \sum_{u_2=1}^{s_2} \alpha_{2(1+u_2)} \left(t_{2i} - k_{2u_2}\right)_+ \\ &\vdots \\ f\left(t_{pi}\right) &= \alpha_{p0} + \alpha_{p1} t_{pi} + \sum_{u_p=1}^{s_p} \alpha_{p(1+u_p)} \left(t_{pi} - k_{pu_p}\right)_+ \end{aligned}\right\} \tag{2}$$

If equation (1) is expressed as a sum of the functions of each predictor, namely:

$$y_i = f\left(t_{1i}\right) + f\left(t_{2i}\right) + \ldots + f\left(t_{pi}\right) + \varepsilon_i ,$$

then the multi-variable spline regression model can be stated as follows:

$$y_i = \alpha_0 + \sum_{h=1}^{p}\left(\alpha_{h1} t_{hi} + \sum_{u_h=1}^{s_h} \alpha_{h(1+u_h)} \left(t_{hi} - k_{hu_h}\right)_+ \right) + \varepsilon_i \tag{3}$$

Equation (3) can be expressed in matrix form, namely:

$$\mathbf{y} = \mathbf{T}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{y} = \left(y_1, y_2, \ldots, y_n\right)^T$ is the response to $i = 1,2,\ldots,n$, $\mathbf{T} = \left(\mathbf{1}, \mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_p\right)$ is a predictor matrix containing knots, $\boldsymbol{\alpha} = \left(\alpha_0, \alpha_{11}, \alpha_{12}, \ldots, \alpha_{1s_1}, \alpha_{21}, \alpha_{22}, \ldots, \alpha_{2s_2}, \ldots, \alpha_{p1}, \alpha_{p2}, \ldots, \alpha_{ps_p}\right)^T$ is regression coefficient of linear multi-variable spline, and $\boldsymbol{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\right)^T$ is an error. As for

$\mathbf{1} = \left(1,1,\ldots,1\right)^T$, $\mathbf{T}_1 = \left(\mathbf{t}_1, \left(\mathbf{t}_1 - k_{11}\right), \left(\mathbf{t}_1 - k_{12}\right), \ldots, \left(\mathbf{t}_1 - k_{1s_1}\right)\right)$, $\mathbf{T}_2 = \left(\mathbf{t}_2, \left(\mathbf{t}_2 - k_{21}\right), \left(\mathbf{t}_2 - k_{22}\right), \ldots, \left(\mathbf{t}_2 - k_{2s_2}\right)\right)$, $\ldots, \mathbf{T}_p = \left(\mathbf{t}_p, \left(\mathbf{t}_p - k_{p1}\right), \left(\mathbf{t}_p - k_{p2}\right), \ldots, \left(\mathbf{t}_p - k_{ps_p}\right)\right)$.

Estimation of multi-variable spline regression parameter $\boldsymbol{\alpha}$ is obtained through the least square method by minimizing the sum of the squares of the error in equation (4).

$$\begin{aligned} \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} &= \left(\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\right)^T \left(\mathbf{y} - \mathbf{T}\boldsymbol{\alpha}\right) \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\alpha}^T\mathbf{T}^T\mathbf{y} + \boldsymbol{\alpha}^T\mathbf{T}^T\mathbf{T}\boldsymbol{\alpha} \end{aligned} \tag{5}$$

Next, equation (5) is derived from the vector $\boldsymbol{\alpha}$ and the resulting derivation is equated to zero.

$$\left.\frac{\partial\left(\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\right)}{\partial\boldsymbol{\alpha}}\right|_{\boldsymbol{\alpha}=\hat{\boldsymbol{\alpha}}} = -2\mathbf{T}^T\mathbf{y} + 2\mathbf{T}^T\mathbf{T}\boldsymbol{\alpha} \tag{6}$$

If equation (6) is equated to zero, then we get:

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y} \tag{7}$$

From equation (7), the estimation of the multi-variable spline regression model is as follows:

$$\hat{\mathbf{y}} = \mathbf{T}\hat{\boldsymbol{\alpha}}$$

where $\hat{\boldsymbol{\alpha}}$ according to equation (7). The Generalized Cross-Validation (GCV) formula for the model is obtained as follows:

$$GCV\left(\mathbf{k}\right) = \frac{MSE\left(\mathbf{k}\right)}{n^{-1}tr\left[\mathbf{I} - \mathbf{A}\left(\mathbf{k}\right)\right]^2} \tag{8}$$

where $\mathbf{A}\left(\mathbf{k}\right) = \mathbf{T}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T$.

