Volume 27, Number 3, September 2026 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

ISSN 1234-7655

Head of Editorial Office/Secretary
Patryk Barszcz, *Statistics Poland, Warsaw, Poland, e-mail: p.barszcz@stat.gov.pl, phone number +48 22 – 608 33 66*

Managing Editor
Adriana Nowakowska, *Statistics Poland, Warsaw, Poland, e-mail: a.nowakowska3@stat.gov.pl*

Technical Assistant
Rajmund Litkowiec, *Statistical Office in Rzeszów, Rzeszów, Poland, e-mail: r.litkowiec@stat.gov.pl*

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

CONTENTS

STATISTICS IN TRANSITION new series, *September 2026*
Vol. 27, No. 3, pp. III

Submission information for Authors

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

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

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

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

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

Manuscript should be submitted electronically to the Editor:

sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

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

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

STATISTICS IN TRANSITION new series, *September 2026*
Vol. 27, No. 3, pp. V–VI

Policy statement

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

***

Abstracting and indexing databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

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

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Emerging Sources Citation Index (ESCI) – Web of Science Core Collection

Electronic Journals Library

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

SCOPUS –Elsevier

TDNet

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

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. VII–XI

From the Editor

On behalf of the entire editorial team, I have the pleasure of presenting to our readers with the September issue consisting of 11 articles. Altogether 25 authors from a large set of countries – USA, Poland, Thailand, India, Zimbabwe, Australia, Botswana, Iraq, and Nigeria – present results of their investigations in a wide spectrum of research areas.

Original research papers

The September issue starts with the paper by **Zirou Zhou** and **Emily Berg** entitled ***Small-area estimation for the Iowa Seat-Belt Use Survey, 2017–2021.*** This study developed an integrated Bayesian small area estimation framework for estimating county-by-year seat-belt use in Iowa from 2017 to 2021. The target of inference was all 99 Iowa counties, although the survey directly sampled road segments from only 15 counties. To address this sparse and complex data structure, the authors considered two related estimation procedures: one for the proportion of belted vehicle miles traveled and one for the proportion of belted vehicle occupants. The framework combines models for one-inflated seat-belt-use proportions with a Poisson log-linear model for occupant counts, incorporates correlated random effects, and uses Bayesian model averaging to account for model uncertainty among retained candidate models. The main empirical finding is that the Bayesian estimates are generally close to the direct estimates in posterior mean, while often producing smaller posterior standard deviations.

Wojciech Łukaszonek, **Marcin Szymkowiak**, and **Waldemar Wołyński** in their article ***Classification of local labor markets in Wielkopolskie province in light of principal component analysis for doubly multivariate data*** evaluate the situation in the local labor markets of the Wielkopolskie province at the county level using principal component analysis for doubly multivariate data, in three variants: classical principal component analysis (PCA), a kernel model (KPCA) and a functional data model (FPCA). The authors applied two different augmented PCA approaches (KPCA and FPCA) to the considered set of variables, and for comparative purposes they also performed a classical PCA. The results obtained allowed to identify the FPCA model (Functional data Principal Component Analysis) as the one with the highest variance explanation. Moreover, the results obtained from the FPCA model show a high degree of alignment with the PCA model results. The methodology proposed in this article is

flexible and can be successfully used in other research areas including, inter alia, poverty, socio-economic development or disability, particularly where both feature variability and temporal dynamics play important roles in understanding complex socioeconomic phenomena.

The next paper prepared by **Wararit Panichkitkosolkul** and **Patchanok Srisuradetchai** entitled ***Confidence intervals for the Double XShanker distribution parameter: a simulation and application study*** examines four two-sided confidence intervals for the parameter θ of the Double XShanker distribution: likelihood-ratio, Wald, bootstrap-t, and bias-corrected and accelerated (BCa). An explicit expression for the observed Fisher information is derived, allowing the Wald interval to be computed directly from the maximum likelihood estimate. Interval performance is evaluated by empirical coverage probability (ECP) and average width (AW) through Monte Carlo experiments over a range of sample sizes and parameter values, and through an application to rainfall data. The likelihood-ratio and Wald intervals remain close to the nominal 0.95 level across most settings considered, including moderate sample sizes. The bootstrap-t and BCa intervals tend to be narrower but exhibit under-coverage when the sample size is small; their coverage improves as n increases. For larger parameter values, coverage deteriorates more noticeably in small samples, particularly for the bootstrap-based intervals. The rainfall example shows patterns consistent with the simulation results.

Namrata Pareek's and **Amit Choudhury's** article ***Inferential aspects of a M/M/1 queuing system using bivariate prior*** presents a Bayesian framework for estimating the performance measures of a single-server Markovian queue. A bivariate prior is employed to naturally incorporate the system's ergodicity condition along with the traditional likelihood function, which fundamentally differentiates this work from that of the existing literature. The Bayesian estimators of the performance measures are presented in the study, and the authors investigate standard errors and credible intervals of the Bayesian estimates. In addition, the predictive distributions for waiting times and the number of customers in the system at departure epoch were derived, offering deeper insights into the system dynamics. A simulation study was also conducted to assess the efficacy and reliability of the proposed methodology. Finally, a real-world scenario was presented to demonstrate the practical applicability of the developed algorithms.

In the next paper, ***The bias of the estimation of income distribution parameters in non-probability surveys: an experiment based on EU-SILC microdata for Poland,*** **Arkadiusz Kozłowski** examines the potential bias of estimates of income distribution parameters using real-life data from a household survey (European Union Statistics on Income and Living Conditions, EU-SILC) and computer simulations. The article investigates the change in the performance of the estimates depending on data defectiveness,

the sample size, the pool of auxiliary variables, sources of auxiliary information, and the estimation method. The non-probability samples had non-ignorable bias intentionally introduced by designing selection mechanisms directly tied to income or the access to a computer and the Internet. The results indicated that the bias of the estimates behaved predictably for simple parameters, such as the median and mean. However, for more complex parameters, the bias patterns were less stable. The use of sociodemographic variables for weighting did not consistently reduce the bias; on the contrary, in certain cases, it led to the aggravation of the bias.

The article by **Fastel Chipepa, Wilbert Nkomo, Nyika Mtemeri,** and **Takesure Nyakuamba** entitled ***Kumaraswamy family of generalized Type II Exponentiated Half Logistic-G distribution: properties and statistical inference*** introduces a new family of distributions (FDs) named the Kumaraswamy type II exponentiated half logistic-G (K-TIIEHL-G), and explores their mathematical and statistical properties, along with specific cases. Notably, this FDs can be represented as an infinite linear combination of exponentiated-G densities enabling the derivation of important statistical properties. The density and hazard rate geometries were investigated for some special cases. The model parameters were determined using the maximum likelihood technique and Monte Carlo simulation experiments were carried out to validate the consistency of the model parameters. The Kumaraswamy type II exponentiated half logistic-Weibull (K-TIIEHL-W), a member of the K-TIIEHL-G family, was applied to two sets of real data. The results indicated that the K-TIIEHL-W model significantly outperformed several established contending equi-parameter models in terms of data fitting.

Ali Salman Habeeb in the paper ***Modeling exchange rate volatility using EM algorithm with hybrid GARCH models*** constructs a three-step hybrid probability model that includes model identification, model estimation, and diagnosis checking using the Box-Jenkins approach. The author explored the performance of two important tools for estimating unknown parameters of the proposed model. The first is Maximum Likelihood (ML), which has a wide reputation in the statistical literature, and is provided in Eviews10 software. Furthermore, it was compared with the second proposed numerical method, namely Expectation-Maximization (EM algorithm), which requires creating a special MATLAB code to implement its performance. Moreover, two methods using information criteria (AIC, BIC, HQ) were compared. In order to show the effectiveness of the hybrid ARCH-GARCH models, the two methods were used to measure the volatility of the weekly exchange rate of the Central Bank of Iraq's sales of the US dollar against the Iraqi dinar from the first week of 2021 to the last week of 2023. It was concluded that the appropriate model is the hybrid MA(1)–GARCH(1,1) model, which is represented by the Moving Average MA(1) of the average equation and the GARCH(1,1) model of the conditional variance equation.

The work presented by **Dariusz Kotlewski**, ***Assessing the contribution of infrastructure to economic growth for total Polish economy and by Polish voivodship in light of KLEMS growth accounting,*** demonstrates the possibility of assessing the role of infrastructure in economic growth with the use of KLEMS growth accounting (otherwise called the KLEMS productivity accounting). A theoretical justification is provided for the validity of the adopted procedure. Then, the methodology and issues related to the extraction of infrastructure data are presented. Finally, a demonstration is provided with some interpretation based on Statistics Poland's data used and processed in the KLEMS productivity accounting for the Polish economy. It was found that when the proposed methodology is applied to spatial aggregations (i.e. countries and provinces) rather than industries then quite sensible data can be obtained even before the envisaged SNA reform.

The next article by **Ajobo Saheed A.**, **Adesina Oluwaseun A.**, **Ohida David I.**, **Alobaloke Kafayat A.**, and **Yaya OlaOluwa. S.**, ***Seasonal ARIMA modeling of inflation in Nigeria*****,** applies the Seasonal Autoregressive Integrated Moving Average (SARIMA) process to model the consumer price indices in Nigeria, specifically headline, food and core inflation. The study analyses monthly inflation rate data for the period January 2003 to December 2023, obtained from the Central Bank of Nigeria's website. The findings in the paper show the best-fitting SARIMA models for each inflation measure. The SARIMA (1,1,0)(2,0,1)12 model is selected as the most appropriate for headline inflation, while the SARIMA(2,1,2)(2,0,2)12 and SARIMA(1,0,1)(2,0,1)12 models are chosen for food inflation and core inflation, respectively. Caution is then required when applying SARIMA models in forecasting inflation, with emphasis on the importance of considering other relevant economic and socio-political factors that may influence inflation trajectories.

Farhan Ansari and **M. J. S. Khan** present ***Moments and estimation of Burr-Hatke Exponential model based on generalized order statistics with real data applications.*** This article discusses a recurrence relation for single and product moments of the Burr-Hatke exponential distribution based on generalized order statistics and record values. Moreover, the authors derived the explicit and exact expressions for the single moment of generalized order statistics from the Burr-Hatke exponential distribution. Furthermore, the classical and Bayesian estimation procedures for estimating the Burr-Hatke exponential distribution parameter are discussed when the data are in the form of records and *gos*. In addition, the asymptotic confidence interval, boot-p confidence interval, and credible interval for the parameter of the Burr-Hatke exponential distribution are discussed. Monte Carlo simulations were conducted to evaluate the performance of the estimators derived from both classical and Bayesian approaches, with a particular

emphasis on their mean squared error. Moreover, two real datasets are included for illustrative purposes in this study.

Research Communicates and Letters

The paper ***Rethinking equivalence scales in the measurement of extreme poverty: evidence from Poland*** prepared by **Hanna Dudek** provides new insights into the measurement of extreme poverty in Poland, focusing on the role of equivalence scales. The key research question is whether the original OECD equivalence scale, currently used by Statistics Poland, is consistent with the scale effects implied by the construction of the subsistence minimum across different household types. Using a one-parameter power equivalence scale, extreme poverty rates are recalculated under alternative assumptions based on anonymized microdata from the 2023 Household Budget Survey. The results indicate that the original OECD scale shows stronger economies than the one embedded in the subsistence minimum values, leading to substantial differences in the estimated extreme poverty rate. These findings underline the importance of a careful selection of equivalence scales in official poverty statistics, as methodological choices directly influence poverty assessment and the targeting of social policy.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 1–23, https://doi.org/10.59139/stattrans-2026-025
Received – 06.06.2025; accepted – 10.07.2026

Small-area estimation for the Iowa Seat-Belt Use Survey, 2017–2021

Zirou Zhou[1], **Emily Berg**[2]

Abstract

This study develops Bayesian small-area estimates of seat-belt use for Iowa counties from 2017 to 2021. The aim is to produce stable county-by-year estimates and measures of uncertainty for all 99 Iowa counties, although only 15 counties were directly sampled in the Iowa Seat-Belt Use Survey. The analysis focuses on two policy-relevant quantities: the percentage of belted vehicle miles traveled and the percentage of belted vehicle occupants. To address the multivariate, longitudinal, and boundary-inflated structure of the data, we compare several Bayesian hierarchical models for driver and passenger seat-belt-use proportions and combine them with a Poisson log-linear model for occupant counts. Correlated random effects are used to borrow strength across counties, years, and road segments, and posterior inference is implemented using Integrated Nested Laplace Approximations. Model performance is evaluated using WAIC, posterior predictive checks, posterior intervals, Bayesian model averaging, and comparisons with direct estimates. The empirical results show that the Bayesian estimates are generally close to direct estimates in posterior mean, while often producing substantially smaller posterior standard deviations. These gains indicate improved stability for local traffic-safety summaries, although they are interpreted as a bias–variance trade-off rather than as uniform dominance over direct estimation.

Key words: small area estimation, Bayesian hierarchical model, seat-belt survey, one-inflated proportion, INLA.

1. Introduction

Reliable local estimates of seat-belt use are important for traffic-safety planning, but they are difficult to obtain when survey samples are small within local domains. The Iowa Seat-Belt Use Survey provides annual information on seat-belt use, but it was designed primarily to support statewide estimation rather than direct county-by-year inference. The research problem addressed in this paper is therefore to estimate county-by-year seat-belt-use parameters for all 99 Iowa counties from 2017 to 2021 when direct county estimates are unstable or unavailable. Only 15 counties are directly sampled, and each sampled county contains a limited number of road segments observed across years. As a result, direct estimators can have large sampling variation, and direct estimates cannot be obtained for non-sampled counties. In this paper, the inferential target is all county-by-year domains in Iowa. Estimates for counties without sampled road segments are interpreted as model-based

[1]Department of Statistics, Iowa State University, 2438 Osborn Dr, Ames, IA, USA. E-mail: zzhou@iastate.edu. ORCID: https://orcid.org/0000-0002-4059-7871.
[2]Department of Statistics, Iowa State University, 2438 Osborn Dr, Ames, IA, USA. E-mail: emilyb@iastate.edu. ORCID: https://orcid.org/0000-0002-0962-005X.

posterior predictions that rely on auxiliary variables and hierarchical borrowing of strength, rather than as design-based direct estimates.

The practical motivation comes from the need to provide the Iowa Governor's Traffic Safety Bureau with more stable summaries of local seat-belt use and possible changes over time. Two related but distinct quantities are of interest. The first is the proportion of vehicle miles traveled for which drivers or passengers are belted, which weights road segments by traffic exposure. The second is the proportion of belted drivers or passengers among vehicle occupants, which depends on both the seat-belt-use proportions and the number of occupants observed on each road segment. These two summaries provide complementary information for program evaluation and traffic-safety planning. The data structure creates several modeling challenges: the responses are multivariate because drivers and passengers are observed jointly; the proportions are bounded and often equal to one; the occupant counts are discrete; and the same road segments are revisited from 2017 to 2021, creating a longitudinal structure.

Small area estimation provides a natural framework for this problem because it combines survey data with model-based information to improve the stability of domain-level estimates (Rao and Molina, 2015). Bayesian hierarchical methods are especially useful when the target parameters are nonlinear functions of unit-level quantities and when uncertainty must be propagated through several stages of prediction (Molina et al., 2014). Recent work has further emphasized Bayesian and unit-level modeling for complex survey data, including informative sampling, non-Gaussian responses, and repeated or longitudinal survey structures (Parker et al., 2023a,b; Vedensky et al., 2025a; McGovern et al., 2026). Previous work has extended small area methods to binary, count, and other non-Gaussian outcomes, including logistic mixed models (González-Manteiga et al., 2007), zero-inflated models with correlated area effects (Lyu et al., 2020), and broader unit-level Bayesian approaches for complex survey data (Parker et al., 2023a; Wu and Stephenson, 2023). Related Bayesian models have also been developed for binomial or beta-binomial structures with spatial or multilevel dependence (Dong and Wakefield, 2021), for applications that jointly model proportions and sample sizes (Oleson et al., 2007), and for longitudinal or multitype survey responses with domain-level dependence (Vedensky et al., 2025a,b). Bayesian model averaging and related Bayesian model-combination strategies have also been used to account for model uncertainty in small area estimation and spatial disease mapping (Aitkin et al., 2009; Hoeting et al., 1999; Mohebbi et al., 2014; Sugasawa et al., 2024). These studies show the value of flexible hierarchical modeling for complex small area problems, but they do not directly address the combination of one-inflated seat-belt-use proportions, occupant counts, repeated road-segment observations, and driver–passenger dependence in a multiyear county-level application.

Several studies are directly related to the Iowa seat-belt survey. Berg (2022) developed databases for small area estimation using the Iowa seat-belt data. Berg (2023) proposed a Bayesian unit-level Poisson model with multivariate random effects for seat-belt counts. Zhou and Berg (2024) developed a unit-level one-inflated beta model for seat-belt-use rates, but the analysis was limited to driver data from one year. These studies provide important foundations, but each addresses only part of the current problem. A count-based model captures traffic volume and occupant counts, but it does not directly model the bounded

seat-belt-use rate. A one-inflated beta model captures the boundary behavior of proportions, but the previous application did not jointly use drivers and passengers across multiple years. The remaining gap is an integrated application-specific Bayesian small area framework that compares alternative proportion models, incorporates occupant counts, and produces county-by-year predictions for both sampled and nonsampled counties.

This paper develops such a framework for the Iowa Seat-Belt Use Survey from 2017 to 2021. We model driver and passenger seat-belt-use proportions using several candidate distributions, including one-inflated beta, beta-epsilon, logit-normal-epsilon, beta-binomial, and binomial specifications. We also model the numbers of drivers and passengers using a Poisson log-linear count model. The proportion and count components are combined to estimate the two policy-relevant small area parameters: the percentage of belted vehicle miles traveled and the percentage of belted occupants. Correlated random effects are used to represent county, county-by-year, and road-segment sources of variation, allowing the model to borrow strength across related domains while preserving the longitudinal and multivariate structure of the survey. The Bayesian implementation is based on latent Gaussian modeling and Integrated Nested Laplace Approximations, which provide computationally feasible posterior inference for these hierarchical models (Rue et al., 2009).

The main contribution of this study is not the introduction of a new general distribution, but the construction and evaluation of an integrated Bayesian small area estimation strategy for a complex traffic-safety survey. Compared with simpler existing models, the proposed framework jointly addresses boundary-inflated proportions, discrete occupant counts, driver–passenger dependence, repeated observations over years, and prediction for nonsampled counties. Model assessment is organized into model selection and goodness-of-fit evaluation. We use WAIC to compare alternative year-effect structures, posterior predictive checks, and empirical coverage rates of posterior intervals to assess distributional fit, and comparisons with direct estimates to evaluate the practical effect of shrinkage. Bayesian model averaging is used to account for remaining model uncertainty among retained models.

The remainder of the paper is organized as follows. Section 2 describes the survey data structure, defines the county-by-year parameters of interest, and presents the Bayesian models for proportions and counts. Section 3 describes model selection, posterior predictive checking, Bayesian model averaging, and the prediction strategy for sampled and nonsampled counties. Section 4 presents the county-by-year estimates and compares the Bayesian predictions with direct estimates. Section 5 discusses the practical findings, limitations, and future extensions, including prior sensitivity analysis, spatial random effects, stronger predictive validation, and alternative count models when overdispersion is present.

2. Methods

This section presents the proposed Bayesian small area estimation framework in a structured order. We first define the county-by-year parameters of interest and present all candidate models considered in the study. We then describe the prior specification, posterior computation, and the criteria used for model selection and goodness-of-fit evaluation. Finally, we describe the Iowa Seat-Belt Use Survey data, the target population and sampling

frame, the auxiliary variables, the exploratory analysis, and the final prediction procedure for sampled and non-sampled counties.

2.1. Methodological framework and parameters of interest

The inferential target is seat-belt use for all 99 Iowa counties from 2017 to 2021. Let $i = 1,\ldots,99$ index counties, $t = 1,\ldots,5$ index survey years, and $j \in U_i$ index road segments in the population frame for county i. The observed survey data are available only for road segments from 15 sampled counties. Therefore, estimates for sampled counties use both observed survey data and model-based borrowing of strength, whereas estimates for non-sampled counties are model-based posterior predictions.

For road segment j in county i and year t, let $N_{d,tij}$ and $N_{p,tij}$ denote the observed numbers of drivers and passengers, respectively. Let $n_{d,tij}$ and $n_{p,tij}$ denote the corresponding numbers of belted drivers and belted passengers. When the denominators are positive, the observed driver and passenger seat-belt-use proportions are

$$y_{d,tij} = \frac{n_{d,tij}}{N_{d,tij}}, \tag{1}$$

$$y_{p,tij} = \frac{n_{p,tij}}{N_{p,tij}}. \tag{2}$$

Passenger proportions are undefined when no passengers are observed. These cases are excluded from the passenger proportion likelihood, while the passenger count information is retained in the count model.

Two types of county-by-year parameters are considered. The first is the vehicle-miles-traveled-weighted seat-belt-use proportion. Let v_{ij} denote vehicle miles traveled for road segment j in county i. The driver and passenger parameters are

$$\nu_{d,it} = \frac{\sum_{j\in U_i} v_{ij} y_{d,tij}}{\sum_{j\in U_i} v_{ij}}, \tag{3}$$

$$\nu_{p,it} = \frac{\sum_{j\in U_i} v_{ij} y_{p,tij}}{\sum_{j\in U_i} v_{ij}}. \tag{4}$$

The second type is the proportion of belted occupants. The driver, passenger, and total-occupant parameters are

$$\vartheta_{d,it} = \frac{\sum_{j\in U_i} N_{d,tij} y_{d,tij}}{\sum_{j\in U_i} N_{d,tij}}, \tag{5}$$

$$\vartheta_{p,it} = \frac{\sum_{j\in U_i} N_{p,tij} y_{p,tij}}{\sum_{j\in U_i} N_{p,tij}}. \tag{6}$$

These parameters are nonlinear functions of both seat-belt-use proportions and occupant counts. Therefore, the full framework combines models for proportions with models for counts and propagates uncertainty through both components.

2.2. Candidate models for seat-belt-use proportions

Let $g \in d, p$ denote driver or passenger. Five candidate models are considered for the seat-belt-use response. These models represent different ways of handling bounded proportions, one-inflation, and binomial count information.

For model m, the common linear predictor for the proportion or probability component is written as

$$\eta_{g,tij}^{(m)} = x_{g,tij}'\beta_g^{(m)} + b_{g,i}^{(m)} + b_{g,it}^{(m)} + b_{g,ij}^{(m)}, \tag{7}$$

where $x_{g,tij}$ contains the fixed-effect covariates, including log vehicle miles traveled, road type, and year-related terms when considered. The random effects $b_{g,i}^{(m)}$, $b_{g,it}^{(m)}$, and $b_{g,ij}^{(m)}$ represent county, county-by-year, and road-segment components, respectively. For each random-effect level, driver and passenger effects are allowed to be correlated:

$$b_\ell^{(m)} \sim \text{MVN}\left(0, \Sigma_\ell^{(m)}\right), \qquad \ell \in i, it, ij. \tag{8}$$

2.2.1 Model 1: one-inflated beta model

The one-inflated beta model separates the probability of observing a boundary value equal to one from the beta-distributed continuous component. Define

$$w_{g,tij} = \begin{cases} 1, & y_{g,tij} = 1, \\ 0, & y_{g,tij} < 1. \end{cases} \tag{9}$$

For observations with $0 < y_{g,tij} < 1$, the beta component is

$$y_{g,tij} \mid w_{g,tij} = 0 \sim \text{Beta}\left(\mu_{g,tij}^{(1)}\phi_g^{(1)}, 1 - \mu_{g,tij}^{(1)}\phi_g^{(1)}\right), \tag{10}$$

and the one-inflation component is

$$w_{g,tij} \sim \text{Bernoulli}\left(p_{g,tij}^{(1)}\right). \tag{11}$$

The two link functions are

$$\text{logit}\left(\mu_{g,tij}^{(1)}\right) = x_{g,tij}'\beta_g^{(1)} + b_{g,i}^{(1)} + b_{g,it}^{(1)} + b_{g,ij}^{(1)}, \tag{12}$$

$$\text{logit}\left(p_{g,tij}^{(1)}\right) = z_{g,tij}'\alpha_g^{(1)} + u_{g,i}^{(1)} + u_{g,it}^{(1)} + u_{g,ij}^{(1)}. \tag{13}$$

This model extends the unit-level one-inflated beta model in Zhou and Berg (2024) by applying it to both drivers and passengers across multiple years.

2.2.2 Model 2: beta-epsilon model

The beta-epsilon model uses a small adjustment to move boundary values away from one. Let

$$y^{\varepsilon}_{g,tij} = y_{g,tij} - \varepsilon, \qquad \varepsilon = 0.001. \tag{14}$$

Then

$$y^{\varepsilon}_{g,tij} \sim \text{Beta}\left(\mu^{(2)}_{g,tij}\phi^{(2)}_{g}, 1-\mu^{(2)}_{g,tij}\phi^{(2)}_{g}\right), \tag{15}$$

$$\text{logit}\left(\mu^{(2)}_{g,tij}\right) = x'_{g,tij}\beta^{(2)}_{g} + b^{(2)}_{g,i} + b^{(2)}_{g,it} + b^{(2)}_{g,ij}. \tag{16}$$

This model handles proportions equal to one through a deterministic boundary adjustment rather than a separate one-inflation component.

2.2.3 Model 3: logit-normal-epsilon model

The logit-normal-epsilon model also uses the adjusted response $y^{\varepsilon}_{g,tij}$, but models the transformed response on the logit scale:

$$\text{logit}\left(y^{\varepsilon}_{g,tij}\right) \sim \text{Normal}\left(\mu^{(3)}_{g,tij}, \sigma^{2(3)}_{g}\right), \tag{17}$$

$$\mu^{(3)}_{g,tij} = x'_{g,tij}\beta^{(3)}_{g} + b^{(3)}_{g,i} + b^{(3)}_{g,it} + b^{(3)}_{g,ij}. \tag{18}$$

This model is included as a normal working model for the transformed proportions.

2.2.4 Model 4: beta-binomial model

The beta-binomial model uses the observed belted counts and denominators directly:

$$n_{g,tij} \mid p^{(4)}_{g,tij} \sim \text{Binomial}\left(N_{g,tij}, p^{(4)}_{g,tij}\right), \tag{19}$$

$$p^{(4)}_{g,tij} \sim \text{Beta}\left(\mu^{(4)}_{g,tij}\frac{1-\rho^{(4)}_{g}}{\rho^{(4)}_{g}}, 1-\mu^{(4)}_{g,tij}\frac{1-\rho^{(4)}_{g}}{\rho^{(4)}_{g}}\right), \tag{20}$$

$$\text{logit}\left(\mu^{(4)}_{g,tij}\right) = x'_{g,tij}\beta^{(4)}_{g} + b^{(4)}_{g,i} + b^{(4)}_{g,it} + b^{(4)}_{g,ij}. \tag{21}$$

The parameter $\rho^{(4)}_{g}$ controls overdispersion relative to the binomial model. This specification is related to beta-binomial small area models discussed in Rao and Molina (2015) and applications such as Dong and Wakefield (2021).

2.2.5 Model 5: binomial model

The binomial model also uses the observed belted counts and denominators, but without the beta-binomial overdispersion layer:

$$n_{g,tij} \sim \text{Binomial}\left(N_{g,tij}, p^{(5)}_{g,tij}\right), \tag{22}$$

$$\text{logit}\left(p^{(5)}_{g,tij}\right) = x'_{g,tij}\beta^{(5)}_{g} + b^{(5)}_{g,i} + b^{(5)}_{g,it} + b^{(5)}_{g,ij}. \tag{23}$$

This model is the most direct count-based specification for the observed number of belted drivers or passengers.

2.3. Model for occupant counts

To estimate the proportion of belted occupants, we also model the numbers of drivers and passengers on each road segment. The count model is specified as a Poisson log-linear model:

$$N_{g,tij} \sim \text{Poisson}\left(\theta_{g,tij}\right), \qquad g \in d, p, \tag{24}$$

$$\log\left(\theta_{g,tij}\right) = x'_{g,tij}\beta_{C,g} + c_{g,i} + c_{g,it} + c_{g,ij}. \tag{25}$$

The log link is used because it is the standard link for Poisson regression and ensures that the mean parameter $\theta_{g,tij}$ is positive. As in the proportion models, the driver and passenger random effects are allowed to be correlated at the county, county-by-year, and road-segment levels:

$$c_{\ell} \sim \text{MVN}\left(0, \Sigma_{C,\ell}\right), \qquad \ell \in i, it, ij. \tag{26}$$

This count component is combined with the proportion component when estimating $\vartheta_{d,it}$, $\vartheta_{p,it}$, and $\vartheta_{o,it}$.

2.4. Model selection and goodness-of-fit evaluation

The evaluation procedure is organized into two parts: model selection and goodness-of-fit evaluation.

Model selection is based on the Widely Applicable Information Criterion (WAIC). WAIC is used to compare alternative year-effect structures, including fixed year effects, random year effects, no year effects, and random slopes for year-by-road-type combinations. The final year-effect structure is selected by considering both WAIC and model complexity.

Goodness-of-fit evaluation is based on posterior predictive p-values and empirical coverage rates of posterior intervals. For each candidate model, replicated data are generated from the posterior predictive distribution and compared with the observed data. We consider moment-based summaries, including the mean, median, standard deviation, skewness, and kurtosis, as well as Pearson-type discrepancy measures. Posterior predictive p-values close to 0 or 1 indicate a lack of fit. Because several models show poor posterior predictive performance for higher-order moments, we also use empirical coverage rates of 95%

posterior intervals as an additional goodness-of-fit diagnostic. These coverage rates are used for model checking and model comparison, not as a separate uncertainty-measurement procedure.

For the one-inflated beta model, a posterior draw can be written as

$$y_{g,tij}^{(l)} = \left(1 - w_{g,tij}^{(l)}\right) y_{g,tij}^{*,(l)} + w_{g,tij}^{(l)}, \tag{27}$$

where $y_{g,tij}^{*,(l)}$ is drawn from the beta component and $w_{g,tij}^{(l)}$ is drawn from the Bernoulli component. The posterior predictive mean and variance are

$$E\left(y_{g,tij}^{(l)}\right) = \left(1 - p_{g,tij}^{(l)}\right) \mu_{g,tij}^{*,(l)} + p_{g,tij}^{(l)}, \tag{28}$$

$$\mathrm{Var}\left(y_{g,tij}^{(l)}\right) = \left(1 - p_{g,tij}^{(l)}\right) \frac{\mu_{g,tij}^{*,(l)}\left(1 - \mu_{g,tij}^{*,(l)}\right)}{\phi_g^{(l)} + 1} + \left(1 - p_{g,tij}^{(l)}\right) p_{g,tij}^{(l)} \left(1 - \mu_{g,tij}^{*,(l)}\right)^2. \tag{29}$$

The Pearson-type discrepancy is then computed using the posterior predictive mean and variance. For Models 2 and 3, posterior predictive checks are computed on the adjusted proportion scale or logit-adjusted scale. For Models 4 and 5, posterior predictive checks are computed using the belted counts $n_{g,tij}$ conditional on $N_{g,tij}$.

2.5. Bayesian model averaging

The posterior predictive checks and empirical coverage rates in Section 3.2 suggest that no single candidate model is uniformly best for both drivers and passengers. Therefore, Bayesian model averaging is used among the retained models. In the final implementation, Models 2 and 5 are retained because they provide more reasonable goodness-of-fit behavior than the other candidate models.

Let $\mathscr{M} = 2,5$ denote the retained model set. For posterior draw l, let $\ell_m^{(l)}$ be the conditional log-likelihood of the observed data under model m. The model weight is computed as

$$\omega_m^{(l)} = \frac{\exp\left(\ell_m^{(l)}\right)}{\sum_{h \in \mathscr{M}} \exp\left(\ell_h^{(l)}\right)}, \qquad m \in \mathscr{M}. \tag{30}$$

For the two retained models, this becomes

$$\omega_2^{(l)} = \frac{\exp\left(\ell_2^{(l)}\right)}{\exp\left(\ell_2^{(l)}\right) + \exp\left(\ell_5^{(l)}\right)}, \qquad \omega_5^{(l)} = 1 - \omega_2^{(l)}. \tag{31}$$

For numerical stability, the weights are computed on the log-likelihood scale and then normalized. The final posterior draws of the small area parameters are obtained by combining the model-specific posterior draws using these weights.

2.6. Final prediction procedure for sampled and non-sampled counties

The final estimates are produced by combining the retained proportion models through Bayesian model averaging and then combining the resulting proportion predictions with the Poisson count predictions. For each posterior draw, the model generates predictions of $y_{d,tij}$, $y_{p,tij}$, $N_{d,tij}$, and $N_{p,tij}$ for road segments in the population frame. These predicted quantities are substituted into the definitions of $\nu_{d,it}$, $\nu_{p,it}$, $\vartheta_{d,it}$, $\vartheta_{p,it}$, and $\vartheta_{o,it}$.

For sampled counties, the prediction uses posterior draws of the fixed effects and available county-level random effects. Segment effects for nonsampled road segments are predicted from their posterior predictive distributions. For nonsampled counties, county-specific effects are not directly observed, so the estimates rely more heavily on fixed effects and the posterior predictive distribution of the county effects. Therefore, nonsampled-county estimates are model-based posterior predictions rather than direct design-based estimates.

3. Prior specification and selection

All candidate models are implemented in a Bayesian hierarchical framework. The prior distributions are specified separately for fixed effects, dispersion parameters, overdispersion parameters, and random-effect covariance matrices. For each model $m = 1, \ldots, 5$ and each occupant group $g \in \{d, p\}$, the regression coefficients in the proportion component are assigned diffuse normal priors,

$$\beta_g^{(m)} \sim \mathrm{N}\left(0, 10^4 I\right). \tag{32}$$

For the one-inflation component in Model 1, the fixed-effect coefficients are assigned the same type of diffuse normal prior,

$$\alpha_g^{(1)} \sim \mathrm{N}\left(0, 10^4 I\right). \tag{33}$$

For the Poisson count model, the fixed-effect coefficients are also assigned diffuse normal priors,

$$\beta_{C,g} \sim \mathrm{N}\left(0, 10^4 I\right). \tag{34}$$

For the beta precision parameters in the one-inflated beta model and the beta-epsilon model, weakly informative priors are assigned on the log scale to ensure positivity:

$$\log\left(\phi_g^{(1)}\right) \sim \mathrm{N}(0, 10^2), \qquad \log\left(\phi_g^{(2)}\right) \sim \mathrm{N}(0, 10^2). \tag{35}$$

For the logit-normal-epsilon model, the residual precision is denoted by

$$\tau_g^{(3)} = \frac{1}{\sigma_g^{2(3)}}, \tag{36}$$

and the prior is specified as

$$\log\left(\tau_g^{(3)}\right) \sim \mathrm{N}(0, 10^2). \tag{37}$$

For the beta-binomial model, the overdispersion parameter satisfies $0 < \rho_g^{(4)} < 1$. Therefore, a weakly informative prior is assigned on the logit scale:

$$\mathrm{logit}\left(\rho_g^{(4)}\right) \sim \mathrm{N}(0, 10^2). \tag{38}$$

The random effects are specified at three levels: county, county-by-year, and road segment. Let

$$\ell \in \{i, it, ij\} \tag{39}$$

denote the random-effect level. For the proportion component, the driver and passenger random effects are modeled jointly as

$$b_\ell^{(m)} = \begin{pmatrix} b_{d,\ell}^{(m)} \\ b_{p,\ell}^{(m)} \end{pmatrix} \sim \mathrm{MVN}\left(0, \Sigma_{B,\ell}^{(m)}\right), \qquad m = 1, \ldots, 5. \tag{40}$$

For the one-inflation component in Model 1, the corresponding driver and passenger random effects are

$$\begin{pmatrix} b_\ell^{(1)} \\ u_\ell^{(1)} \end{pmatrix} = \begin{pmatrix} b_{d,\ell}^{(1)} \\ b_{p,\ell}^{(1)} \\ u_{d,\ell}^{(1)} \\ u_{p,\ell}^{(1)} \end{pmatrix} \sim \mathrm{MVN}\left(0, \Sigma_{U,\ell}^{(1)}\right). \tag{41}$$

For the Poisson count model, the driver and passenger count random effects are modeled as

$$c_\ell = \begin{pmatrix} c_{d,\ell} \\ c_{p,\ell} \end{pmatrix} \sim \mathrm{MVN}\left(0, \Sigma_{C,\ell}\right). \tag{42}$$

The covariance matrices are assigned inverse-Wishart-type priors. Equivalently, in the INLA implementation, the corresponding precision matrices are assigned Wishart priors. For each covariance matrix

$$\Sigma_{r,\ell}, \qquad r \in \{B, C\}, \tag{43}$$

we define

$$\Omega_{r,\ell} = \Sigma_{r,\ell}^{-1}, \tag{44}$$

and assign

$$\Omega_{r,\ell} \sim \text{Wishart}_2(\nu, R), \qquad \nu = 5, \qquad R = \begin{pmatrix} 0.01 & 0 \\ 0 & 0.01 \end{pmatrix}. \tag{45}$$

For $\Sigma_{U,\ell}$, we can define $\Omega_{U,\ell} = \Sigma_{U,\ell}^{-1}$, and assign

$$\Omega_{U,\ell} \sim \text{Wishart}_4(\nu, R), \qquad \nu = 11, \qquad R = \begin{pmatrix} 0.01 & 0 & 0 & 0 \\ 0 & 0.01 & 0 & 0 \\ 0 & 0 & 0.01 & 0 \\ 0 & 0 & 0 & 0.01 \end{pmatrix}. \tag{46}$$

This prior allows the two to four occupant-specific random effects to be correlated while keeping the covariance structure weakly informative. The use of this inverse-Wishart-type covariance prior is motivated by the prior sensitivity analysis in Section 3.1, where the inverse-Wishart specification better recovered the positive dependence between the driver and passenger random effects than the half-Cauchy alternatives.

Posterior inference is implemented using Integrated Nested Laplace Approximations (INLA) for latent Gaussian models (Rue et al., 2009). Posterior means are used as point estimates, and posterior standard deviations and 95% credible intervals are used to summarize uncertainty.