2.3. Estimates of confidence intervals with variance are known

We use the pivotal quantity approach (Toulis, 2017) in estimating the confidence interval for the multi-variable spline regression coefficient. First, we determine the expected value of $\hat{\boldsymbol{\alpha}}$, namely:

$$\begin{aligned} E(\hat{\boldsymbol{\alpha}}) &= E\left(\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y}\right) \\ &= \left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{T}\boldsymbol{\alpha} + E(\boldsymbol{\varepsilon}) \\ &= \boldsymbol{\alpha} \end{aligned} \tag{9}$$

Second, we determine the variance of $\hat{\boldsymbol{\alpha}}$, namely:

$$\begin{aligned} Var(\hat{\boldsymbol{\alpha}}) &= Var\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{y} \\ &= \sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\mathbf{T}^T\mathbf{T}\left(\mathbf{T}^T\mathbf{T}\right)^T \\ &= \sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1} \end{aligned} \tag{10}$$

Based on equations (9) and (10), the estimate of $\hat{\boldsymbol{\alpha}}$ follows the normal distribution $\hat{\boldsymbol{\alpha}} \sim N\left(\boldsymbol{\alpha}, \sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1}\right)$, namely mean $\boldsymbol{\alpha}$ and variance $\sigma^2\left(\mathbf{T}^T\mathbf{T}\right)^{-1}$.

Given a form of transformation from $Z_{hu_h} = \dfrac{\hat{\alpha}_{hu_h} - \alpha_{hu_h}}{\sqrt{\sigma^2 d_{hh}}}$, $d_{hh} = \text{diag}\left(\mathbf{T}^T\mathbf{T}\right)^{-1}$ and $Z_{hu_h} \sim N(0,1)$. This means that the expectation and variance values of Z_{hu_h} are 0 and 1. The effect σ^2 is known to cause $Z_{hu_h}(t_1, t_2, \ldots, t_n)$ to be the pivotal quantity for parameter α, so the confidence interval can be stated as follows:

$$P\left(a \le Z_{hu_h}(t_1, t_2, \ldots, t_n) \le b\right) = 1 - \gamma, \tag{11}$$

where a and b are real number elements, $a < b$.
Equation (11) can also be stated as:

$$P\left(a \le \frac{\hat{\alpha}_{hu_h} - \alpha_{hu_h}}{\sqrt{\sigma^2 d_{hh}}} \le b\right) = 1 - \gamma \tag{12}$$

Equation (12) can be worked on to become:

$$\alpha_{hu_h} = \hat{\alpha}_{hu_h} - a\sqrt{\sigma^2 d_{hh}} \text{ dan } \alpha_{hu_h} = \hat{\alpha}_{hu_h} - b\sqrt{\sigma^2 d_{hh}}. \tag{13}$$

Based on equation (13), the confidence interval for the multi-variable linear spline regression parameters is:

$$P\left(\hat{\alpha}_{hu_h} - b\sqrt{\sigma^2 d_{hh}} \le \alpha_{hu_h} \le \hat{\alpha}_{hu_h} - a\sqrt{\sigma^2 d_{hh}}\right) = 1 - \gamma$$

Furthermore, the shortest confidence interval is obtained from conditional optimization by the Lagrange method, namely $\min_{a,b\in R}\{a,b\} = \min_{a,b\in R}\left\{(b-a)\sqrt{\sigma^2 d_{hh}}\right\}$ with the constraint function $g(b)-g(a)-(1-\gamma)=0$, where g is the cumulative probability distribution $N(0,1)$. Furthermore, the Lagrange function can be expressed as:

$$F(a,b,\lambda) = (b-a)\sqrt{\sigma^2 d_{hh}} + \lambda\left(g(b)-g(a)-(1-\gamma)\right) \tag{14}$$

where λ is the Lagrange constant. From the results of the partial derivative of the parameter a,b,λ, we get $a=-b$, which satisfies the equation. If it is substituted into equation (14), then it is obtained:

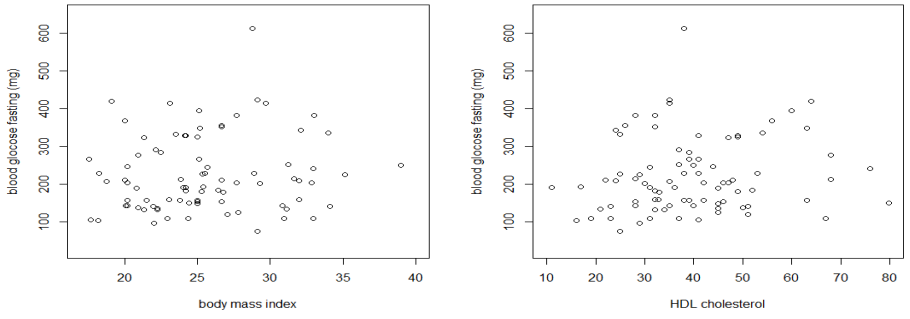
$$P\left(\hat{\alpha}_{hu_h} - b\sqrt{\sigma^2 d_{hh}} \le \alpha_{hu_h} \le \hat{\alpha}_{hu_h} + b\sqrt{\sigma^2 d_{hh}}\right) = 1-\gamma \tag{15}$$

For $b = Z_{\gamma/2}$, equation (15) can also be stated as:

$P\left(\hat{\alpha}_{hu_h} - Z_{\gamma/2}\sqrt{\sigma^2 d_{hh}} \le \alpha_{hu_h} \le \hat{\alpha}_{hu_h} + Z_{\gamma/2}\sqrt{\sigma^2 d_{hh}}\right) = 1-\gamma$, where γ is the level of significance used in research that researchers usually use 0.05.