3.1. Prior sensitivity test

To evaluate the sensitivity of posterior inference to the prior specification for the county-level covariance structure, we compared three prior choices: a half-Cauchy prior with scale 2.5, a half-Cauchy prior with scale 1000, and an inverse-Wishart prior with hyperparameters $(5, 0.01, 0.01, 0)$. The simulation study was based on the one-inflated beta regression model defined in Section 2.2.1, specifically equations (9) to (13). The beta precision parameter was fixed at $\phi = 42$. The true values of the random-effect parameters were chosen to match values estimated from the empirical data analysis, with

$$\sigma_b = 0.09164, \qquad \sigma_m = 0.077, \qquad \rho = 0.6. \tag{47}$$

Table 1 reports the prior sensitivity results on the standard deviation scale. The two half-Cauchy priors give nearly identical posterior summaries. Both half-Cauchy specifications overestimate the marginal standard deviations of the county-level random effects. For example, the posterior medians of σ_b and σ_m are about 0.355 and 0.539, respectively, compared with the true values 0.09164 and 0.077. In addition, the posterior median of the correlation is close to zero, and the posterior interval is wide, indicating that the half-Cauchy priors do not recover the positive dependence between the two county-level random-effect components.

Table 1. Prior sensitivity analysis for county-level standard deviations and correlation

Prior		True	Posterior median	95% posterior interval
Half-Cauchy(2.5)	σ_b	0.09164	0.3550	(0.2548, 0.5054)
	σ_m	0.0770	0.5390	(0.2814, 1.1581)
	ρ	0.6000	0.0000	(-0.6278, 0.6308)
Half-Cauchy(1000)	σ_b	0.09164	0.3550	(0.2548, 0.5054)
	σ_m	0.0770	0.5390	(0.2814, 1.1581)
	ρ	0.6000	0.0000	(-0.6278, 0.6308)
Inverse-Wishart	σ_b	0.09164	0.1734	(0.0698, 0.3824)
	σ_m	0.0770	0.0890	(0.0000, 0.5379)
	ρ	0.6000	0.8616	(-0.5171, 0.9971)

The inverse-Wishart prior gives a different pattern. The posterior medians of the standard deviations are closer to the true values, especially for the second component, and the posterior median of the correlation is positive. This suggests that the inverse-Wishart prior better preserves the positive dependence structure in the simulation. However, the posterior interval for ρ remains wide, and the posterior interval for σ_m is also wide. This indicates that the county-level covariance structure is still weakly identified.

Overall, the sensitivity analysis shows that the beta precision parameter is relatively stable across prior specifications, but the county-level variance and correlation parameters are sensitive to the covariance prior. We use the inverse-Wishart prior in the subsequent analysis because it better captures the positive dependence between the driver and passenger county-level random effects, which is important for the proposed multivariate small area model.

4. Survey data exploratory analysis

The Iowa Seat-Belt Use Survey is conducted annually to support estimates of seat-belt use. The data used in this study cover 2017–2021. A probability-proportional-to-size sample of road segments was selected from 15 Iowa counties, and the selected road segments were revisited across years. Although only 15 counties were directly sampled, the inferential target is all 99 Iowa counties. Therefore, estimates for the remaining 84 counties are model-based posterior predictions that rely on auxiliary variables and the fitted hierarchical model.

For each sampled road segment and year, observers recorded the number of drivers, passengers, and total occupants, as well as the corresponding numbers wearing seat belts. The data therefore contain both proportions and counts. The proportions are bounded between 0 and 1 and often equal to 1, motivating the one-inflated and epsilon-adjusted proportion models. The count outcomes are nonnegative integers, motivating the Poisson count model.

The main auxiliary variables are vehicle miles traveled, road type, year, county, and road-segment identifiers. Vehicle miles traveled is used both as a covariate in the regression models and as a weight in the definition of the vehicle-miles-traveled parameter. Road type is included because seat-belt-use patterns and traffic volume may differ across road classes.

Year is considered because the survey is longitudinal and because changes in seat-belt use over time are of interest for traffic-safety planning.

4.1. Exploratory analysis of seat-belt-use proportions

Following Zhou and Berg (2024), Figure 1 examines the relationship between the logit of observed seat-belt-use proportions and log vehicle miles traveled. The plots suggest an approximately linear association between the logit of the proportion and log vehicle miles traveled. They also suggest that the relationship may vary by road type and year. These patterns motivate the inclusion of log vehicle miles traveled, road type, and year-related terms in the candidate models.

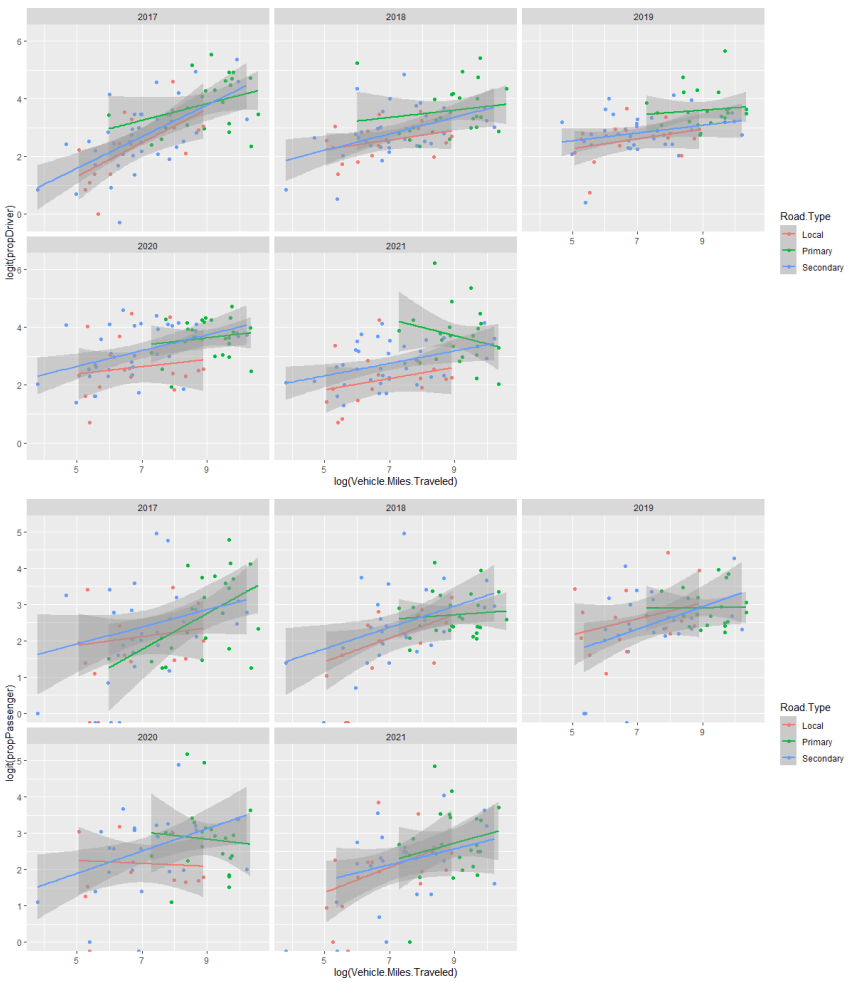

Figure 1. Scatter plots for the logit of proportions for drivers and passengers

4.2. Exploratory analysis of occupant counts

To assess whether a Poisson count model is reasonable, vehicle miles traveled were stratified into ten groups based on deciles. For each group, the sample mean and standard deviation of the count outcome were calculated. The counts were also grouped by road type. We then regressed the logarithm of the standard deviation on the logarithm of the group mean. Under a Poisson mean–variance relationship, this log–log regression has a slope close to 0.5.

Table 2. Box–Cox transformation regression: log standard deviation versus log mean

Grouping	**Outcome**	λ
Road type	Total driver count	0.4627
Road type	Total passenger count	0.5233
VMT category	Total driver count	0.4772
VMT category	Total passenger count	0.5957

As shown in Table 2, the estimated slopes are close to 0.5, suggesting that the observed mean–variance relationship is broadly consistent with a Poisson structure. Therefore, the Poisson log-linear model is used as the baseline count model.

5. Model evaluation results

5.1. Model selection for year effects

We first evaluate whether year effects should enter the model as fixed effects, random effects, both, or neither. This comparison is conducted using the binomial model, which is later shown to be one of the retained models.

Table 3. Posterior summaries for fixed year effects

	Year	Mean	SD	95% posterior interval
Drivers	2018	-0.133	0.154	(-0.438, 0.170)
	2019	-0.100	0.155	(-0.405, 0.204)
	2020	0.154	0.156	(-0.152, 0.461)
	2021	-0.186	0.155	(-0.491, 0.120)
Passengers	2018	-0.033	0.163	(-0.354, 0.286)
	2019	0.141	0.164	(-0.182, 0.464)
	2020	0.134	0.164	(-0.188, 0.457)
	2021	0.034	0.165	(-0.290, 0.358)

The posterior intervals in Table 3 contain zero for all fixed year effects, suggesting limited evidence of strong year-specific fixed effects. We then compare alternative year-effect structures using WAIC.

Table 4. WAIC for alternative year-effect structures.

Model	WAIC
Fixed year effects	3636.05
Random year effects	3632.49
No year effects	3634.34
Random slopes for years and road type	3630.16

Although the random-slope model has the smallest WAIC in Table 4, the differences among the candidate structures are small. Therefore, parsimony and interpretability are also considered. Figure 2 shows posterior density plots for the variance and correlation parameters associated with random year effects.

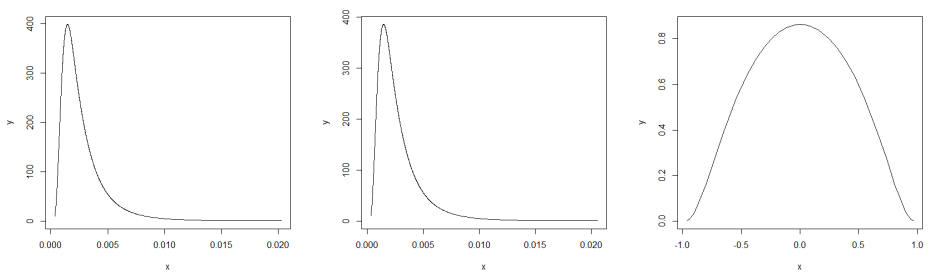

Figure 2. Posterior densities for σ^2 and ρ in the random year-effect model

The variance and correlation parameters in Figure 2 are nearly zero, indicating that the random year effects add little extra variation. Thus, the final model excludes the year effect for both fix and random effects as it may be an unnecessary random-effect component.

5.2. Model comparison

Posterior predictive p-values are used to compare the five candidate response models with random slopes for road types. The posterior p-values in Table 5 compare observed summaries with summaries generated from posterior predictive draws.

Table 5. Posterior predictive p-values for alternative models

Specification		Mean	Median	SD	Skewness	Kurtosis	Pearson test
Model 1	Driver	22.8%	22.6%	0.9%	99.9%	0.1%	45.5%
	Passenger	59.87%	12.47%	0.03%	100%	0%	0%
Model 2	Driver	57.8%	87.2%	3.07%	100%	0.67%	0.87%
	Passenger	50.5%	94.3%	8.2%	100%	0%	6.3%
Model 3	Driver	51.8%	99.7%	0%	0%	0%	0%
	Passenger	26.53%	52.63%	100%	0%	0%	100%
Model 4	Driver	78.2%	80.47%	62.47%	37.43%	36%	56.23%
	Passenger	99.87%	7.8%	56.3%	1.23%	2.7%	37.53%
Model 5	Driver	80.9%	60.03%	30.7%	2.7%	5.06%	35.43%
	Passenger	81.23%	11.4%	98.78%	100%	100%	99.17%

Table 5 suggests that no model is uniformly best across all summaries and both response groups. Model 1 works reasonably for some driver summaries but performs poorly for some passenger summaries. Models 2 and 5 provide more stable performance across both drivers and passengers.

Since the posterior p-values are not great for the higher order moments, we also evaluate the empirical coverage of 95% posterior intervals as a goodness-of-fit test. The results are shown in Table 6.

Table 6. Coverage rates of 95% posterior intervals

Specification		Coverage rate
Model 1	Driver	89.95%
	Passenger	73.94%
Model 2	Driver	95.45%
	Passenger	96.28%
Model 3	Driver	87.08%
	Passenger	53.10%
Model 4	Driver	88.51%
	Passenger	44.73%
Model 5	Driver	94.73%
	Passenger	97.27%

Models 2 and 5 have coverage rates closest to the nominal 95% level for both drivers and passengers. Therefore, these two models are retained for Bayesian model averaging.

Table 7. Summary of model checking and final model decision

Model	Response form	Checking result	Decision
M1	One-inflated beta	Reasonable for some driver summaries; weak for passenger summaries	Not retained
M2	Beta-epsilon	Reasonable PPCs; near-nominal coverage	Retained
M3	Logit-normal-epsilon	Poor PPCs for several summaries	Not retained
M4	Beta-binomial	Weak passenger coverage; unstable checks	Not retained
M5	Binomial	Reasonable PPCs; near-nominal coverage	Retained

Table 7 summarizes the model-checking evidence and the final model decision. The results indicate that no single candidate model is uniformly adequate for all summaries and both response groups. Model 1 captures some features of the driver proportions but performs weakly for passenger summaries. Model 3 has poor posterior predictive performance for several summaries, indicating that the logit-normal working model does not adequately reproduce the observed data. Model 4 allows extra-binomial variation, but its passenger coverage rate is weak, and its checking results are less stable. In contrast, Models 2 and 5 provide more balanced goodness-of-fit behavior across drivers and passengers, with reasonable posterior predictive checks and empirical coverage rates close to the nominal level. Therefore, these two models are retained and combined through Bayesian model averaging in the final estimation procedure.

5.3. Bayesian model averaging results

Figures 3 and 4 show two selected county examples comparing Models 2 and 5. Model 5 performs better than Model 2 in some counties, while Model 2 performs better in others. This supports the use of Bayesian model averaging rather than selecting a single retained model.

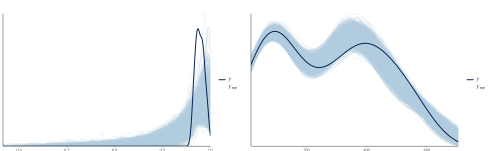

Figure 3. Selected county example where Model 5 performs better than Model 2

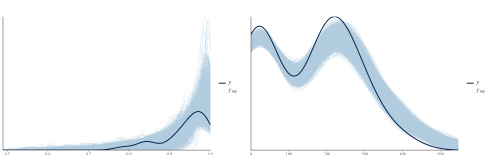

Figure 4. Selected county example where Model 2 performs better than Model 5

The posterior predictive checks indicate that no single candidate model is uniformly best for both drivers and passengers across all counties. Therefore, Bayesian model averaging is used among the retained models. In the final implementation, Models 2 and 5 are retained because they provide reasonable posterior predictive checks and credible interval behavior.

6. Results and Comparison

This section reports the final Bayesian small area estimates and compares them with direct estimates when available. The results are organized around the two parameters of interest: the proportion of belted vehicle miles traveled and the proportion of belted occupants. The supplement includes figures and tables for state-level estimations of the proportion of belted vehicle miles traveled, along with comparisons to direct estimations. Similarly, for the proportion of belted occupants, the supplement provides figures for county-level estimations, as well as state-level figures and tables, with comparisons to direct estimations. Additionally, a comparison between the two approaches is included in the supplement.

6.1. Proportion of vehicle miles traveled

For drivers and passengers, the vehicle-miles-traveled-weighted parameters $\nu_{d,it}$ and $\nu_{p,it}$ are estimated using posterior draws from the Bayesian model averaging procedure.

6.1.1 Posterior means and credible intervals

Figure 5 shows the posterior means and 95% credible intervals for the 15 sampled counties across years. All county results are included in the supplementary material. The county-level estimates show moderate year-to-year variation, but most credible intervals overlap substantially, suggesting limited evidence of strong county-specific temporal changes.

Figure 6 shows year-to-year differences based on VMT. The reference line at zero helps assess whether annual changes are meaningfully different from zero. Most intervals overlap zero, suggesting that annual fluctuations are generally within posterior uncertainty.

6.1.2 Comparison with direct estimates

Figure 7 in the supplement compares Bayesian estimates with direct estimates. The posterior means are generally close to the direct estimates, indicating that the model preserves the main empirical pattern in the observed data. The posterior standard deviations are often smaller than the direct-estimator standard deviations, reflecting the stabilizing effect of borrowing strength across counties, years, and road segments.

The reduction in posterior standard deviation should be interpreted as a bias–variance trade-off. The Bayesian model improves stability by shrinking noisy county-year estimates toward patterns explained by covariates and random effects. However, shrinkage can also introduce bias if important local variation is not captured by the model. Therefore, the comparison in Figure 7 should be interpreted as evidence of improved precision, not as proof that the Bayesian estimates are uniformly superior to direct estimates.

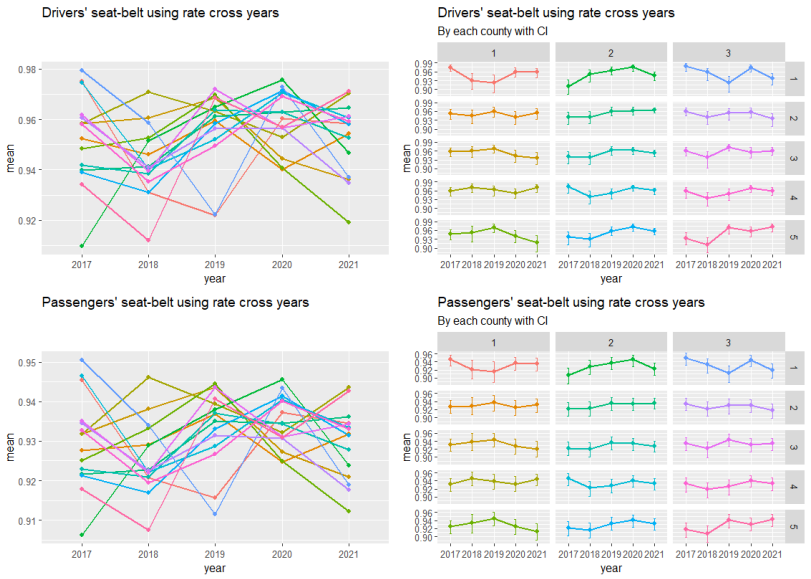

Figure 5. Posterior means and 95% credible intervals for 15 sampled counties across years

6.1.3 State-level estimation

The state-level posterior intervals for the vehicle-miles-traveled-weighted proportions are shown in Table 8 in the supplement. Driver estimates are generally higher and more stable than passenger estimates.

6.1.4 Estimation for all 99 counties

Heat maps for the lower bound, posterior mean, and upper bound of the 95% posterior intervals are provided in the Supplemental material. Moran's I tests are used to assess whether the county-level posterior means show spatial dependence. The results are shown in Table 9 in the supplement.

The Moran's I tests do not show strong evidence of spatial dependence for most years. Some borderline results may reflect spatial patterning in vehicle miles traveled or road characteristics, which are included as auxiliary information in the prediction model.

6.2. Proportion of belted occupants

The proportion of belted occupants is estimated by combining predictions from the proportion models with predictions from the Poisson count model. This procedure estimates $\vartheta_{d,it}$ and $\vartheta_{p,it}$.

Based on the counts, the changes in proportions over the years, the 95% posterior intervals, and comparisons against direct estimations, including potential outliers, are also included in the supplementary material.

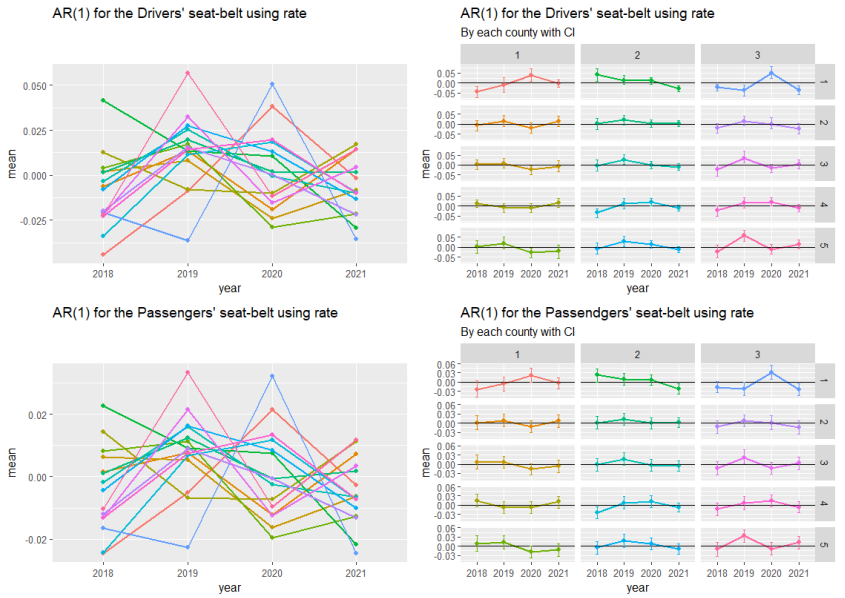

Figure 6. Posterior means and 95% credible intervals for year-to-year differences in 15 sampled counties

6.2.1 Estimation for all 99 counties

Moran's I tests for the belted-occupant county means are shown in Table 11 in the supplement.

The Moran's I results are similar to those from the vehicle-miles-traveled procedure. Overall, there is limited evidence of strong residual spatial dependence in the county-level posterior means, although spatial random effects remain a useful direction for future work.

7. Discussion and conclusion

This study developed an integrated Bayesian small area estimation framework for estimating county-by-year seat-belt use in Iowa from 2017 to 2021. The target of inference was all 99 Iowa counties, although the survey directly sampled road segments from only 15 counties. To address this sparse and complex data structure, we considered two related estimation procedures: one for the proportion of belted vehicle miles traveled and one for the proportion of belted vehicle occupants. The framework combines models for one-inflated seat-belt-use proportions with a Poisson log-linear model for occupant counts, incorporates correlated random effects, and uses Bayesian model averaging to account for model uncertainty among retained candidate models.

The main empirical finding is that the Bayesian estimates are generally close to the direct estimates in posterior mean, while often producing smaller posterior standard deviations. In many county-year domains, the hierarchical model reduces posterior uncertainty

by borrowing strength across counties, years, and road segments. This is practically useful because direct county estimates can be unstable when only a small number of road segments are observed. The proposed framework, therefore, provides a structured model-based approach for producing more stable local summaries of seat-belt use, including predictions for nonsampled counties where direct estimates are unavailable.

However, these gains should be interpreted carefully. A smaller posterior standard deviation does not automatically imply better estimation if the shrinkage introduces bias. Thus, the improvement from the Bayesian model should be viewed as a bias–variance trade-off rather than as uniform dominance over direct estimation. The model improves stability by shrinking noisy county-year estimates toward patterns explained by the covariates and random effects, but this shrinkage may also reduce local variation that is not fully captured by the model. For non-sampled counties, the estimates are model-based posterior predictions that depend on the assumed hierarchical structure and available auxiliary variables, rather than design-based direct estimates.

The model comparison results also show that no single response distribution performed best for all aspects of the data. The Iowa seat-belt survey contains several challenging features, including boundary values in the proportions, discrete occupant counts, repeated road-segment observations, and dependence between driver and passenger outcomes. Although the retained models provide reasonable summaries, they remain simplified representations of the underlying data-generating process. In particular, county-specific year trends are difficult to estimate precisely with only five annual waves and a limited number of sampled counties.

Another limitation concerns the passenger analysis. Observations with zero passengers were excluded from the passenger-use component. Therefore, passenger-related estimates should be interpreted as conditional on at least one passenger being present, rather than as unconditional estimates over all vehicles. If passenger presence is associated with road type, traffic volume, time of day, county characteristics, or seat-belt behavior itself, this conditioning may affect the passenger estimates and county-year comparisons.

Future work should extend the model in several directions. First, spatial random effects could be incorporated to borrow information across neighboring counties and to better reflect geographic dependence in traffic-safety behavior (Muchie et al., 2022). Second, stronger predictive validation, such as cross-validation, bootstrap comparisons, or design-based benchmarking, would help assess whether the reduction in posterior uncertainty is accompanied by acceptable bias. Third, prior sensitivity analysis requires improvement, particularly for beta-epsilon and Bernoulli regression. The current prior sensitivity test is simplistic and relies on a single simulation iteration for beta one-inflated regression. Finally, alternative count models, such as negative binomial, zero-inflated, or hurdle-type formulations, should be considered if over-dispersion or structural zeros are present in the occupant counts. These extensions would strengthen the reliability and interpretability of Bayesian small area estimates for future seat-belt surveys.

Supplementary Material

The supplementary material is available at https://doi.org/10.5281/zenodo.21275327.

References

Aitkin, M., Liu, C. C., and Chadwick, T., (2009). Bayesian model comparison and model averaging for small-area estimation. *The Annals of Applied Statistics*, 3(1), pp. 199–221. doi:10.1214/08-AOAS205.

Berg, E., (2022). Construction of databases for small area estimation. *Journal of Official Statistics*, 38, pp. 673–708.

Berg, E., (2023). Small area prediction of seat-belt use rates using a Bayesian hierarchical unit-level Poisson model with multivariate random effects. *Stat*, 12, e544.

Dong, T. Q., Wakefield, J., (2021). Modeling and presentation of vaccination coverage estimates using data from household surveys. *Vaccine*, 39, pp. 2584–2594.

González-Manteiga, W., M. J. Lombardía, I. Molina, D. Morales, andSantamaría, L., (2007). Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics & Data Analysis*, 51(5), pp.2720–2733.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T., (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4), pp. 382–417. doi:10.1214/ss/1009212519.

Lyu, X., E. J. Berg, and Hofmann, H., (2020). Empirical Bayes small area prediction under a zero-inflated lognormal model with correlated random area effects. *Biometrical Journal*, 62, pp.1859–1878.

McGovern, A., Wilson, K., and Wakefield, J., (2026). Direct-assisted Bayesian unit-level modeling for small area estimation of rare event prevalence. *Journal of Survey Statistics and Methodology*, 14(1), pp. 178–208. doi:10.1093/jssam/smaf037.

Mohebbi, Mohammadreza, Rory Wolfe, and Forbes A., (2014). Disease mapping and regression with count data in the presence of overdispersion and spatial autocorrelation: a Bayesian model averaging approach. *International journal of environmental research and public health*, 11, No. 1 , pp. 883–902.

Molina, I., B. Nandram and Rao, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, 8(2), pp. 852–885.

Muchie, K. F., A. K. Wanjoya and Mwalili, S. M., (2022). On hierarchical Bayesian spatial small area model for binary data under spatial misalignment. *Journal of Probability and Statistics*, 3865626.

Oleson, J. J., C. He, D. Sun and Sheriff, S., (2007). Bayesian estimation in small areas when the sampling design strata differ from the study domains. *Survey Methodology*, 33, pp. 173–185.

Parker, P. A., Janicki, R. and Holan, S. H., (2023a). A comprehensive overview of unit-level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology*, 11(4), pp. 829–857. doi:10.1093/jssam/smad020.

Parker, P. A., Janicki, R. and Holan, S. H., (2023b). Comparison of unit-level small area estimation modeling approaches for survey data under informative sampling. *Journal of Survey Statistics and Methodology*. doi:10.1093/jssam/smad022.

Rao, J. N. K., Molina, I. , (2015). *Small Area Estimation*. Hoboken, NJ: John Wiley & Sons.

Rue, H., S. Martino and Chopin, N., (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B* 71, pp.319–392.

Sugasawa, S., Kobayashi, G., and Kawakubo, Y., (2024). Bayesian benchmarking small area estimation via entropic tilting. arXiv:2407.17848.

Vedensky, D., Parker, P. A. and Holan, S. H., (2025a). Bayesian unit-level models for longitudinal survey data under informative sampling: An analysis of expected job loss using the Household Pulse Survey. *Journal of the Royal Statistical Society: Series A, Statistics in Society*.

Vedensky, D., Parker, P. A. and Holan, S. H., (2025b). Bayesian unit-level modeling of categorical survey data with a longitudinal design. arXiv:2502.15112.

Wu, S. M., Stephenson, B. J. K., (2023). Bayesian estimation methods for survey data with potential applications to health disparities research. *WIREs Computational Statistics*, 16(1), e1633. doi:10.1002/wics.1633.

Zhou, Z., Berg, E., (2024). A unit-level one-inflated beta model for small area prediction of seat-belt use rates. *Journal of Applied Statistics*, pp. 1–24.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 25–46, https://doi.org/10.59139/stattrans-2026-026
Received – 12.08.2025; accepted – 07.07.2026

Classification of local labor markets in the Wielkopolskie voivodship in the light of principal component analysis for doubly multivariate data

Wojciech Łukaszonek[1], **Marcin Szymkowiak**[2], **Waldemar Wołyński**[3]

Abstract

The main aim of this article is to evaluate the situation in the local labor markets of the Wielkopolskie voivodship at the county level using principal component analysis for doubly multivariate data, in three variants: classical principal component analysis (PCA), a kernel model (KPCA) and a functional data model (FPCA). This will be done using doubly multivariate data, i.e. data that are characterized by both the variability of features and the variability over time. Since the use of classical PCA for doubly multivariate data raises some problems related to the appropriate estimation of the covariance matrix, the authors propose the use of a non-linear kernel model and a model for functional data to overcome this inconvenience.

The analysis of local labor markets in the Wielkopolskie voivodship based on 12 observed variables between 2004 and 2022 showed that similar clustering results were achieved for all the considered approaches. However, a substantial increase in the explained variance was only achieved in relation to PCA in the case of FPCA. Equally important is that unlike classical PCA, which treats each time point as a separate variable, FPCA models the entire time trajectory as a functional object. This preserves the temporal continuity and allows the capture of important dynamics such as trends, seasonal patterns, and structural breaks that characterize labor markets.

The methodology proposed in this article is flexible and can be successfully used in other research areas including poverty, socio-economic development or disability, particularly where both feature variability and temporal dynamics play important roles in understanding complex socioeconomic phenomena.

Keywords: local labor markets, doubly multivariate data, principal component analysis, functional principal component analysis, kernel principal component analysis, cluster analysis

1. Introduction

Labor markets represent a pivotal component of contemporary economies, functioning as a nexus where the supply and demand of the workforce interact to determine employment

[1]Interfaculty Institute of Mathematics and Statistics, Calisia University & Statistical Office in Poznań, Poznań, Poland. E-mail: w.lukaszonek@uniwersytetkaliski.edu.pl. ORCID: https://orcid.org/0000-0001-8109-2901.

[2]Institute of Informatics and Quantitative Economics, Poznań University of Economics and Business & Statistical Office in Poznań, Poznań, Poland.
E-mail: marcin.szymkowiak@ue.poznan.pl. ORCID: https://orcid.org/0000-0003-3432-4364.

[3]Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland.
E-mail: waldemar.wolynski@amu.edu.pl. ORCID: https://orcid.org/0000-0002-0777-9163.

patterns, wage levels, and overall economic development (Bosworth et al., 1996). The analysis of labor market dynamics has become increasingly significant in contemporary economic research due to its profound implications for policymaking, regional development strategies, and social welfare programs (Martin and Morrison, 2003; Hartal et al., 2023). A comprehensive understanding of the spatial and temporal heterogeneity of labor markets is imperative for effective policy formulation to address issues such as unemployment, skills mismatches, and regional disparities that persist despite broader economic growth (Blanchard et al., 1992; Elhorst, 2003; Kanbur and Svejnar, 2009; McQuaid et al., 2013; Woźniak, 2021; Yenilmez and Esin, 2016).

In Poland, the transition from a centrally planned economy to a market-oriented system has resulted in substantial transformations in labor market structures, giving rise to distinct regional patterns that reflect both historical legacies and contemporary economic forces (Huber, 2007; Jenkins, 2001; Keane and Prasad, 2006). The country's administrative division into voivodeships (provinces) and counties (poviats) has resulted in local labor markets with varying characteristics, challenges, and development trajectories. The aforementioned disparities have been further influenced by factors such as EU accession, global economic crises, technological change, and demographic shifts including migration patterns (Ciżkowicz et al., 2016; Fihel and Okólski, 2019) .

The Wielkopolskie voivodship, with its diverse economic structure encompassing both industrial centers and agricultural regions, presents a particularly interesting case for examining local labor market variations. Within this voivodship, counties exhibit significant disparities in employment rates, sectoral composition, wage levels, and labor productivity. These disparities have consequences for the economic performance of individual territories and also contribute to broader patterns of socioeconomic divergence or convergence within the region (Smętkowski, 2018; Churski et al., 2020; Kwiatkowski and Szymańska, 2022, 2024).

The inherent complexity of labor market analysis is attributable to its multidimensional nature, wherein a multitude of agents (factors) exert a simultaneous influence on market conditions (Fischer and Nijkamp, 2014). Conventional analytical methodologies frequently encounter difficulties in adequately capturing the intricacies of such data, particularly when dealing with information that displays both cross-sectional (across counties) and longitudinal (over time) variability - a phenomenon referred to as "doubly multivariate data"[4]. The application of classical statistical methods is rendered significantly more challenging by such data structures, with issues including dimensionality, multicollinearity and temporal dependencies.

Among many different statistical methods classical principal component analysis (PCA) has been utilized as a dimension reduction technique in economics and regional studies for a considerable period (Jolliffe, 2002; Del Campo et al., 2008). However, classical PCA encounters substantial limitations when applied to doubly multivariate data, particularly in its

[4]The term "doubly multivariate data" is used here in its established sense from the literature on multivariate repeated measurements, i.e. data that are multivariate on two levels: several variables observed on each unit at several time points. It should be distinguished from spatiotemporal data, since the methods applied in this study do not model spatial dependence: each county is treated as an independent observation represented by a $p \times T$ matrix, and the spatial dimension enters only at the visualization (cartogram) stage.

inability to properly account for the temporal structure of observations and challenges in estimating appropriate covariance matrices (Hall, 2018; Wang et al., 2016). These limitations can result in suboptimal identification of underlying patterns and potentially misleading classifications of local labor markets.

Recent methodological advances have introduced more sophisticated variants of PCA that are better suited to handle the complexities of doubly multivariate data. Among these, kernel model for PCA (KPCA) has been shown to offer the advantage of capturing nonlinear relationships through kernel functions (Schölkopf et al., 1998; Hoffmann, 2007). In contrast, functional data model for PCA (FPCA) treats time-varying characteristics as continuous functions rather than discrete observations, thereby preserving the temporal continuity inherent in observed processes (Ramsay and Silverman, 2005; Shang, 2014).

The objective of this article is to employ more advanced principal component analysis techniques (KPCA and FPCA) preceding the classification of local labor markets in the Wielkopolskie voivodship at the county level. The aim of this study is to compare these approaches designed specifically for doubly multivariate data with results obtained from classical PCA. The present analysis is conducted over the period between 2004 and 2022, encompassing significant economic events including Poland's accession to the EU, the global financial crisis, and the pandemic caused by the COVID-19 virus. This enables the capture of both structural and cyclical dimensions of labor market dynamics (Lewandowski and Magda, 2023; Ciołek, 2021).

Beyond its methodological contributions, this research has substantial practical implications for the formulation of regional policy. The identification of clusters of counties with analogous labor market characteristics and trajectories is predicated on the analysis, with the objective being the development of tailored interventions that address specific local challenges (Pike et al., 2016). Furthermore, the methodology outlined here provides a flexible framework that can be extended to other domains of socioeconomic research in which spatial and temporal dimensions play crucial roles (Krzyśko et al., 2021, 2022).

The remainder of this article is structured as follows: Section 2, 3 and 4 provide a brief description of the principal component approaches used in our study. Section 5 and 6 describe the data and research procedure respectively. Section 7 presents the empirical results and classification of local labor markets in the Wielkopolskie voivodship using different PCA approaches. Finally, Section 8 concludes with policy recommendations and directions for future research.

2. Principal component analysis

The primary statistical method implemented in this study is based on classical principal component analysis, which can be conceptualized as a multivariate technique for transforming observable primary variables into a new set of mutually orthogonal variables, referred to as principal components (Gray, 2017; Kassambara, 2017). In this approach, the original data is linearly transformed into a new coordinate system such that the axes (principal components) correspond to orthogonal directions that successively capture the maximum possible variance in the data and can be easily identified and interpreted. The number of components that can be determined is equal to the number of original variables. These new

variables are linear combinations of observable variables:

$$Z_k = a_{k1}X_1 + a_{k2}X_2 + \ldots + a_{kp}X_p, \quad k = 1, \ldots, p, \tag{1}$$

where $a_{k1}, a_{k2}, \ldots, a_{kp}$ are the elements of the eigenvector corresponding to the k-th largest eigenvalue of the covariance matrix $\mathbf{\Sigma}$ of the vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ and p denotes the number of variables. The successive principal components explain a decreasing amount of the total variance of the variables. It can be demonstrated that the sum of the variances of the principal components is equal to the sum of the variances of the primary variables. The methodology for determining the principal components is based on the calculation of the eigenvalues and the eigenvectors of the covariance matrix $\mathbf{\Sigma}$. In practice by limiting the analysis to the first two principal components, it is possible to represent the observed objects on a plane, thereby making it easier to identify the relationships that exist between them. The percentage of variance carried out by the first two principal components determines the degree of variation of the original variables reflected in the visualization (Rencher and Christensen, 2012).

In the classical approach, the construction of principal components is based on n observations being p-dimensional vectors. The incorporation of time-variation in the aforementioned procedure results in the augmentation of observations with an additional dimension corresponding to T moments of time. The subsequent analysis is predicated on n observations (objects) being $p \times T$ dimensional matrices. The initial step to perform PCA is vectorization, which yields objects that are vectors of $p \cdot T$ coordinates. In this case, the construction of the principal components can be distilled to the estimation of the covariance matrix $\mathbf{\Sigma}$ of degree $p \times T$. The condition for the correct estimation of this matrix is that $n > p \cdot T$. It means that the estimator of covariance matrix $\mathbf{\Sigma}$ will be positively definite, otherwise ($n \leq p \cdot T$) it will be non-negative definite. This does not hinder the construction of principal components, but it weakens the model's specification.

In this paper, we present two potential solutions to the aforementioned problem:

- a model based on non-linear transformation carried out by the kernel function (Kernel model - KPCA),
- a model based on transformation into functional data (Functional data model - FPCA).

The proposed models not only allow a significant reduction in multidimensionality, but also often allow for the satisfaction of sample size requirements. Moreover, the kernel model allows non-linearity to be exploited and functional data model captures temporal continuity. In addition, both models (when used correctly) significantly reduce computation time and power requirements as well.