3. Analysis and discussion

Blood sugar data of type 2 DM patients were analysed using non-parametric regression, in this case, multi-variable linear spline regression. This is because changes in a person's blood sugar can change very quickly and do not follow a certain trend. The condition is also shown in Figure 1 through scatter plots between fasting blood sugar factor (y) with body mass index/BMI (t_1), high-density lipoprotein/HDL cholesterol (t_2), low-density lipoprotein/LDL cholesterol (t_3), and triglycerides/TG (t_4). Based on Figure 1, we can see a data plot that does not follow a certain parametric pattern, for example linear, quadratic, cubic, and others. Therefore, we analysed this data using a non-parametric multi-variable linear spline regression approach.



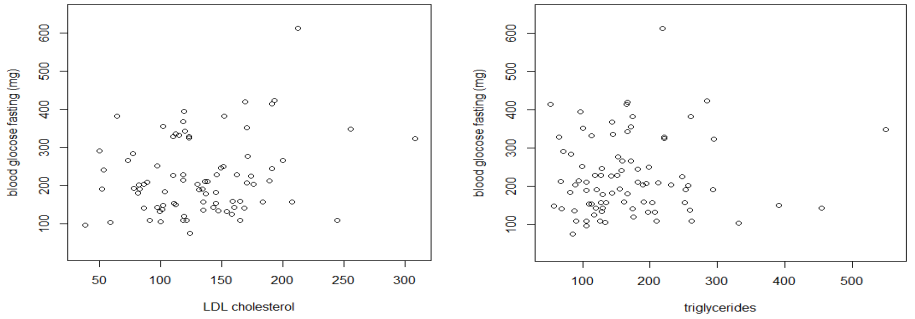


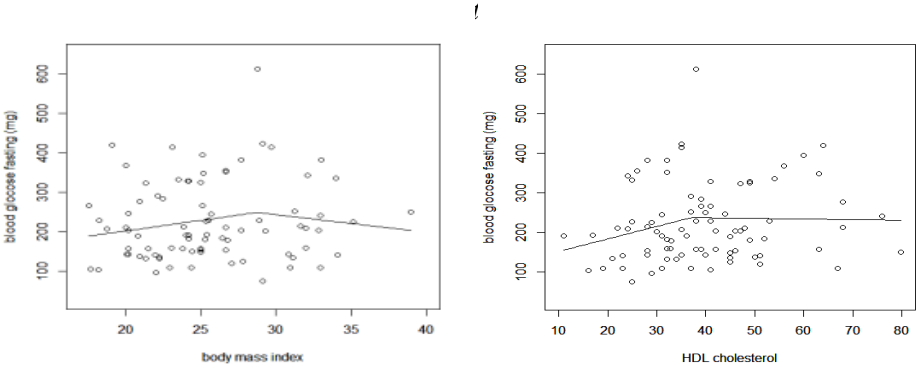
Figure 1. Scatter plot between blood sugar with body mass index, HDL cholesterol, LDL cholesterol, and triglycerides factors

Spline regression is related to the optimal knot point and the number of knots, so we need to find the values of these parameters through the minimum GCV value. Here, we show a comparison of GCV values from the use of 1, 2, and 3-knot points as in Table 1.

Table 1. GCV values at 1, 2, 3 knots

Number of knots	**1**	**2**	**3**
GCV value	**10,552.75**	11,033.11	10,568.35

Based on Table 1, we can see that the number of knot points that give the minimum GCV value in the multi-variable linear spline model is one knot. Therefore, the blood sugar data of type 2 DM patients were modelled with a linear spline multi-variable approach through one knot on all the predictor variables involved. Next, we obtain a multi-variable linear spline regression model as follows:



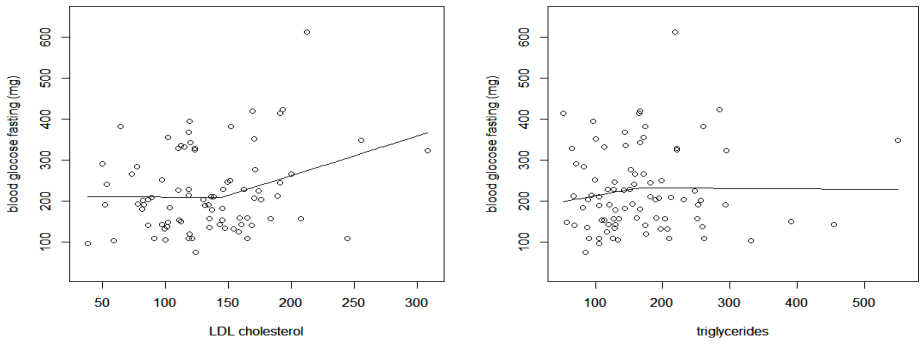


Figure 2. Estimation of multi-variable linear spline regression models with 1-knot point

The regression curve according to the model is shown in Figure 2. The optimal number of knot points is obtained at one knot for each predictor. The relationship of fasting blood glucose with Body Mass Index (BMI) is optimal at the knot point of 28.76 kg/m. These results indicate that there are two patterns of changes in fasting blood glucose based on that factor. We can see that it increases when the BMI is below the knot point, but after that, it decreases. This means that the increase based on BMI is in the pre-obesity stage. Excessive accumulation of fat in the body can cause insulin resistance which affects blood glucose levels. For variable t_2, the optimal HDL cholesterol at the one-knot point is 35 mg/dL. This shows that when HDL cholesterol reaches the point of knot, there is the accumulation of fat in certain organs. The incident triggers an increase in fasting blood glucose which can cause atherosclerosis or narrowing of the blood vessels and heart. Furthermore, fasting blood glucose has decreased slowly, which indicates an attempt to decrease after HDL increases. This certainly shows a positive trend from patients in controlling their blood sugar.

For variable t_3, the optimal LDL cholesterol at the one-knot point is 145 mg/dL. The pattern that is formed is that fasting blood glucose decreases slowly until LDL reaches that point. This means that diabetes patients keep their fasting blood glucose to prevent other diseases. In the next pattern, we see that fasting blood sugar levels increase very sharply when LDL cholesterol is more than the knot point. These results indicate the need to evaluate the patient's efforts to lower blood sugar when LDL is high. Furthermore, the optimal model for the variable t_4, namely triglycerides, is obtained at the knot point 167 mg/dL. These results indicate fasting blood glucose increases when the triglycerides are below the knot point, and after that, it decreases slowly. This means that for high triglycerides, there are patients' efforts to keep their blood sugar from rising.

Table 2. Estimated confidence intervals from spline regression coefficients for blood sugar data for type 2 DM patients at a confidence level of 95%

Variable	Parameter	Estimation of Parameter	Lower limit	Upper limit
	β_0	-76.58	-375.79	222.62
t_1	β_{11}	6.95	-1.67	15.57
t_1	β_{12}	-12.60	-34.70	9.50
t_2	β_{21}	3.91	-1.54	9.36
t_2	β_{22}	-4.27	-11.22	2.68
t_3	β_{31}	-0.38	-1.39	0.62
t_3	β_{32}	1.47	-0.26	3.20
t_4	β_{41}	0.36	-0.44	1.16
t_4	β_{42}	-0.59	-1.64	0.47

Point estimation results that have been obtained, proceed to the estimation of the confidence interval so that estimation results can be more accurate to explain the condition of the population. However, we first test the residual assumptions, especially the normal distribution. The normality assumption test is based on the Kolmogorov Smirnov test (Steinskog, Tjostheim and Kvamsto, 2007). We get the p-value, $p = 0.06 > \gamma = 0.05$, which means the residual is normally distributed. Point estimation results from the best multi-variable spline model for fasting blood glucose at the Hasanuddin University Teaching Hospital were used to construct confidence intervals for regression parameters. The results of the estimated upper and lower limit obtained are in Table 2. The table shows the model regression coefficient interval that contains a lower limit and an upper limit. These interval values provide an overview of the condition of the diabetic patient population, especially for the patient's blood sugar model based on body mass index, HDL cholesterol, LDL cholesterol, and triglycerides.

4. Conclusions

The estimation results of the regression coefficient interval of the multi-variable spline model provide interval values from the lower limit to the upper limit at the 95% confidence level. Confidence intervals represent the possible values in a diabetic patient population related to fasting blood glucose, BMI, LDL cholesterol, HDL cholesterol, and triglycerides. For type 2 diabetes, the confidence interval of the regression coefficient in Table 2 can be used as reference material for further research. The influence of these four factors of changes in blood sugar is shown by the pattern of changes in each regression curve obtained from a multi-variable simultaneous model.

Body mass index, cholesterol, and triglyceride factors can all cause an increase in blood sugar due to the accumulation of fat and cholesterol. We can pay attention to the results of the analysis, which show not an upward trend that occurs in the regression curve, but there are variations in the pattern that occurs. Segmentation of pattern changes can be demonstrated by multi-variable spline due to the presence of optimal knot parameters obtained from the minimum GCV value. The up and down patterns that occur in one regression curve indicate that many factors do indeed affect blood sugar in DM type 2 patients. Therefore, we suggest that further research adds a higher dimension to the variable by taking into account assumptions that can be violated for the data dimensions the big one.

Acknowledgment

Many thanks to the Deputy of Research and Development Strengthening, Ministry of Research and Technology/National Agency for Research and Innovation, the Republic of Indonesia for the Basic Research with the research contract No: 7/E1/KP.PTNBH/2021 dated March 8, 2021.

References

Aprilia, B., Islamiyati, A., Anisa, (2019). Platelet modeling based on hematocrit in DHF patients with spline quantile regression, *International Journal of Academic and Applied Research*, Vol. 3(12), pp. 51–54.

Chamidah, N., Gusti, K. H., Tjahjono, E., Lestari, B., (2019). Improving of classification accuracy of cyst and tumor using local polynomial estimator, *Telkomnika*, Vol. 17, pp. 1492–1500.

Chamidah, N., Saifudin, T., (2013). Estimation of children growth curve based on kernel smoothing in multi-response nonparametric regression, *Applied Mathematical Sciences*, Vol. 7(37), pp. 1839–1847.

David, J.C., Tom, G. F., Douglas, N., (2001). Confidence intervals for nonparametric curve estimates toward more uniform pointwise coverage, *Journal of the American Statistical Association*, Vol. 96(453), pp. 233–246.