3. Kernel model

This approach uses a computed centered kernel matrix to non-linearly map the input data into a Hilbert space (Schölkopf et al., 1998). Let us consider a dataset $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$,

$\boldsymbol{X}_i \in \mathbb{R}^{T\times p},\ i = 1,2,\ldots,n$, where p is the number of variables and T is the number of time points in which these variables are observed. Initially, the dataset is mapped into Hilbert space $\mathscr{H}$ by $\phi\colon \mathbb{R}^{T\times p} \to \mathscr{H}$.

To perform linear principal component analysis in feature space $\mathscr{H}$, we need to find eigenvalues $\lambda \geq 0$ and eigenvectors $u \in \mathscr{H}$ of the empirical covariance operator, given by:

$$C = \frac{1}{n}\sum_{i=1}^{n} \langle \psi(\boldsymbol{X}_i), \psi(\boldsymbol{X}_i)\rangle_{\mathscr{H}} \tag{2}$$

satisfying:

$$Cu = \lambda u, \tag{3}$$

where $\langle u,u\rangle_{\mathscr{H}} = 1$ and $\psi(\boldsymbol{X}_i) = \phi(\boldsymbol{X}_i) - \bar{\phi},\ \bar{\phi} = \frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{X}_i),\ i = 1,2,\ldots,n$.

Equation (3) leads us to conclude that the following coefficients $\alpha_1,\ldots,\alpha_n$ exist and

$$u = \sum_{i=1}^{n} \alpha_i \psi(\boldsymbol{X}_i). \tag{4}$$

We can equivalently consider n equations of the form:

$$\langle \psi(\boldsymbol{X}_i), Cu\rangle_{\mathscr{H}} = \lambda \langle \psi(\boldsymbol{X}_i), u\rangle_{\mathscr{H}}. \tag{5}$$

If we combine (3), (4) and (5), we get

$$\tilde{\boldsymbol{K}}\boldsymbol{\alpha} = n\lambda\boldsymbol{\alpha}, \tag{6}$$

where $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_n)'$, $\tilde{\boldsymbol{K}} = \boldsymbol{HKH}$, $\boldsymbol{K} = k(\boldsymbol{X}_i,\boldsymbol{X}_j)$, $\boldsymbol{H} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n',\ \mathbf{1}_n \in \mathbb{R}^{n\times 1}$ and $\boldsymbol{H}$ is centering matrix and $k()$ is an appropriately selected kernel function. The obtained matrix $\tilde{\boldsymbol{K}}$ is the basis for the construction of the principal components (by means of its eigenvectors $\boldsymbol{\alpha}_i$ and eigenvalues λ_i) (Krzyśko et al., 2018).

4. Functional data model

Let us suppose that we are observing a p-dimensional stochastic process:

$$\boldsymbol{X}(t) = (X_1(t), X_2(t),\ldots,X_p(t))',\ \boldsymbol{X}(t) \in L_2^p(I),\ \mathrm{E}(\boldsymbol{X}(t)) = \mathbf{0}, \tag{7}$$

where $L_2(I)$ is the Hilbert space of square integrable functions on the interval I. Furthermore, suppose that the k-th component of the process $\boldsymbol{X}(t)$ can be represented by a finite number of basis functions $\{\varphi_b\}$:

$$X_k(t) = \sum_{b=0}^{B_k} c_{kb}\varphi_b(t),\ t \in I, \tag{8}$$

where c_{kb} are random variables such that $\mathrm{E}(c_{kb}) = 0$, $\mathrm{Var}(c_{kb}) < \infty$. (e.g. Ramsay, Silverman, 2005). The process $\boldsymbol{X}(t) \in L_2^p(I)$ can be written as:

$$\boldsymbol{X}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c},\ t \in I, \tag{9}$$

where $\boldsymbol{c} = (c_{10}, \ldots, c_{1B_1}, \ldots, c_{p0}, \ldots, c_{pB_p})'$, $\boldsymbol{c} \in \mathbb{R}^{B+p}$, $B = B_1 + B_2 + \cdots + B_p$, $\boldsymbol{\Phi}(t) = \mathrm{diag}(\boldsymbol{\varphi}'_{B_1}(t), \ldots, \boldsymbol{\varphi}'_{B_p}(t))$, $\boldsymbol{\varphi}_{B_k}(t) = (\varphi_0(t), \varphi_1(t), \ldots, \varphi_{B_k}(t))'$, $k = 1, 2, \ldots, p$. Note that all information about process $\boldsymbol{X}(t)$ is in the random vector $\boldsymbol{c}$ whereby:

$$\boldsymbol{\mu}_{\boldsymbol{X}}(t) = \boldsymbol{\Phi}(t)\boldsymbol{\mu}_{\boldsymbol{c}},\ t \in I, \quad \boldsymbol{\Sigma}_{\boldsymbol{X}}(s,t) = \boldsymbol{\Phi}(s)\boldsymbol{\Sigma}_{\boldsymbol{c}}\boldsymbol{\Phi}'(t),\ s,t \in I \tag{10}$$

with $\boldsymbol{\mu}_{\boldsymbol{c}} = \mathrm{E}(\boldsymbol{c})$ i $\boldsymbol{\Sigma}_{\boldsymbol{c}} = \mathrm{Var}(\boldsymbol{c})$. Our further assumption is that the random process $\boldsymbol{X}(t)$ has a representation given by formula (9). The principal component analysis is based only on the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{c}}$ of the process $\boldsymbol{X}(t)$, however, it is not known in practice. We are, therefore, forced to estimate this matrix using an independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of a random process $\boldsymbol{X}(t)$. Assuming that $\boldsymbol{x}_i$ is of the form (9) we obtain $\hat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_1, \cdots, \hat{\boldsymbol{c}}_n)$, then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{c}} = \frac{1}{n}\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}'$. The obtained covariance matrix is the basis for the construction of the principal components (by means of the relationship with the eigenvalues of the matrix) (Górecki et al., 2014, 2018).

5. Labor market data

To classify local labor markets in Poland in the Wielkopolskie voivodship at the county level a dataset of $p = 12$ variables presented in Table 1 has been used.

Table 1. List of variables

Variable	Description	Type
1	share of the registered unemployed with higher education in registered unemployed total	D
2	share of the registered unemployed out of job for longer than 1 year in registered unemployed total	D
3	share of the newly registered unemployed in the registered unemployed total	D
4	share of the registered unemployed aged less than 25 in the registered unemployed total	D
5	share of the registered unemployed aged 55 and older in the registered unemployed total	D
6	share of the employed in the population in the working age	S
7	average monthly gross salary (in PLN)	S
8	entities registered in REGON per 1000 population	S
9	entities newly registered in REGON per 1000 population	S
10	entities removed from REGON per 1000 population	D
11	registered unemployment rate	D
12	job offers per 1000 population in the working age	S

Note. The register REGON fulfills the function of the national official Register of National Economy Entities. It is the only integrated register in Poland covering all of the national economy entities.

All variables taken into account in the empirical study come from the Polish database called Local Data Bank - BDL[5], which is the main source of information about the social, economic, demographic and environmental conditions of Poland. All the characteristics considered in analysis were selected with a view to obtain a relatively complete description of the labor market situation in the Wielkopolskie voivodship. These variables effec-

[5] `https://bdl.stat.gov.pl`

tively capture multiple dimensions of labor market volatility and dynamics, thus allowing for a thorough assessment of economic conditions across counties. The selected variables form a well-balanced framework that covers four critical aspects of labor market analysis.

I. Labor Supply Characteristics (Variables 1–5)

These variables provide detailed insights into the composition and characteristics of the unemployed population. The inclusion of educational level (variable 1) reveals skill mismatches and human capital availability. Duration of unemployment (variable 2) indicates structural issues and the effectiveness of active labor market policies. The share of newly registered unemployed (variable 3) shows labor market volatility and cyclical patterns. Age structure variables (variables 4 and 5) capture generational challenges and age-specific employment barriers, crucial for understanding demographic transitions in the labor market.

II. Labor Market Performance Indicators (Variables 6, 11, 12)

These core indicators directly measure market efficiency and balance. The working-age employment rate (variable 6) provides geographical context relative to population potential. The registered unemployment rate (variable 11) serves as the primary measure of labor market slack. Job offers per working-age population (variable 12) indicate labor demand intensity and market dynamics.

III. Economic Activity and Entrepreneurship (Variables 8–10)

These variables capture the entrepreneurial ecosystem and business dynamics. REGON entity density (variable 8) reflects business development maturity. New registrations (variable 9) indicate entrepreneurial activity and economic growth potential. Entity removals (variable 10) reveal business sustainability and market conditions.

IV. Labor Market Value and Productivity (Variable 7)

Average gross salary serves as a proxy for productivity levels, cost of living, and overall economic development across counties.

The data covers a period of time from 2004 to 2022 ($T = 19$) and describes in detail all $n = 35$ counties in this region. Values of all variables were unitized by the method of zero unitization (Jajuga and Walesiak, 2000). The main reason of this approach was that they were expressed in different units and had different value levels. In order to do this, it was necessary to isolate the stimulants and destimulants in the group of all potential variables, which are appropriately labeled as S and D in Table 1.

6. Research procedure

The study, which is based on a comprehensive analysis of all counties within the Wielkopolskie voivodship, employs a sample which consists of $n = 35$ observations[6] described by $p = 12$ statistical features (variables) measured at $T = 19$ moments in time. Therefore, each observation (county) is represented by a matrix of dimensions $p \times T = 12 \times 19$. As mentioned in Section 2, in classical PCA the correct estimation of covariance matrix $\mathbf{\Sigma}$

[6]This is the number of all counties in the Wielkopolskie voivodship in Poland.

requires that $n > p \cdot T$. It means that the estimator of covariance matrix $\mathbf{\Sigma}$ will be positively definite, otherwise ($n \leq p \cdot T$) it will be non-negative definite. In our case, we have $35 \ll 12 \cdot 19 = 228$. While this does not hinder the construction of the principal components, it does weaken the specification of the model. Therefore, to overcome the problems mentioned above in the estimation of the covariance matrix $\mathbf{\Sigma}$ and the aspiration to encapsulate continuity in the temporal domain, the authors used two enhanced models to obtain the principal components: the kernel model and the functional data model. The research procedure consisted of several steps including preparing the data for analysis, carrying out principal component analysis using different models, determining the relevant dendrograms and cartograms followed by interpretation of the results. This procedure has been described in detail in the Diagram 1.

Diagram 1. Research procedure

Step 1: Preparation of variables
Obtaining all the most important variables describing the labor market from the Local Data Bank (`bdl.stat.gov.pl`) and calculated into relevant indices.

Step 2: Using a zero-unitarization method
Making variables comparable using their type (stimulants - S and destimulants - D) (Jajuga and Walesiak, 2000).

Step 3: Construction of principal components

- using classical principal component analysis (PCA)
- using Gaussian kernel - (Kernel model - KPCA)

$$\mathrm{k}(\boldsymbol{X}_i, \boldsymbol{X}_j) = \exp(-\gamma \left\| \boldsymbol{X}_i - \boldsymbol{X}_j \right\|_F^2), \quad \gamma > 0, \quad \|A\|_F = \sqrt{tr(A'A)}$$

- using functional data transformation based on Fourier basis system in the space $L_2([0,T])$ - (Functional data model - FPCA)

$$\varphi_0(x) = \frac{1}{\sqrt{T}},\ \varphi_{2k-1}(x) = \frac{\sqrt{2}}{\sqrt{T}} \sin \frac{2\pi k x}{T},\ \varphi_{2k}(x) = \frac{\sqrt{2}}{\sqrt{T}} \cos \frac{2\pi k x}{T},\ k = 1, 2, \ldots,$$

where $\|A\|_F = \sqrt{tr(A'A)}$ denotes the Frobenius norm

Step 4: Cluster analysis
Using hierarchical Ward's method (Ward, 1963) based on the first two principal components to identify similar groups, create appropriate dendrograms and cartograms.

Step 5: Interpretation of results
Evaluation of the quality of the results obtained, recognition of differences and similarities and complex interpretation.

7. Results

The variance explained for the first three principal components across all models implemented in this study is presented in Table 2. The findings reveal significant variations in the efficacy of these methods when applied to the doubly multivariate labor market data from the Wielkopolskie voivodship.

Table 2. Variance of original data explained by principal components

Model		Component 1	Component 2	Component 3
PCA	% of variance	43.52%	12.97%	7.34%
	cumulative %	43.52%	56.49%	63.83%
KPCA	% of variance	16.17%	12.11%	8.97%
	cumulative %	16.17%	28.28%	37.25%
FPCA	% of variance	48.51%	14.29%	7.55%
	cumulative %	48.51%	62.80%	70.35%

FPCA demonstrates superior performance in capturing the underlying structure of labor market indicators. The first component derived from FPCA alone explains 48.51% of the total variance in the original dataset, significantly outperforming both classical PCA (43.52%) and especially KPCA (16.17%). This marked difference demonstrates the advantage of the functional approach in properly accounting for the temporal structure inherent in the data, which spans from 2004 to 2022.

The cumulative explained variance for the first two components further reinforces the methodological advantages of FPCA. While classical PCA achieves a respectable 56.49% explained variance with two components, FPCA increases this to 62.80%. In contrast, KPCA demonstrates substantially weaker performance, with its first two components explaining a mere 28.28% of the total variance. Interestingly, the incremental contribution of the second and third principal components remains relatively consistent across all three methods. The second component adds approximately 12–14% of the total variance, while the third component contributes 7–9%. This suggests that FPCA's primary methodological advantage lies in its ability to capture dominant patterns of variation in the first component more effectively, likely due to its better preservation of the temporal continuity of labor market indicators across the study period.

The relatively poor performance of KPCA is somewhat surprising given its theoretical advantages in capturing non-linear relationships. This finding suggests that for the labor market indicators examined in this study, the temporal structure of the data (which FPCA is specifically designed to address) represents a more significant source of variation than any non-linear relationships between variables (which KPCA aims to capture). It is also important to note that, despite the fact that KPCA displayed the poorest performance in terms of goodness measures, it seems to have minimal influence on the final classification results. This provides valuable methodological guidance for future studies of regional labor markets using similar indicator sets. However, due to the low quality of the measure of explained variance, the results obtained using KPCA method will not be presented graphically in the remainder of this article.

The results of this study hold significant implications for the classification of local labor markets in the Wielkopolskie voivodship. The considerably elevated explained variance achieved by FPCA suggests that this approach captures significant patterns in the data that other methods may partially miss, potentially resulting in more accurate and informative clustering of counties. Consequently, the subsequent cluster analysis and interpretation will prioritize the components derived from the FPCA method, although comparisons with classical PCA will be maintained to ensure the robustness of the findings.

Figures 1 and 2 present the projection of counties onto the plane of the first two principal components using PCA and FPCA methods respectively. Projecting the results onto a plane R^2 locates the counties relative to the four quadrants of the Cartesian coordinate system. The homogeneous clusters of counties themselves were subsequently identified by means of Ward's hierarchical method (see Figure 3 and 4). Once the FPCA projection is reoriented so that both methods share the same sign convention for the principal components[7], the same counties occupy the same quadrants under both approaches. For example, the counties of Poznań, poznański, Leszno and Kalisz are located in the fourth quadrant of the coordinate system in both the PCA and the FPCA methods, which makes the two projections directly comparable.

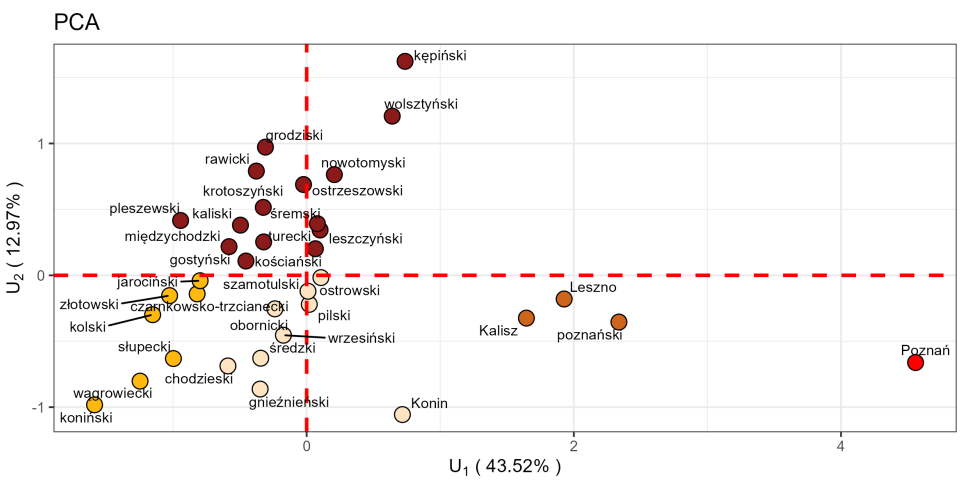

Figure 1. Counties projection on the plane of the first two principal components in the PCA method

To facilitate the analysis of the obtained results dendrograms using hierarchical Ward's method based on the first two principal components were created to identify homogeneous clusters of counties in terms of the labor market situation. In line with the adopted research design, five clusters were distinguished in both dendrograms - the dashed lines in Figures 3 and 4 mark the corresponding cut-off. Five groups were retained because they provide an

[7]The sign of the principal components is determined only up to multiplication by -1. The FPCA scores were therefore reoriented to match the orientation of the PCA projection, which does not affect the explained variances, the variable contributions or the clustering results.

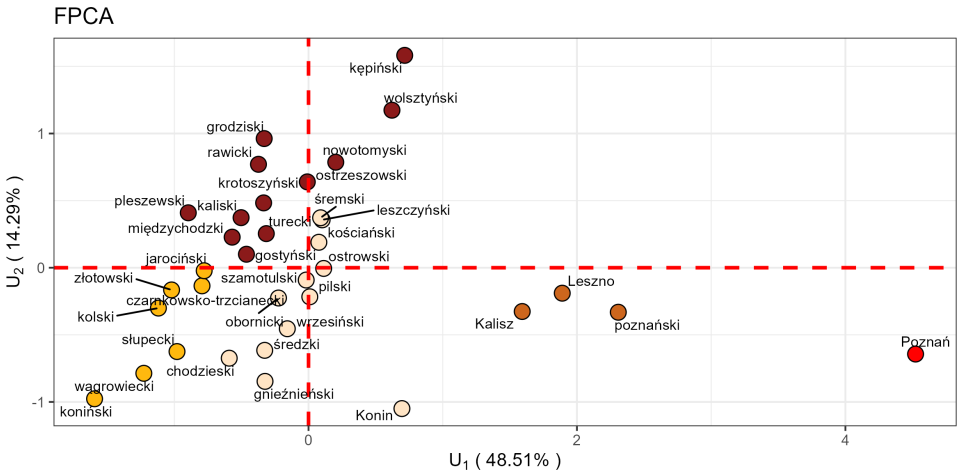

Figure 2. Counties projection on the plane of the first two principal components in the FPCA method

interpretable and substantively meaningful partition of the local labor markets, isolating in particular the city of Poznań as a singleton and the former provincial cities together with the poznański county as a distinct second group.

In order to better compare the obtained clusters, the colors used in the dendrograms correspond to those used in Figures 1 and 2. It can be noted that the clusters obtained by Ward's method mostly coincide with the clusters using the projections of the first two principal components (for PCA and FPCA approaches respectively). From this point of view, we will base our analysis of the results on the dendrograms.