Islamiyati, A., (2022). Spline longitudinal multi-response model for the detection of lifestyle-based changes in blood glucose of diabetic patients. *Current Diabetes Reviews*, Published on: 14 January.

Islamiyati, A., Fatmawati, Chamidah, N., (2018). Estimation of covariance matrix on bi-response longitudinal data analysis with penalized spline regression, *Journal of Physics: Conference Series*, Vol. 979, 012093.

Islamiyati, A., Fatmawati, Chamidah, N., (2020). Changes in blood glucose 2 hours after meals in Type 2 diabetes patients based on length of treatment at Hasanuddin University Hospital, Indonesia, *Rawal Medical Journal*, Vol. 45(1), pp. 31–34.

Islamiyati, A., Kalondeng, A., Sunusi, N., Zakir, M., Amir, A. K., (2022). Biresponse nonparametric regression model in principal component analysis with truncated spline estimator. *Journal of King Saud University-Science*, Vol. 34, 101892, pp. 1–9.

Islamiyati, A., Sunusi, N., Kalondeng, A., Fatmawati, F., Chamidah, N., (2020). Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model, *Journal of Sciences: Islamic Republic of Iran*, Vol. 31(2), pp. 175–183.

Islamiyati, A., Raupong, Anisa, (2019). Use of penalized spline linear to identify change in pattern of blood sugar based on the weight of diabetes patients, *International Journal of Academic and Applied Research*, Vol. 3(12), pp. 75–78.

Islamiyati, A., Fatmawati, Chamidah, N., (2019). Ability of covariance matrix in bi-response multi-predictor penalized spline model through longitudinal data simulation, *International Journal of Academic and Applied Research*, Vol. 3(3), pp. 8–11.

Islamiyati, A, Fatmawati, Chamidah, N., (2020). Penalized spline estimator with multi smoothing parameters in bi-response multi-predictor regression model for longitudinal data, *Songklanakarin Journal of Science and Technology*, Vol. 42(4), pp. 897–909.

Lestari, B., Budiantara, I. N., Chamidah, N., (2019). Smoothing parameter selection method for multiresponse nonparametric regression model using smoothing spline and kernel estimators approaches. *Journal of Physics: Conference Series*, Vol. 1397, 012064.

Mao, W., Zhao, L. H., (2003). Free-knot polynomial splines with confidence intervals, *Journal of the Royal Statistical Society Series B,* Vol. 65(4), pp. 901–919.

Mardianto, M. F. F., Tjahjono, E., Rifada, M., (2020). Statistical modelling for prediction of rice production in Indonesia using semiparametric regression based on three forms of Fourier series estimator, *ARPN Journal of Engineering and Applied Sciences*, Vol. 14(15), pp. 2763–2770.

Toulis, P., (2017). A useful pivotal quantity, *The American Statistician*, Vol. 71(3), pp. 272–274.

Saegusa, T., (2020). Confidence band for a distribution function with merged data from multiple source, *Statistic in Transition*, Vol. 21(4), pp. 144–158.

Singh, S., (2015). Genetics of type 2 diabetes: advances and future prospect, *Journal of Diabetes and Metabolism*, Vol. 6(4), pp. 1–8.

Steinskog, D. J., Tjostheim, D. B., Kvamsto, N. G. A., (2007). A cautionary note on the use of the Kolmogorov–Smirnov Test for normality, *American Meteorological Society,* Vol. 135(3), pp. 1151–1157.

Wang, Y., Wahba, G., (2003). *Bootstrap confidence interval for smoothing spline and their comparison to Bayesian confidence intervals,* Technical Report University of Wisconsin-Madison USA, No. 193.

APPENDIX

Data on patients with type 2 diabetes mellitus at the Hasanuddin University Teaching Hospital, Makassar, Indonesia, 2014-2018.

Patient Number	Blood Sugar	BMI	HDL	LDL	Triglycerides
1	230	18.26	41	146	151
2	414	23.11	35	191	165
3	212	20	48	136	182
4	229	28.89	38	162	118
5	212	26.67	22	138	106
6	352	26.67	32	170	100
7	137	22.22	45	135	88
8	150	24.44	79.8	112	391
9	132	22.22	32	99	197
10	420	19.11	64	169	166
11	135	31.11	21	147	128
12	368	20	56	118	143
13	154	26.67	28	111	113
14	225	35.11	29	174	247
15	356	26.67	26	102	172
16	132	21.33	34	154	207
17	184	26.40	52	103	81
18	157	23.80	38	207	203
19	251	38.95	40	151	198
20	324	21.35	47	308	294
⋮	⋮	⋮	⋮	⋮	⋮
83	109	32.97	23	165	126
84	336	34	54	112	145

Data sharing information statement.

This study does not involve direct human interaction. Data were collected from patient medical records with the approval of hospital management. For the purposes of developing this study, the authors may share data through personal contacts.

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1 pp. 213

Statistics Poland and the Polish Statistical Association
are pleased to announce the organization of the 3rd Congress of Polish Statistics,
which will take place from 26 to 28 April 2022 in Krakow.

This year's Congress will commemorate the 110th anniversary
of the Polish Statistical Association.