Upon careful examination, it becomes apparent that both dendrograms exhibit remarkably similar hierarchical structures. The classical principal component analysis (PCA) dendrogram and the functional principal component analysis (FPCA) dendrogram display nearly identical clustering patterns, with objects grouped in highly consistent arrangements across both methods. This high degree of concordance is particularly notable given the distinct theoretical foundations of the two approaches. The branching patterns in both dendrograms reveal consistent cluster formations, with the corresponding objects joining at similar heights.

This suggests that, despite classical PCA treating each measurement as an independent variable and FPCA preserving the functional continuity of the data, both methods essentially capture the same underlying structure in our dataset. The primary clusters identified by both approaches are highly aligned, indicating robust groupings that persist regardless of the analytical method employed. In both approaches, Poznań forms a singleton (one-element cluster) as the capital of the Wielkopolskie voivodship and a city with county rights. The isolation of Poznań as a singular cluster in both dendrograms highlights its distinctive labor market characteristics within the Wielkopolskie voivodship. As both the provincial capital and a city with county rights, Poznań has a significantly better labor market situation than

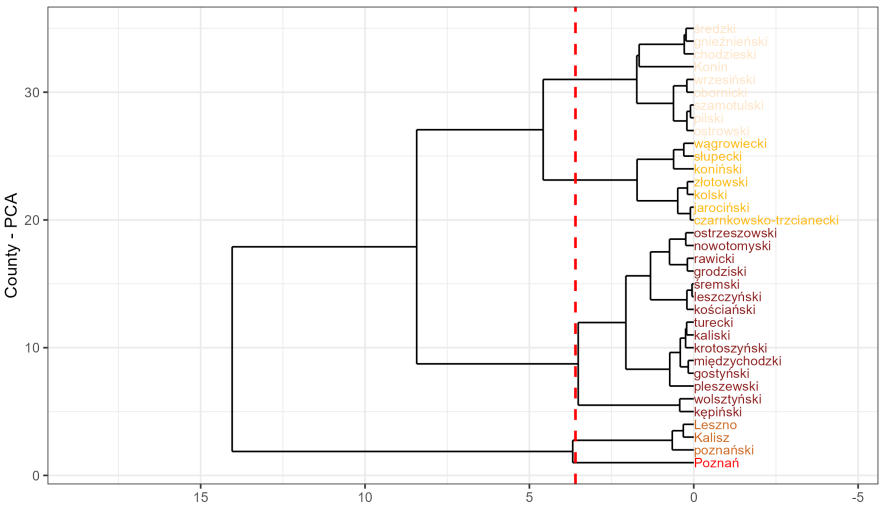

Figure 3. A dendrogram for counties in the PCA method

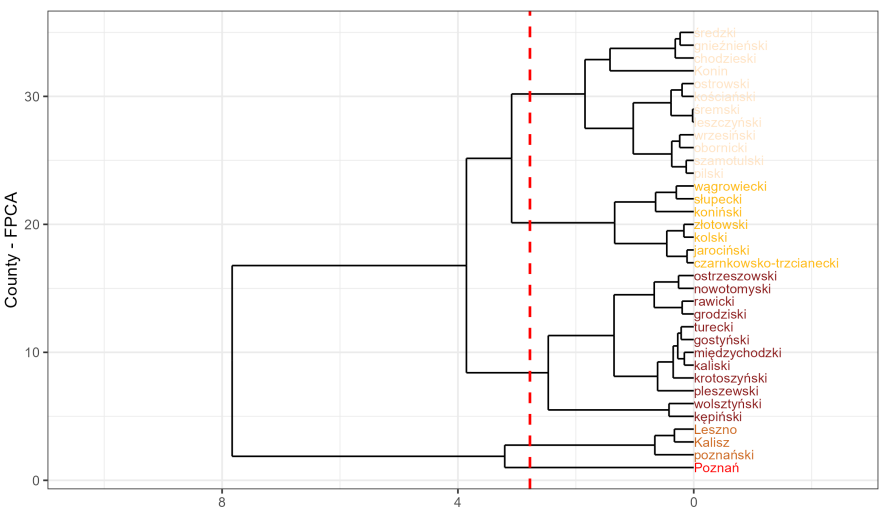

Figure 4. A dendrogram for counties in the FPCA method

surrounding counties, driven by several key factors. With one of the lowest unemployment rates not only in the region but also in Poland, Poznań consistently outperforms the provincial average. This advantageous position reflects the city's robust and diverse economic base, which effectively absorbs the available labor force. Its status as an administrative, educational and business hub creates high demand for workers across various sectors and qualification levels. Wage levels in Poznań are substantially higher than in other counties in

the Wielkopolskie voivodship. This is due to the concentration of high-value-added sectors, such as IT, business services and advanced manufacturing, which tend to offer more competitive compensation packages. The presence of numerous multinational corporations also contributes to elevated salary standards across the local labor market.

Compared to other counties, which often rely more heavily on agriculture or single industries, Poznań's economy exhibits greater sectoral diversification. The city boasts a balanced mix of the services, manufacturing and construction sectors. This diversification provides greater economic resilience and stability in employment patterns, which is why Poznań stands out in our analysis. The presence of major universities and research institutions creates a cycle of self-reinforcing human capital development. These institutions employ large numbers of highly qualified staff and produce graduates who enter the local labor market. This attracts knowledge-intensive businesses and further enhances the city's economic position. Historically, Poznań's position at the intersection of major transport routes (including the Berlin-Warsaw corridor) has fostered trade and economic development. The city's modern logistics infrastructure supports its role as a regional business center, creating employment opportunities in transportation, logistics, and related services. The consistent identification of Poznań as a distinct cluster using both the PCA and FPCA methodologies confirms that these labor market characteristics reflect robust structural differences. This finding emphasizes the significant economic disparities within the voivodship and highlights Poznań's role as the dominant economic center of the Wielkopolskie, with labor market conditions that differ quantitatively and qualitatively from those in surrounding areas.

Our hierarchical clustering analysis identified a second cluster consistently across both analytical approaches, comprising the cities of Kalisz and Leszno together with the poznański county surrounding the capital. This grouping displays labor market characteristics that position these areas favorably compared with the remaining counties of the Wielkopolskie voivodship, although they remain distinct from the exceptional status of Poznań. Kalisz and Leszno held provincial capital status prior to Poland's 1999 administrative reform. This historical role fostered substantial administrative, educational and business infrastructure and important regional functions (administrative offices, courts and specialized services) that continue to provide stable employment for qualified personnel. The poznański county, which surrounds the provincial capital, benefits markedly from its proximity to Poznań and exhibits the characteristics of suburban economic development. It is worth noting that the city of Konin, although also a former provincial capital, does not join this second cluster in our analysis but falls into one of the intermediate groups, which is consistent with the comparatively weaker position of the eastern part of the voivodship reported by other authors.

Further analysis identified three additional clusters comprising the remaining counties of the Wielkopolskie voivodship. Two of them form an intermediate tier: they have a broadly comparable overall standing but are differentiated mainly along the second functional principal component, i.e. by the structure and persistence of unemployment rather than by overall economic vitality. The first of these intermediate groups gathers counties located closer to the regional centers, including the city of Konin together with chodzieski, gnieźnieński, kościański, leszczyński, obornicki, ostrowski, pilski, szamotulski, średzki, śremski and wrzesiński (Cluster 3 in Table 3). The second comprises counties with predominantly indus-

trial or agricultural-industrial profiles, such as gostyński, grodziski, kaliski, kępiński, krotoszyński, międzychodzki, nowotomyski, ostrzeszowski, pleszewski, rawicki, turecki and wolsztyński (Cluster 2 in Table 3). Among them, kępiński has developed a nationally recognised furniture industry (the so-called Kępno furniture cluster), while wolsztyński combines one of the lowest unemployment rates in the voivodship with a strong entrepreneurial base in which industry and services clearly dominate over agriculture. The final cluster (Cluster 4 in Table 3), consisting of czarnkowsko-trzcianecki, jarociński, kolski, koniński, słupecki, wągrowiecki and złotowski, displays the least favorable labor market conditions. These are predominantly peripheral, agricultural counties located in the northern and eastern parts of the voivodship, characterized by lower productivity, lower average wages, limited demand for highly qualified specialists, and negative demographic trends such as the out-migration of young, educated working-age residents.

The clusters created in this way correspond to the cartogram shown in Figure 5, which illustrates the spatial differentiation of counties in the Wielkopolskie voivodship based on Ward's method and the two first principal components obtained in FPCA[8]. The cartogram illustrates the spatial distribution of the five labor market clusters identified through our hierarchical clustering analysis, revealing clear geographical patterns across the Wielkopolskie voivodship. Cluster 5 represents the most favorable labor market conditions and is concentrated exclusively in the city of Poznań, forming the region's economic core. Cluster 1, with strong but secondary performance, comprises the poznański county encircling the capital together with the cities of Leszno in the south-west and Kalisz in the south-east. Clusters 3 and 2 form an intermediate tier covering most of the central part of the voivodship: they share a broadly comparable overall standing and are differentiated mainly by the structure of unemployment rather than by overall economic vitality. Cluster 3 groups counties located closer to the regional centers, forming a belt around the Poznań agglomeration, whereas Cluster 2 gathers counties with more pronounced industrial or agricultural-industrial profiles, located predominantly in the southern and western parts of the voivodship. Cluster 4 displays the least favorable labor market conditions and occupies the northern and eastern peripheries, distant from the major economic centers and transport corridors. This spatial arrangement clearly reflects a core-periphery structure modified by historical development patterns: labor market performance generally declines with increasing distance from Poznań, except where former provincial cities and transport corridors create pockets of better economic conditions.

Our classification can be compared, at the level of the general spatial pattern, with recent studies of the same 35 counties that use different variables and methods. Kwiatkowski and Szymańska (2024), ranking the counties by a synthetic taxonomic development index for 2010–2022, place the city of Poznań at the top, followed by the cities of Leszno and Kalisz and the poznański county in their highest group - which, together with Poznań, constitute our most favorable clusters (the city of Konin falls into one of the intermediate groups in our analysis) - whereas the koniński county and other peripheral eastern counties (kaliski, słupecki, złotowski) form their lowest group, mirroring the core-periphery gradient obtained here.

[8]We present only one cartogram due to the similarities of the obtained dendrograms in the PCA and FPCA methods. The full composition of all clusters is given in Table 3.

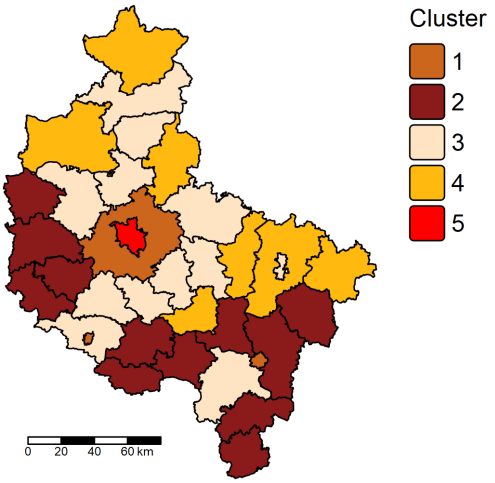

Figure 5. Clusters of counties in Ward's method based on the first two functional principal components (FPCA)

Table 3. Composition of the five clusters of counties (FPCA model, Ward's method), ordered from the most to the least favorable labor market conditions

Cluster 5	Cluster 1	Cluster 3	Cluster 2	Cluster 4
Poznań	Kalisz	chodzieski	gostyński	czarnkowsko-trzcianecki
	Leszno	gnieźnieński	grodziski	jarociński
	poznański	Konin	kaliski	kolski
		kościański	kępiński	koniński
		leszczyński	krotoszyński	słupecki
		obornicki	międzychodzki	wągrowiecki
		ostrowski	nowotomyski	złotowski
		pilski	ostrzeszowski	
		szamotulski	pleszewski	
		średzki	rawicki	
		śremski	turecki	
		wrzesiński	wolsztyński	

Note. Cluster numbers and the ordering of columns correspond to the colours and the legend used in the cartogram (Figure 5). The two intermediate clusters (Cluster 3 and Cluster 2) are comparable in overall standing and are differentiated mainly by the structure of unemployment (the second functional principal component). The assignment shown is based on the FPCA model; under PCA, the counties kościański, leszczyński and śremski move from Cluster 3 to Cluster 2.

A parallel study of registered unemployment in the same counties likewise identifies the city of Poznań, the poznański and wolsztyński counties among the most favorable and the koniński county as the least favorable (Kwiatkowski and Szymańska, 2022). At the national

scale, Kwiatkowski and Kwiatkowska (2020) place the Wielkopolskie among the voivodeships with the most favorable county labor markets and confirm a persistent west-east gradient. The placement of individual counties varies across these studies, reflecting their different variable sets (e.g. unemployment dynamics versus the broader set of twelve indicators used here) and grouping procedures, nevertheless the overarching picture is highly consistent, i.e. the dominance of Poznań and its agglomeration, the relatively strong position of the former voivodeship cities, and the weaker performance of the peripheral eastern counties. Obtained through synthetic indicators and panel econometric models rather than the functional approach adopted in this paper, this convergence supports the robustness of the spatial structure revealed by our classification.

In Table 4, the percentage contributions of variables to the first and second functional principal components reveal distinct structural patterns characterizing local labor markets in the Wielkopolskie voivodship. The analysis demonstrates that these markets can be effectively differentiated along two primary dimensions, each capturing fundamentally different aspects of regional economic performance.

The first functional principal component (U_1) emerges as an indicator of economic dynamism versus unemployment vulnerability. This component is predominantly shaped by business registration activity, with entities registered in REGON per 1000 population contributing 19.2% and newly registered entities adding 12.3%. Alongside these entrepreneurial indicators, the component is shaped by labor market measures pointing in the same direction, namely the share of employed persons in the working-age population (16.2%), as well as by unemployment measures acting in the opposite direction, including the registered unemployment rate (13.4%). Notably, the substantial contribution of unemployed individuals with higher education (15.6%) suggests that U_1 distinguishes regions where even skilled workers face employment challenges from those with robust economic opportunities. This component effectively separates economically vibrant areas characterized by high business formation and low unemployment from regions experiencing limited entrepreneurial activity and persistent employment difficulties.

The second functional principal component (U_2) focuses specifically on unemployment structure and persistence rather than overall economic vitality. The registered unemployment rate dominates this component with a 26.7% contribution, while structural unemployment characteristics play crucial roles: youth unemployment under 25 years (18.2%), long-term unemployment exceeding one year (13.1%), and the flow of newly registered unemployed (9.1%). This component appears to differentiate labor markets based on the severity and structural nature of unemployment problems, particularly emphasizing age-related employment dynamics and unemployment duration.

Several insights emerge from this decomposition. First, business registration activity, particularly through REGON variables, represents the most powerful differentiator of local labor markets, suggesting that entrepreneurial ecosystem strength fundamentally shapes regional economic performance.

Table 4. Variables in FPCA having the largest percentage contribution to the construction of the first (U_1) and the second (U_2) functional principal components

Variable	Description	Share (in %)
	Functional principal component U_1	
8	entities registered in REGON per 1000 population	19.2
6	share of the employed in the population in the working age	16.2
1	share of the registered unemployed with higher education in the registered unemployed total	15.6
11	registered unemployment rate	13.4
9	entities newly registered in REGON per 1000 population	12.3
4	share of the registered unemployed aged less than 25 in the registered unemployed total	11.2
5	share of the registered unemployed aged 55 and older in the registered unemployed total	5.0
10	entities removed from REGON per 1000 population	2.5
2	share of the registered unemployed out of job for longer than 1 year in the registered unemployed total	1.4
7	average monthly gross salary	1.3
12	job offers per 1000 population in the working age	1.2
3	share of the newly registered unemployed in the registered unemployed total	0.75
	Functional principal component U_2	
11	registered unemployment rate	26.7
4	share of the registered unemployed aged less than 25 in the registered unemployed total	18.2
6	share of the employed in the population in the working age	13.9
2	share of the registered unemployed out of job for longer than 1 year in the registered unemployed total	13.1
3	share of the newly registered unemployed in the registered unemployed total	9.1
9	entities newly registered in REGON per 1000 population	5.1
7	average monthly gross salary	4.2
10	entities removed from REGON per 1000 population	3.0
1	share of the registered unemployed with higher education in the registered unemployed total	2.9
5	share of the registered unemployed persons aged 55 and older in the registered unemployed total	1.5
8	entities registered in REGON per 1000 population	1.3
12	job offers per 1000 population in the working age	1.0

Second, the two components capture complementary rather than overlapping dimensions, with U_1 addressing the entrepreneurship-unemployment spectrum and U_2 focusing on unemployment's internal structure. Third, while youth unemployment significantly influences regional differentiation through U_2, older worker unemployment (55+) contributes minimally to both components. Finally, average gross salary shows limited explanatory power in both components, indicating that wage levels may be less important for distinguishing local labor markets compared to employment rates and business formation activity. This functional principal component structure suggests that regional labor markets are primarily characterized by their capacity for business creation and their specific unemployment patterns rather than wage differentials.

8. Conclusions

In this study, we aimed to explore the possibilities offered by principal component analysis (PCA) for doubly multivariate data in the classification of local labor markets for all 35 counties of the Wielkopolskie voivodship in Poland. We applied two different augmented PCA approaches (KPCA and FPCA) to the considered set of variables, and for comparative purposes, we also performed a classical PCA. The results obtained allowed us to identify the FPCA model (Functional data Principal Component Analysis) as the one with the highest variance explanation (for the first two and three principal components, see Table 2). Moreover, the results obtained from the FPCA model show a high degree of alignment with the PCA model results (treated as the reference). The projections of the counties onto the plane of the first two principal components differ only slightly when their relative location is taken into account, and the resulting cluster structure is largely consistent across the two methods. The single most favorable cluster (the city of Poznań) and the second cluster (the poznański county together with the cities of Kalisz and Leszno) are identical under both approaches. The differences are confined to three counties: kościański, leszczyński and śremski which are assigned to one of the intermediate clusters under PCA and to the other under FPCA, meaning that 32 out of 35 counties (more than 91%) are classified consistently. This high degree of agreement is regarded as both satisfactory and encouraging for further research on classification employing FPCA. It is essential to emphasize that the FPCA method introduces a unique perspective to the analysis performed, namely the treatment of observations as continuous phenomena in the time dimension (continuous functions), in contrast to the discrete approach underlying classical PCA.

The identification of distinct clusters with homogeneous labor market characteristics enables policymakers to move beyond one-size-fits-all approaches toward targeted interventions that address the specific structural challenges and opportunities within each group. This differentiated policy approach is particularly crucial in the context of limited public resources, where the effectiveness of interventions depends heavily on their alignment with local economic conditions and labor market dynamics. The temporal dimension captured through FPCA provides additional value by revealing the evolutionary trajectories of different regions, enabling policymakers to anticipate future challenges and design appropriate interventions. Furthermore, the replicable methodology presented here offers a practical framework for other regions and countries seeking to develop sophisticated, data-driven ap-

proaches to local labor market analysis and policy design. The demonstrated superiority of FPCA over classical methods suggests that incorporating temporal continuity into regional analysis can substantially improve policy targeting accuracy and resource allocation efficiency.

The limitations of this study include the fact that generalizing the obtained results is not justified, as the analysis was based on a single, specific example. A worthwhile consideration would be to verify the superiority of the FPCA model over KPCA and to conduct simulation studies. Another limitation is the arbitrary selection of parameters in the individual models, such as the kernel function, Fourier basis, or the number of its expansions. However, it is important to note that these parameters were not selected randomly but were based on previous studies (Górecki et al., 2014, 2018; Krzyśko et al., 2018, 2022). Moreover, the choice of Fourier basis functions may not optimally capture non-periodic structural breaks that characterize labor market responses to economic crises or policy changes. The relatively poor performance of KPCA may partly reflect suboptimal kernel parameter selection rather than inherent methodological limitations.