For more information on this event look at:

https://kongres2022.stat.gov.pl/en/
https://kongres2022.stat.gov.pl/en/programme

On behalf of the IAOS authorities we are pleased to announce the organization of the conference which will take place from 26 to 28 April 2022 at the Convention Center in Cracow, Poland

The general theme of the IAOS-2022 Conference is:
Worthy Information for Challenging Times.

For more information on this event look at:
https://www.iaos-isi.org/index.php/latestnews/288-announcing-the-iaos-2022-conference
https://www.iaos-isi.org/index.php/latestnews/289-more-information-on-the-2022-iaos-conference-in-krakow

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 3 pp. 215

A New Role for Statistics: The Joint Special Issue of "Statistics in Transition New Series" (SiTns) and " Statystyka Ukraïny" (SU) is planned for June 2022

On behalf of the SiTns' Editorial Board and myself, and of Professor Oleksandr Osaulenko – Editor-in-Chief of the SU – I would like to invite interested researchers and practitioners to submit manuscripts on statistical production under war conditions and the role of statistics in describing the effects of foreign aggression on the functioning of the economy and society – with particular emphasis on humanitarian crisis and the degradation of people's well-being.

Inspired by the tragedy of the heroic Ukrainian nation currently experiencing war devastation, the interest in new challenges and tasks posed for statistics matches the demand for knowledge about consequences of barbarism that already seemed to be excluded from the civilized space, forever. It is not only about the European area but also about every other region or country that has similar experiences to Ukraine.

Wishing in this way to demonstrate the contribution of statistics to such knowledge, we would also like to not only record the horrors of the outrageous aggression, but also to provide statistical evidence on the human empathy and cross-border or transnational solidarity, as well as concern for the common values of all societies which want to live in peace.

Manuscripts should be submitted by May 15, 2022 electronically to the Editorial Office: sit@stat.gov.pl

More details will be provided on our web pages:
https://sit.stat.gov.pl/Index
https://su-journal.com.ua/index.php/journal

Włodzimierz Okrasa
Editor
Statistics in Transition new series

Oleksandr Osaulenko
Editor
Statystyka Ukraïny

STATISTICS IN TRANSITION new series, March 2022
Vol. 23, No. 1, pp. 217–222,

About the Authors

Abdulkadir Sauta Saidu is an Associate Professor of Biostatistics at the Department of Statistics and Operations Research, Modibbo Adama University, Yola, Nigeria. His research interests are statistical inference, data analysis, stochastic processes and models, statistical quality control, and multivariate analysis. He has published over 30 research papers in international/national journals and conferences. He is a member of many professional bodies.

Abu-Shawiesh Moustafa Omar Ahmed is a Professor of Statistics in the Department of Mathematics at the Hashemite University (HU) in Jordan. He has also taught at the King Faisal University, Saudi Arabia and Nizwa University, Oman. Dr. Abu-Shawiesh has diverse research interests, mainly robust statistical methods, statistical process control, artificial neural networks, statistical inference and time series analysis. Since 1997, he has over 36 full research articles published in different internationally well-reputed statistical journals. Dr. Abu-Shawiesh is a member of editorial boards for many international statistical journals, including Annals of Management Science and Journal of Mathematics and Statistics. He has also served as a peer reviewer to many international journals in statistics, mathematics and management.

Adepoju Akeem Ajibola is a Lecturer I at the Department of Statistics, Kano University of Science and Technology, Wudil, Kano State, Nigeria. Simultaneously, he holds the office of Student Industrial Work Experience Scheme (SIWES) and the Departmental Strategic Office in the same Department. His main research interest area includes fuzzy system, statistical quality control, distribution theory, Bayesian method/analysis, multivariate analysis, probability distribution and data analysis. He has published 12 papers in reputable journals and conference. He is an invited reviewer to Statistics in Transition (New series). He is currently a member of Royal Statistical Society Nigeria Local Group (RSS), Professional Statisticians Society of Nigeria (PSSN), and Nigeria Statistical Association (NSA).

Anisa Kalondeng is an Associate Professor in the Statistics Department at the Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. Her research fields of study are experimental design, categorical data analysis, and marketing research.

Chaturvedi Aditi received her PhD in Statistics from the Department of Statistics, Babasaheb Bhimrao Ambedkar University, Lucknow, India. She works in the broad

area of statistical inference and her primary area of research is reliability and life testing. She has presented her work in various international conferences. She received R.S. Verma Best Paper Award in Twenty-Seventh International Conference of FIM in Conjunction with Third convention of IARS on IMSCT organized by the Department of Statistics, University of Jammu, Jammu, India in 2018 and Best Theoretical Research Paper award in International Virtual Conference on Prof. C.R. Rao's School of Thought on Statistical Sciences organized by the Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University and the Department of Mathematics and Statistics Mrs. A.V.N. College, Vishakapatnam in 2020.

Chaudhuri Arijit, honorary Visiting Professor, Indian Statistical Institute, India, after serving there as a Full Professor for 25 consecutive years. He published 11 text books/monographs on survey sampling in USA (8), Europe (2) and India (1) and about 150 peer-reviewed research papers in international journals, by both himself alone and a few jointly with students and colleagues, and visited with academic assignments many universities/institutions abroad intermittently during 1973-2009.