References

Blanchard, O. J., Katz L. F., Hall R. E. and Eichengreen, B., (1992). Regional evolutions. *Brookings papers on economic activity*, 1992(1), pp. 1–75.

Bosworth, D. L., Dawkins, P. and Stromback, T., (1996). *The economics of the labor market*. Longman Harlow.

Churski, P., Herodowicz, T., Konecka-Szydłowska, B. and Perdał, R. A., (2020). *Teoretyczny i praktyczny wymiar polityki rozwoju zorientowanej terytorialnie* [Theoretical and practical dimension of place-base territorial policy], 9(201), Polska Akademia Nauk.

Ciołek, D., (2021). Changes in the labour market during the COVID-19 pandemic and their spatial interactions - evidence from monthly data for Polish LAU. *Geographia Polonica*, 94(4), pp. 523–538.

Ciżkowicz, P., Kowalczuk, M., Rzońca, A., (2016). Heterogeneous determinants of local unemployment in Poland. Post-Communist Economies, 28(4), pp. 487–519.

Del Campo, C., Monteiro, C. M., and Soares, J. O., (2008). The European regional policy and the socio-economic diversity of European regions: A multivariate analysis. *European Journal of Operational Research*, 187(2), pp. 600–612.

Elhorst, J. P., (2003). The mystery of regional unemployment differentials: Theoretical and empirical explanations. *Journal of Economic Surveys*, 17(5), pp. 709–748.

Fihel, A., Okólski, M., (2019). Population decline in the post-communist countries of the European Union. *Population & Societies*, 567(6), pp. 1–4.

Fischer, M. M., Nijkamp, P., (Eds.), (2014). *Handbook of regional science (Vol. 3)*. Heidelberg: Springer.

Górecki, T., Krzyśko, M., Waszak, Ł. and Wołyński, W., (2014). Methods of reducing dimension for functional data. *Statistics in Transition New Series*, 15(2), pp. 231–242.

Górecki, T., Krzyśko, M., Waszak, Ł. and Wołyński, W., (2018). Selected statistical methods of data analysis for multivariate functional data. *Statistical Papers*, 59, pp. 153–182.

Gray, V., (2017). *Principal component analysis: methods, applications and technology*. Nova Science Publishers.

Hall, P., (2018). *Principal component analysis for functional data: methodology, theory, and discussion*, in Frédéric Ferraty, and Yves Romain (eds), Benchmark Methods for FDA, in Frédéric Ferraty, and Yves Romain (eds),The Oxford Handbook of Functional Data Analysis, Oxford Handbooks (2010; online edn, Oxford Academic, 8 Aug. 2018), pp. 210–234.

Hartal, S., Kourtit, K., Carmen Pascariu, G., (2023). Spatial labour markets and resilience of cities and regions: a critical review. Regional Studies, 57(12), pp. 2359–2372.

Hoffmann, H., (2007). Kernel PCA for novelty detection. *Pattern recognition*, 40(3), pp. 863–874.

Huber, P., (2007). Regional labour market developments in transition: A survey of the empirical literature. *The European Journal of Comparative Economics*, 4(2), pp. 263–298.

Jajuga, K., Walesiak, M., (2000). *Standardisation of data set under different measurement scales*. In Classification and Information Processing at the Turn of the Millennium: Proceedings of the 23rd Annual Conference of the Gesellschaft für Klassifikation eV, University of Bielefeld, March 10–12, 1999, pp. 105–112. Berlin, Heidelberg: Springer Berlin Heidelberg.

Jenkins, R. M., (2001). *Labor Markets and Economic Transformation in Postcommunist Europe*. In: Berg, I., Kalleberg, A.L. (eds) Sourcebook of Labor Markets. Plenum Studies in Work and Industry, pp. 135–162. Springer, Boston, MA.

Jolliffe, I. T., (2002). *Principal component analysis for special types of data*. In: Principal Component Analysis. Springer Series in Statistics. Springer, New York, NY.

Kanbur, R., Svejnar, J., (2009). *Labor markets and economic development*. Taylor & Francis.

Kassambara, A., (2017). *Practical guide to principal component methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*, Vol. 2. Sthda.

Keane, M. P., Prasad, E. S., (2006). Changes in the structure of earnings during the Polish transition. *Journal of Development Economics*, 80(2), pp. 389–427.

Krzyśko, M., Łukaszonek, W., Ratajczak, W. and Wołyński, W., (2018). Nonlinear principal component analysis for geographically weighted temporal – spatial data. *Acta Universitatis Lodziensis. Folia Oeconomica*, 4(337), pp. 169–181.

Krzyśko, M., Nijkamp, P., Ratajczak, W., Wołyński, W., (2022). Multidimensional economic indicators and multivariate functional principal component analysis (MFPCA) in a comparative study of countries' competitiveness. *Journal of Geographical Systems*, 24(1), pp. 49–65.

Krzyśko, M., Wołyński, W., Szymkowiak, M., and Wojtyła, A., (2021). A spatio-temporal analysis of the health situation in Poland based on functional discriminant coordinates. *International Journal of Environmental Research and Public Health*, 18(3), 1109.

Kwiatkowski, E., Kwiatkowska, E., (2020). Zróżnicowanie poziomu i charakteru bezrobocia w przekroju powiatów w Polsce. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie* [Differentiation of the level and nature of unemployment across poviats in Poland. *Krakow Review of Economics and Management*], 3(987), pp. 7–29.

Kwiatkowski, E., Szymańska, A., (2022). Zróżnicowanie poziomu i charakteru bezrobocia w powiatach województwa wielkopolskiego [Differentiation of the level and nature of unemployment in the poviats of the Wielkopolskie voivodeship]. *Rozwój Regionalny i Polityka Regionalna*, (62), pp. 131–149.

Kwiatkowski, E., Szymańska, A., (2024). Efekty grawitacyjne i zróżnicowanie poziomu rozwoju gospodarczego powiatów w województwie wielkopolskim w latach 2010–2022 [Gravity effects and the differentiation of the level of economic development of poviats in the Wielkopolskie voivodeship in 2010–2022]. *Rozwój Regionalny i Polityka Regionalna*, 17(72), pp. 11–32.

Lewandowski, P., Magda, I., (2023). *The labor market in Poland, 2000–2021*. IZA World of Labor 2023, p. 426.

Martin, R.L., Morrison, P. S., (Eds.), (2003). *Geographies of labour market inequality*. Routledge

McQuaid, R. W., Green, A. E. and Danson, M., (Eds.), (2013). *Employability and local labor markets*. Routledge.

Pike, A., Rodríguez-Pose, A., and Tomaney, J., (2016). *Local and regional development*. Routledge.

Ramsay, J.O., Silverman, B. W., (2005). *Functional data analysis*, 2nd ed. Springer, Berlin.

Rencher, A.C., Christensen, W. F., (2012). *Methods of Multivariate Analysi*s, Third Edition, John Wiley & Sons.

Schölkopf, B., Smola, A. and Muller, K. B., (1998). Nonlinear component analysis as a kernel eigenvalues problem. *Neural Computation*, 10(5), pp. 1299–1319.

Shang, H. L., (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98, pp. 121–142.

Smętkowski, M., (2018). The role of exogenous and endogenous factors in the growth of regions in Central and Eastern Europe: the metropolitan/non-metropolitan divide in the pre-and post-crisis era. *European Planning Studies*, 26(2), pp. 256–278.

Wang, J.L., Chiou, J. M., and Müller, H. G., (2016). Functional data analysis. *Annual Review of Statistics and its Application*, 3(1), pp. 257–295.

Ward, J. H., (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), pp. 236–244.

Woźniak, M., (2021). Spatial matching on the urban labor market: estimates with unique micro data. *Journal for Labour Market Research*, 55(1), 11.

Yenilmez, F., Esin, K. (Eds.), (2016). *Handbook of research on unemployment and labor market sustainability in the era of globalization*. IGI Global.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 47–67, https://doi.org/10.59139/stattrans-2026-027
Received – 28.10.2025; accepted – 04.05.2026

Confidence intervals for the Double XShanker distribution parameter: a simulation and application study

Wararit Panichkitkosolkul[1], **Patchanok Srisuradetchai**[2]

Abstract

We examine four two-sided confidence intervals for parameter θ of the Double XShanker distribution: the likelihood-based, Wald-type, bootstrap-t, and the bias-corrected and accelerated (BCa) bootstrap intervals. An explicit expression for the observed Fisher information is derived, allowing the Wald interval to be computed directly from the maximum likelihood estimate. Interval performance is evaluated by empirical coverage probability (ECP) and average width (AW) through the Monte Carlo experiments over a range of sample sizes and parameter values. The likelihood-ratio and Wald intervals remain close to the nominal 0.95 level across most settings considered, whereas the bootstrap-t and BCa intervals tend to be narrower but show undercoverage in small samples. An additional sensitivity analysis using $B = 1000, 2000$, and 5000 bootstrap resamples indicates that the bootstrap-based conclusions are stable with respect to the number of resamples. The methods are also illustrated using rainfall data from the Los Angeles Civic Center, where the empirical interval estimates are consistent with the simulation findings.

Keywords: parameter estimation, confidence interval, Monte Carlo simulation, coverage probability, bootstrap.

1. Introduction

Accurate parameter estimation is central to statistical modeling, as it supports reliable inference and informed decision-making. This is particularly important in lifetime or survival analysis, where inference is based on time-to-event outcomes such as failure times, survival probabilities, and hazard rates. In practical fields including medical research, engineering, and actuarial science, imprecise or biased estimates may lead to misleading conclusions regarding system reliability, prognosis, or risk assessment.

To accommodate the diverse shapes observed in empirical lifetime data, various generalized and mixed probability distributions have been proposed to increase modeling flexibility (Nadarajah & Kotz, 2006). For example, Shukla (2018) introduced the Ram Awadh distribution as a two-component mixture combining an exponential distribution with scale parameter θ and a gamma distribution with shape parameter σ and common scale parameter θ. Messaadia and Zeghdoudi (2018) proposed the Zeghdoudi distribution, constructed from

[1]Thammasat University Research Unit in Mathematical Sciences and Applications, Pathum Thani, Thailand. E-mail: wararit@mathstat.sci.tu.ac.th. ORCID: https://orcid.org/0000-0001-8315-8185.

[2]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani, Thailand. E-mail: patchanok@mathstat.sci.tu.ac.th. ORCID: https://orcid.org/0000-0002-9925-9615.

a mixture of $Gamma(2,\theta)$ and $Gamma(3,\theta)$ distributions. Related developments include compounded models such as the Gamma zero-truncated Poisson distribution (Niyomdecha et al., 2023), which further illustrate the continued interest in flexible lifetime distributions. More recently, Etaga et al. (2023) introduced the Double XShanker distribution by combining the exponential distribution with the XShanker distribution. This construction yields a flexible one-parameter model capable of representing a range of lifetime data patterns.

Although the Double XShanker distribution has been studied from a distributional perspective, the finite-sample behaviour of confidence intervals for its parameter has received little attention. Interval estimation in non-standard or flexible models may exhibit non-negligible small-sample effects. For instance, Srisuradetchai and Dangsupa (2022) examined likelihood-based and score-type CIs for the geometric parameter in a zero-inflated geometric model and demonstrated the importance of coverage accuracy in small samples.

Motivated by this gap, this study considers four methods for constructing two-sided CIs for the parameter of the Double XShanker distribution: the likelihood-based interval, the Wald-type interval, the bootstrap-*t* interval, and the bias-corrected and accelerated (BCa) bootstrap interval. Their performance is assessed through Monte Carlo simulation across a range of sample sizes and parameter values. An empirical application to rainfall data is also provided to illustrate the methods in practice.

2. Methodology

This section defines the XShanker and Double XShanker models, outlines maximum-likelihood estimation for θ, and specifies the four confidence-interval constructions used in the simulation and data analysis.

2.1. The XShanker and the Double XShanker distributions

The XShanker distribution is formulated as a mixture of the exponential and Shanker distributions (Shanker, 2015), combined using specific mixing probabilities. In this construction, the exponential component has rate parameter θ, and the Shanker component is parameterized by θ. Let X denote a random variable following the XShanker distribution with parameter θ. The probability density function (pdf) of the XShanker distribution can be derived using a mixed model comprising two components, each weighted by their respective mixing probabilities, as follows:

$$f_{\text{XShanker}}(x;\theta) = p \cdot f_{\text{Exp}}(x;\theta) + (1-p) \cdot f_{\text{Shanker}}(x;\theta),$$

where $f_{\text{Exp}}(x;\theta) = \theta e^{-\theta x}$ is the pdf of the exponential distribution with rate θ, $f_{\text{Shanker}}(x;\theta) = \frac{\theta^2}{\theta^2+1}(\theta+x)e^{-\theta x}$ is the pdf of the Shanker distribution with parameter θ, and the mixing proportion is $p = \theta^2/(\theta^2+1)$. The pdf of the XShanker distribution is expressed as follows:

$$f_{\text{XShanker}}(x;\theta) = \frac{\theta^2}{(\theta^2+1)^2}\left[\theta^3+2\theta+x\right]e^{-\theta x}, \quad x>0,\ \theta>0. \tag{1}$$

Figure 1 illustrates the pdf plot of the XShanker distribution for various parameter values.

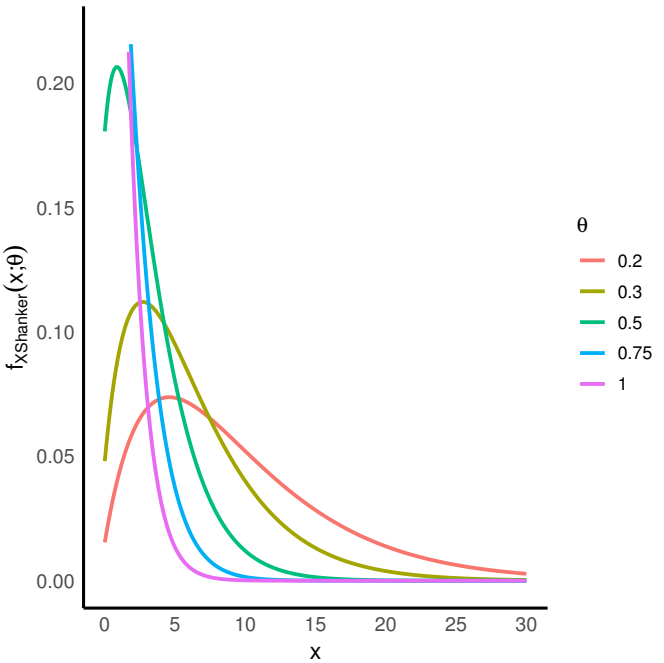

Figure 1. The pdf plot of the XShanker distribution for different parameter settings

Etaga et al. (2023) introduced the Double XShanker distribution, a one-parameter probability distribution developed as a more flexible extension of the XShanker distribution. Let $X \sim$ Double XShanker(θ). The density can be written as

$$f_{\text{Double XShanker}}(x;\theta) = p \cdot f_{\text{Exp}}(x;\theta) + (1-p) \cdot f_{\text{XShanker}}(x;\theta),$$

where $f_{\text{Exp}}(x;\theta)$ is the pdf of the exponential distribution with rate θ, $f_{\text{XShanker}}(x;\theta)$ is the pdf of the XShanker distribution, given in Equation (1) and the mixing proportion is $p = \theta^2/(\theta^2+1)$. The pdf of the Double XShanker distribution is mathematically defined in the equation below

$$f_{\text{Double XShanker}}(x;\theta) = \frac{\theta^2}{(\theta^2+1)^3}\left(\theta^5+3\theta^3+3\theta+x\right)e^{-\theta x}, \quad x>0,\ \theta>0.$$

Figure 2 provides a graphical representation of the pdf of the Double XShanker distribution, showcasing how the distribution behaves under a range of different parameter settings.

2.2. Point parameter estimation

We estimate θ by maximum likelihood. For a sample $x_1,\ldots,x_n$ from the Double XShanker model, the likelihood and score equations are given below.

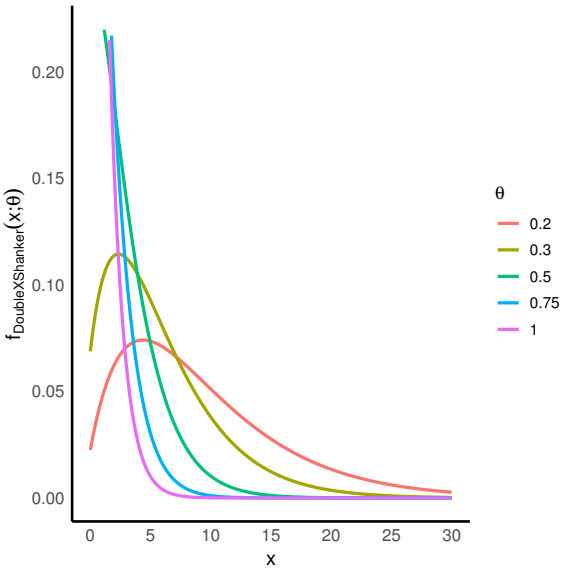

Figure 2. The pdf plot of the Double XShanker distribution for different parameter settings

Step 1: Likelihood

The likelihood function $L(\theta;x_i)$ represents the joint probability of observing a random sample drawn from the Double XShanker distribution. It is defined as:

$$L(\theta;x_i)=\prod_{i=1}^{n}f(x_i;\theta)=\left(\frac{\theta^2}{(\theta^2+1)^3}\right)^n\prod_{i=1}^{n}\left(\theta^5+3\theta^3+3\theta+x_i\right)\exp\left(-\theta\sum_{i=1}^{n}x_i\right).$$

Step 2: Log-likelihood function

Due to the mathematical complexity involved in directly differentiating the likelihood function, the log-likelihood function is employed to facilitate the differentiation process. The log-likelihood function is given by

$$\log L(\theta;\mathbf{x})=2n\log\theta-3n\log(\theta^2+1)+\sum_{i=1}^{n}\log(\theta^5+3\theta^3+3\theta+x_i)-\theta\sum_{i=1}^{n}x_i.$$

Step 3: Score

To determine the value of θ that maximizes the log-likelihood function, the derivative of $\log L(\theta;x_i)$ with respect to θ is computed. This yields the score function $S(\theta;x_i)$, which is expressed as follows:

$$S(\theta;\mathbf{x})=\frac{\partial}{\partial\theta}\log L(\theta;\mathbf{x})=\frac{2n}{\theta}-\frac{6n\theta}{\theta^2+1}+\sum_{i=1}^{n}\frac{5\theta^4+9\theta^2+3}{\theta^5+3\theta^3+3\theta+x_i}-\sum_{i=1}^{n}x_i.$$

Step 4: MLE

The value of θ is obtained by solving the score equation $S(\theta;\mathbf{x}) = 0$.

This process identifies the critical points that may correspond to potential maxima of the log-likelihood function. Given the absence of a closed-form solution for the maximum likelihood (ML) estimator of the parameter, numerical iterative methods are employed to address the resulting non-linear optimization problem (Nwry et al., 2021). In this study, ML estimation was carried out using the Newton–Raphson method in R. For numerical stability, the iteration was initialized at $1/\bar{x}$, the reciprocal of the sample mean, which provides a natural scale-based starting value for positive lifetime data. The computations were implemented using the `maxLik` package (Henningsen & Toomet, 2011; Myung, 2003).