Chiroma Haruna is an Associate Professor in University of Hafr Al Batin, college of Computer Science and Enginering, Saudi Arabia. He is an Associate Editor in IEEE Access (ISI WoS/Scopus indexed), IAES international Journal of artificial intelligence, (Scopus indexed) and Telecommunication, Computing, Electronic and Control Journal. He has a special interest in technology enhanced learning. He has published over 100 academic articles. He is an invited reviewer to 18 ISI WoS indexed journals. His research interest includes: machine learning with emphasis on deep learning, natured inspired algorithms and their applications in internet vehicles, self driving vehicles, big data analytics emerging cloud computing architecture, NLP and IoT.

Chipepa Fastel is a Lecturer at the Department of Mathematics and Statistical Sciences, Faculty of Science, Botswana International University of Science and Technology. He has authored and co-authored over thirty (30) research papers. His main areas of interest include: distributional theory, survival analysis, biostatistical methods and biometry. He is a member of two editorial boards: Annals of Immunology & Immunotherapy (AII) and International Journal of Mathematics, Statistics and Operations Research.

Golam Kibria B. M. is a Professor in the Department of Mathematics and Statistics at Florida International University (FIU). Dr. Kibria's research interests include applied statistics, distribution theory, quality control, linear model, ridge regression and statistical inference. Since 1993, he has about 225 full research articles published in different internationally well-reputed statistical journals and co-authored two books. Dr. Kibria has supervised as a major (or co-major) professor two PhD, 20 masters and

22 undergraduate students at FIU. He has served over 60 Ph.D. and MSc thesis committees at FIU and abroad. He is the recipient of Faculty Award for Excellence in Research & Creative Activities and Top Scholar Award among others from FIU. Dr. Kibria is a member of editorial boards of over 35 international statistical journals. He is the elected Fellow of Royal Statistical Society.

Goldoust Mehdi received his Ph.D. in Statistics from Amirkabir University of Technology (Tehran Polytechnic) in 2018. He is an Asistant Professor at the Department of Statistical Methods, Faculty of Mathematics, Azad University of Behbahan. Simultaneously she holds the position of an expert in the Amar Jooyan Institute in Ahvaz, Iran. His main areas of interest include distribution theory, reliability, and Entropy.

Hanif Muhammad completed his Master's degree from New South Wales University, Australia in Multistage Cluster Sampling. He completed his PhD in Statistics from the University of Punjab, Lahore, Pakistan. He has over 40 years of research experience. He is an author of over 200 research papers and 10 books. He has served as a Professor in various parts of the world, i.e. Australia, Libya, Saudi Arabia, and Pakistan. He is presently a Professor of Statistics and Vice-Rector (Research) at NCBA & E, Lahore, Pakistan. He is a HEC approved supervisor.

Islamiyati Anna is an Associate Professor at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. Her field of research is nonparametric regression, multivariate analysis, categorical data analysis, computational statistics, and statistical modelling in general. Currently, Anna Islamiyati continues to develop statistical research for applications in the health sector.

Janus Jakub is an Assistant Professor at the Department of Macroeconomics, Cracow University of Economics, Poland. His research interests cover monetary economics and open economy macroeconomics. In 2018, he defended the doctoral dissertation on unconventional monetary policy exit strategies, which was awarded the highest prize in the National Bank of Poland's competition for PhD dissertations. Recently, he has focused on cross-border relationships in interest rates, country risk, and exchange rates. Dr. Janus takes part in international research projects and internships, e.g. at the University of Lisbon and at the University of Heidelberg. He serves as an associate editor in Entrepreneurial Business and Economics Review and a conference secretary in Workshop on Macroeconomic Research.

Jibasen Danjuma is a Professor at the Department of Statistics & Operations Research, Modibbo Adama University, Yola. Simultaneously, his is the Coordinator, Intellectual Property and Technology Transfer Office of the same University. His main areas of interest include: Estimation of elusive and hard-to-reach population, stochastics modelling and categorical data analysis. Currently, he is a member of two editorial

boards: Bagale Journal of Physical and Life Sciences & Nigerian Annals of Pure and Applied Sciences.

Kowalczyk Barbara is an Associate Professor at the Collegium of Economic Analysis, SGH Warsaw School of Economics. Her main research interests include: mathematical statistics, survey sampling, survey methodology, application of statistical theory in social and economic studies. She is an author and co-author of many scientific articles in national and international journals, research studies, presentations at the international conferences, reviews, and two books. In addition to scientific activities she conducts a wide range of teaching activities focused around statistics, including cooperation with international universities.

Kumar Surinder is currently the Head of Department of Statistics, Babasaheb Bhimrao Ambedkar University (A central university), Lucknow, India. He has 21 years of teaching experience and 26 years of research experience in various research fields of statistics such as sequential analysis, reliability theory, business statistics and Bayesian inference. Prof. Kumar has published over 60 research publications in various journals of national and international repute. He is a member of various professional bodies such as ISPS, ICAR-IASRI and Indian Association for Reliability and Statistics (IARS).

Lahiri Partha is a Professor of Survey Methodology and Mathematics at the University of Maryland College Park and an adjunct research professor at the Institute of Social Research, University of Michigan, Ann Arbor. His areas of research interest include data linkages, Bayesian statistics, survey sampling and small-area estimation. Dr. Lahiri served on the editorial board of a number of international journals and on many advisory committees, including the U.S. Census Advisory committee and U.S. National Academy panel. He has also served as an advisor or consultant for various international organizations such as the United Nations and the World Bank. He is a Fellow of the American Statistical Association and the Institute of Mathematical Statistics and an elected member of the International Statistical Institute. Dr. Lahiri is the recipient of the 2020 SAE award for his outstanding contribution to the research, application, and education of small area estimation.

Moakofi Thatayaone is a PhD Statistics Student, Faculty of Science, Botswana International University of Science and Technology. His area of expertise include: distributional theory and reliability analysis. He has authored and co-authored over ten (10) research papers.

Mohammadpour Adel received his B.Sc., M.Sc., and Ph.D. in Statistics from Shiraz University in 1993, 1995, and 2000, respectively. Since 2000, he has been a Faculty Member with the Department of Mathematics and Computer Science, Amirkabir University of Technology (Tehran Polytechnic). He was post-doc in Laboratoire des

signaux et systems at Supélec (2003-2006). His current research interests include statistical inference for heavy-tailed and large-scale data.

Oluyede Broderick is a Full Professor of Mathematics and Statistics, and former Director of the Statistical Consulting Unit (SCU) in the Department of Mathematical Sciences, Georgia Southern University, Statesboro, Georgia, USA. He is currently a Full Professor of Mathematics and Statistics at Botswana International University of Science and Technology (BIUST). He has over thirty years of research and teaching experience at Bowling Green State University, Georgia State University, University of Georgia, Oklahoma State University, Georgia Southern University and BIUST. He has authored and co-authored over one hundred and fifty (150) research papers. His research interests include multivariate statistical analysis, distribution theory, reliability theory, survival analysis, categorical data analysis, biostatistics, order restricted inference, stochastic dependence and weighted distributions. He is a member of editorial boards of several journals.

Raupong is an Associate Professor at the Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Indonesia. His research fields of study are regression models, experimental design, and linear models.

Samaddar Sonakhya is working as Assistant Director as part of the Indian Statistical Service at the Ministry of Social Justice and Empowerment, Govt. of India. Her main areas of interest are development of new sampling practices, small area estimation, income inequality, poverty and official statistical system. She is a former Junior Research Fellow at Indian Statistical Institute, Kolkata.

Sen Aditi is currently a PhD student and Research Assistant at the Department of Mathematics, University of Maryland, College Park. She has a master's degree in Statistics from the University of Calcutta in India. She has had four years of industry experience in data and analytics working as Data Scientist for the Hong Kong and Shanghai Banking Corporation. Her main work involved building analytical solutions and statistical models for the retail banking domain using languages and software like SAS, R and Python. During the graduate study, her areas of interests lie in statistical topics like small area estimation, missing data techniques with applications to sample surveys from various domains.

Sinsomboonthong Juthaphorn is an Associate Professor at Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand. Her interested research areas are statistical inference, probability theory, quality control, statistical analysis and data science. Dr. Juthaphorn is the Head of Doctor of Philosophy Program in Statistics at Kasetsart University.

ul-Amin Muhammad Noor received his PhD degree from NCBA&E, Lahore, Pakistan. He has working experience in various universities for teaching and research that

includes the Virtual University of Pakistan, University of Sargodha, Pakistan, and the University of Burgundy, France. He is currently working as an Assistant Professor at COMSATS University Islamabad Lahore Campus. His research interests include sampling techniques and control charting techniques. He is a HEC approved supervisor.

Wieczorkowski Robert received his PhD degree in mathematics from the Institute of Mathematics, Warsaw University of Technology, Poland, in 1995. Currently, he is a consultant at Programming and Coordination of Statistical Surveys Department, Statistics Poland. His main areas of interest include: application of survey sampling theory in social and agricultural surveys, computational statistics and application of numerical methods in socio-economic studies. Dr Robert Wieczorkowski is an co-author of many methodological publications published by Statistics Poland.

Yabaci Ayşegül is a graduate of Yıldız Technical University, Department of Statistics. She completed her master's and doctoral studies at Uludağ University Faculty of Medicine, Department of Biostatistics. He is currently working as a Research Assistant in the Department of Biostatistics at Bezmialem Vakıf University Faculty of Medicine. His research interests are survival analysis, Effect Size, and Fuzzy Logic and Clustering.

Yasmeen Uzma is a PhD from the National College of Business Administration & Economics, Lahore, Pakistan. She has worked at the University of Waterloo, Canada, National University of Modern Languages, Lahore Campus and COMSATS University Islamabad. Currently, she is working as an Assistant Professor at the Institute of Molecular Biology and Biotechnology, The University of Lahore, Lahore Campus. Her research interest is sampling procedures and biostatistics. She is a Higher Education Commission (HEC) approved supervisor.

© Copyright by Polish Statistical Association, Statistics Poland and the authors, some rights reserved. CC BY-SA 4.0 licence

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).