2.3. Confidence intervals

This section presents four approaches for constructing confidence intervals (CIs) for the parameter of the Double XShanker distribution: the likelihood-based CI, the Wald-type CI, the bootstrap-t CI, and the bias-corrected and accelerated (BCa) bootstrap CI.

Likelihood-based Confidence Interval

The likelihood-based CI is a fundamental statistical technique for parameter estimation. This approach is based on the likelihood function, which measures the probability of observing the given data under specific parameter values within a statistical model. The key idea is to determine a range of parameter values that are most consistent with the observed data, while ensuring that this range corresponds to a specified confidence level. This is typically accomplished by maximizing the log-likelihood function with respect to the parameter of interest and then identifying the parameter values that fall within the confidence region defined by the likelihood ratio criterion.

The ML estimator, denoted by $\hat{\theta}$, is obtained by solving the score equation:

$$S(\theta;\mathbf{x}) = 0.$$

The ML estimator represents the parameter value that is most consistent with the observed data under the assumed statistical model. Once the ML estimate is obtained, a likelihood-based confidence interval can be constructed around it. This method relies on the likelihood ratio, which is defined as:

$$\lambda(\theta) = \frac{L(\theta;\mathbf{x})}{L(\hat{\theta};\mathbf{x})}.$$

Under regularity conditions, Wilks' theorem states that the statistic $-2\log\lambda(\theta)$ asymptotically follows a chi-square distribution, with degrees of freedom equal to the number of parameters estimated—typically one in the context of this study (Severini, 2000; Casella & Berger, 2002; Kummaraka & Srisuradetchai, 2023). The likelihood-ratio interval is defined by the set of θ values that are not rejected by the likelihood-ratio test at level α:

$$-2\log\left(\frac{L(\theta;\mathbf{x})}{L(\hat{\theta};\mathbf{x})}\right) \leq \chi^2_{1-\alpha,1}.$$

For the Double XShanker model, this set can be written explicitly as

$$\left\{\theta : -2\log\left[\left(\frac{\theta}{\hat{\theta}}\right)^{2n}\left(\frac{\hat{\theta}^2+1}{\theta^2+1}\right)^{3n}\prod_{i=1}^{n}\frac{\theta^5+3\theta^3+3\theta+x_i}{\hat{\theta}^5+3\hat{\theta}^3+3\hat{\theta}+x_i} \times \exp\left\{-(\theta-\hat{\theta})\sum_{i=1}^{n}x_i\right\}\right] \leq \chi^2_{1-\alpha,1}\right\},$$

where $\chi^2_{1-\alpha,1}$ is the $(1-\alpha)$ quantile of the chi-square distribution with one degree of freedom. In this study, the ML estimator $\hat{\theta}$ is obtained by numerically solving the score equation $S(\theta;\mathbf{x}) = 0$ using the Newton–Raphson method, as implemented in the `maxLik` package (Henningsen and Toomet, 2011) in R. The score equation is given by

$$S(\theta;\mathbf{x}) = \frac{2n}{\theta} - \frac{6n\theta}{\theta^2+1} + \sum_{i=1}^{n}\frac{5\theta^4+9\theta^2+3}{\theta^5+3\theta^3+3\theta+x_i} - \sum_{i=1}^{n}x_i = 0.$$

Figure 3 illustrates the plot of the negative twice log-likelihood ratio $-2\log\lambda(\theta)$ versus θ (solid blue line), the ML estimate (dashed red line), and the 95% likelihood-based CI (solid green line) for a simulated sample of size 50 from the Double XShanker distribution with $\theta = 1$.

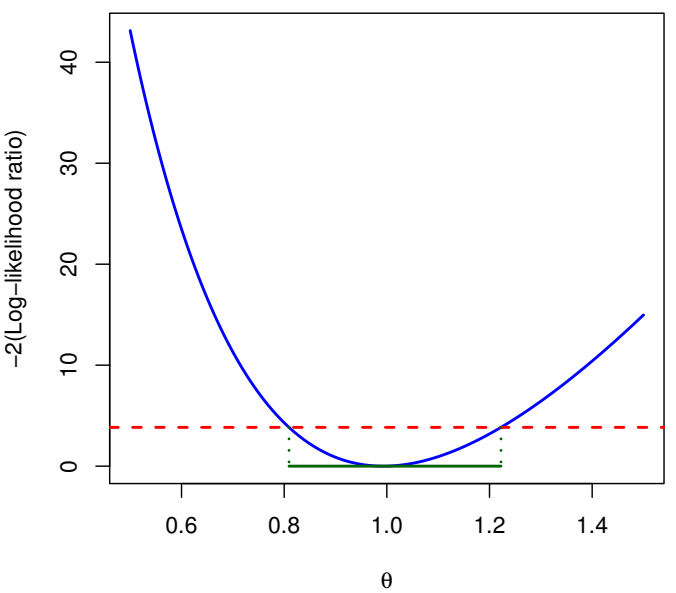

Figure 3. The plot of $-2\log\lambda(\theta)$ versus θ.

Wald-type Confidence Interval

The Wald-type CI for estimating the parameter of the Double XShanker distribution is derived from the ML estimator, denoted by $\hat{\theta}$. This CI is constructed using a second-order Taylor series expansion of the log-likelihood function $\log L(\theta;x)$ around $\hat{\theta}$. The approximation relies on the Wald statistic, which is based on the second derivative of the log-likelihood function, as the first derivative evaluates to zero at the ML estimate (Srisuradetchai & Tonprasongrat, 2022):

$$\log L(\theta;x) \approx \log L(\hat{\theta};x) - \frac{1}{2} I(\hat{\theta})(\theta-\hat{\theta})^2,$$

where $I(\hat{\theta})$ is the observed Fisher information evaluated at $\hat{\theta}$. Under asymptotic conditions, where the sample size is sufficiently large, the Wald statistic provides an effective quadratic approximation to the likelihood ratio test (LRT) statistic (Pawitan, 2001). For the Double XShanker distribution, the first and second derivatives of the log-likelihood function are derived as follows:

$$\frac{\partial}{\partial\theta}\log L(\theta;x_i) = \frac{2n}{\theta} - \frac{6n\theta}{\theta^2+1} + \sum_{i=1}^{n}\left(\frac{5\theta^4+9\theta^2+3}{\theta^5+3\theta^3+3\theta+x_i}\right) - \sum_{i=1}^{n} x_i$$

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log L(\theta;\mathbf{x}) = &-\frac{2n}{\theta^2} - \frac{6n}{\theta^2+1} + \frac{12n\theta^2}{(\theta^2+1)^2} \\ &+ \sum_{i=1}^{n}\frac{20\theta^3+18\theta}{\theta^5+3\theta^3+3\theta+x_i} - \sum_{i=1}^{n}\frac{(5\theta^4+9\theta^2+3)^2}{(\theta^5+3\theta^3+3\theta+x_i)^2}. \end{aligned} \tag{2}$$

Consequently, the observed Fisher information is estimated as:

$$\hat{I}(\hat{\theta}) = -\frac{\partial^2}{\partial\theta^2}\log L(\theta;x_i)\Big|_{\theta=\hat{\theta}}. \tag{3}$$

The Wald-type CI for θ at a confidence level of $(1-\alpha)100\%$ is given by

$$\hat{\theta} \pm z_{1-(\alpha/2)}\sqrt{\hat{I}^{-1}(\hat{\theta})},$$

where $z_{1-(\alpha/2)}$ denotes the $(1-(\alpha/2))^{\text{th}}$ quantile of the standard normal distribution, and $\hat{I}^{-1}(\hat{\theta})$ is the inverse of the observed Fisher information given in Equation (3).

Bootstrap-*t* Confidence Interval

The bootstrap-*t* CI is a resampling-based method that quantifies parameter uncertainty by integrating the bootstrap procedure with the t-distribution framework. Unlike the conventional percentile bootstrap, the bootstrap-*t* approach accounts for the estimator's variability via its standard error, thereby enhancing accuracy and robustness—particularly in small

samples or when the estimator's sampling distribution deviates from normality (Davison and Hinkley, 1997; Efron & Tibshirani, 1993).

This method is particularly useful when the analytical form of the sampling distribution is complex or unknown. The confidence interval is constructed from the empirical distribution of a standardized statistic—referred to as the bootstrap-t statistic—which serves as an approximation to the pivotal quantity used in parametric inference.

The procedure for constructing a bootstrap-t CI involves the following steps (Panichkitkosolkul & Srisuradetchai, 2022):

Step 1: Initial Estimation

Draw a random sample $X = (X_1, \ldots, X_n)$ from the population. Compute the point estimate $\hat{\theta}$ of the parameter of interest, such as the ML estimate.

Step 2: Bootstrap Sampling

Generate B bootstrap samples, X_b^*, $b = 1, 2, \ldots, B$, by sampling with replacement from the original dataset X.

Step 3: Parameter Estimation for Each Sample

For each bootstrap sample X_b^*, compute the bootstrap replicate of the estimator, denoted as $\hat{\theta}_b^*$.

Step 4: Construct the bootstrap-t Statistic

For each bootstrap replicate, calculate the standardized statistic (bootstrap-t) as:

$$t_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\widehat{\mathrm{SE}}_b^*},$$

where $\widehat{\mathrm{SE}}_b^* = \sqrt{\hat{I}^{-1}(\hat{\theta}_b^*)}$ is the bootstrap estimate of the standard error of $\hat{\theta}_b^*$, often obtained via the observed Fisher information (see Equation (3)), with the parameter substituted by $\hat{\theta}_b^*$.

Step 5: Construct the Empirical Distribution

Repeat Steps 2–4 a total of B times to obtain the empirical distribution of the t_b^* values. This distribution approximates the sampling distribution of the standardized estimator.

Step 6: Quantile Calculation

Determine the lower and upper quantiles $t_{\alpha/2}^*$ and $t_{1-\alpha/2}^*$ corresponding to the $\alpha/2$ and $1-(\alpha/2)$ levels of the empirical bootstrap-t distribution. That is,

$$t_{\alpha/2}^* = \inf\left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{[t_b^* \le t]} \ge \alpha/2 \right\}$$

and

$$t_{1-(\alpha/2)}^* = \inf\left\{ t : \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{[t_b^* \le t]} \ge 1-(\alpha/2) \right\},$$

where $\mathbf{1}_{[\cdot]}$ is the indicator function.

Step 7: Construct the bootstrap-t Confidence Interval

The two-sided $(1-\alpha)100\%$ bootstrap-t confidence interval is given by

$$\left[\hat{\theta} - t^*_{1-\alpha/2}\widehat{\mathrm{SE}}(\hat{\theta}),\ \hat{\theta} - t^*_{\alpha/2}\widehat{\mathrm{SE}}(\hat{\theta})\right],$$

where $\widehat{\mathrm{SE}}(\hat{\theta}) = \sqrt{\hat{I}^{-1}(\hat{\theta})}$ is the standard error of the original estimator $\hat{\theta}$.

Bias-corrected and accelerated (BCa) bootstrap confidence interval

The bias-corrected and accelerated (BCa) bootstrap CI is an advanced extension of the basic bootstrap method, designed to correct for both bias and skewness in the empirical distribution of the estimator. This approach is particularly effective in small-sample scenarios or when the estimator's distribution deviates substantially from normality. The BCa bootstrap method enhances the accuracy of interval estimates by incorporating two key adjustments: (1) a bias-correction factor that accounts for asymmetry in the proportion of bootstrap replicates less than the original estimate, and (2) an acceleration parameter that adjusts for the skewness in the estimator's sampling distribution (Bittmann, 2021; Efron, 1987; Panichkitkosolkul & Srisuradetchai, 2022). Unlike conventional methods that rely on normality or symmetry assumptions, the BCa bootstrap approach is nonparametric and distribution-free, offering improved robustness and broader applicability.

The construction of the BCa bootstrap CI proceeds through the following steps:

Step 1: Initial Estimation

Draw a random sample $X = (X_1, \ldots, X_n)$ from the population. Estimate the parameter of interest $\hat{\theta}$ using an appropriate method, such as maximum likelihood.

Step 2: Bootstrap Resampling

Generate B bootstrap samples X^*_b, $b = 1, 2, \ldots, B$, by sampling with replacement from the original dataset X.

Step 3: Bootstrap Estimation

For each bootstrap sample, compute the bootstrap estimator $\hat{\theta}^*_b$, resulting in a distribution of bootstrap estimates $\left\{\hat{\theta}^*_1, \hat{\theta}^*_2, ..., \hat{\theta}^*_B\right\}$.

Step 4: Bias Correction Factor (z_0)

Compute the proportion $\hat{p}$ of bootstrap estimates that are less than the original estimate $\hat{\theta}$:

$$\hat{p} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{[\hat{\theta}^*_b < \hat{\theta}]}.$$

The bias correction factor z_0 is then obtained as the standard normal quantile:

$$z_0 = \Phi^{-1}(\hat{p}),$$

where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal cumulative distribution function.

Step 5: Acceleration Constant (a)

Estimate the acceleration constant a, which reflects the rate of change of the standard error of the estimator with respect to the data. It is typically obtained using the jackknife method:

$$a = \frac{\sum_{i=1}^{n} \left(\bar{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^3}{6\left[\sum_{i=1}^{n} \left(\bar{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^2\right]^{3/2}},$$

where $\hat{\theta}_{(i)}$ is the leave-one-out estimate, and $\bar{\theta}_{(\cdot)}$ is their average.

Step 6: Adjusted Percentiles

The corrected quantiles α_1 and α_2 for the confidence interval are calculated as:

$$\alpha_1 = \Phi\left(z_0 + \frac{z_0 + z_{\alpha/2}}{1 - a(z_0 + z_{\alpha/2})}\right) \quad \text{and} \quad \alpha_2 = \Phi\left(z_0 + \frac{z_0 + z_{1-(\alpha/2)}}{1 - a(z_0 + z_{1-(\alpha/2)})}\right),$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the $(\alpha/2)^{\text{th}}$ and $(1-(\alpha/2))^{\text{th}}$ quantiles of the standard normal distribution, respectively.

Step 7: BCa Bootstrap Confidence Interval Construction

Finally, the two-sided $(1-\alpha)100\%$ BCa bootstrap CI is constructed using the adjusted percentiles α_1 and α_2 as:

$$\left[\hat{\theta}^*_{(\alpha_1)}, \hat{\theta}^*_{(\alpha_2)}\right],$$

where $\hat{\theta}^*_{(\alpha_1)}$ and $\hat{\theta}^*_{(\alpha_2)}$ are $(\alpha_1)^{\text{th}}$ and $(\alpha_2)^{\text{th}}$ quantiles of sorted bootstrap estimates $\hat{\theta}^*_b$.

3. Simulation studies and results

This study evaluates four two-sided CIs for the parameter of the Double XShanker distribution: likelihood-based, Wald-type, bootstrap-t, and bias-corrected and accelerated (BCa) bootstrap intervals. The performance of these methods was assessed through a Monte Carlo simulation study (Robert & Casella, 2010). For each simulated sample, the maximum likelihood estimate was obtained numerically by solving the score equation. To improve the stability of the numerical optimization, the initial value of θ was chosen as the reciprocal of the sample mean, $1/\bar{x}$, which provides a natural scale-based starting value for positive lifetime data.

The performance of the proposed confidence intervals was evaluated using two standard measures: empirical coverage probability (ECP) and average width (AW). These measures were examined across a range of sample sizes and true parameter values. Specifically, sample sizes $n = 10, 20, 30, 50, 100$, and 200 were used, with true parameter values $\theta = 0.20, 0.30, 0.50, 0.75, 1.00, 1.50, 2.00$, and 2.50. For each combination of n and θ, the simulation was replicated 2,000 times. In the main simulation study, the bootstrap-t and BCa bootstrap intervals were constructed using $B = 1000$ bootstrap resamples.

To assess the sensitivity of the bootstrap-based intervals to the number of bootstrap resamples, an additional analysis was conducted for selected representative cases. The settings $(n,\theta) = (10,0.5),(10,2.5),(50,0.5),(50,2.5),(200,0.5)$, and $(200,2.5)$ were considered, covering small, moderate, and large sample sizes, as well as moderate and relatively large parameter values. For each setting, the bootstrap-t and BCa intervals were recomputed using $B = 1000, 2000$, and 5000 bootstrap resamples. To ensure a fair comparison across bootstrap sizes, 5000 bootstrap resamples were generated once for each Monte Carlo sample, and the results for $B = 1000$ and $B = 2000$ were obtained from nested subsets of these resamples. Since the likelihood-based and Wald-type intervals do not depend on B, they were not included in this sensitivity comparison.

The simulation results, summarized in Table 1 and visualized in Figures 4 and 5, compare the ECPs and AWs of the four 95% two-sided CI methods. Across most parameter values and sample sizes, the likelihood-based and Wald-type intervals maintained coverage close to the nominal level of 0.95. Their performance was especially stable for moderate and large sample sizes.

By contrast, the bootstrap-t and BCa bootstrap intervals showed lower ECPs in small samples, particularly when $n = 10, 20$, or 30. For example, at $\theta = 0.2$ and $n = 10$, the ECPs of the bootstrap-t and BCa bootstrap intervals were 0.8850 and 0.8915, respectively, which are below the nominal level. Similar undercoverage was also observed for other parameter values in small samples. As the sample size increased, however, the coverage of the bootstrap-based intervals improved. For instance, when $\theta = 0.2$ and $n = 200$, the ECPs increased to 0.9445 for the bootstrap-t interval and 0.9500 for the BCa bootstrap interval.

With respect to AW, all methods showed the expected decrease in interval width as the sample size increased. The bootstrap-t and BCa bootstrap intervals were generally narrower than the likelihood-based and Wald-type intervals, especially in small samples. This narrower width was accompanied by lower coverage, indicating a trade-off between interval precision and coverage accuracy.

Overall, the likelihood-based and Wald-type methods provided more reliable coverage in small to moderate samples. The bootstrap-t and BCa bootstrap intervals were often shorter, but their coverage was less stable when n was small. These findings suggest that likelihood-based and Wald-type intervals are preferable when coverage accuracy is the main criterion, whereas bootstrap-based intervals may be useful in larger samples when shorter intervals are desired.

The sensitivity results for the number of bootstrap resamples are presented in Table 2. Increasing B from 1000 to 2000 and 5000 produced only small changes in both ECP and AW. The qualitative conclusions remained unchanged: the bootstrap-t and BCa bootstrap intervals continued to show undercoverage when $n = 10$, while their coverage was closer to the nominal level for $n = 50$ and $n = 200$. These results indicate that the main conclusions based on $B = 1000$ are stable with respect to the number of bootstrap resamples considered in this sensitivity analysis.

Table 1. Empirical coverage probability and average width of the 95% two-sided CIs for the parameter of the Double XShanker distribution

		Empirical Coverage Probability (ECP)				Average Width (AW)			
θ	n	Likelihood	Wald	Boot-t	BCa	Likelihood	Wald	Boot-t	BCa
0.20	10	0.9475	0.9510	0.8850	0.8915	0.1720	0.1717	0.1527	0.1584
	20	0.9415	0.9405	0.9025	0.9065	0.1203	0.1202	0.1121	0.1137
	30	0.9390	0.9370	0.9160	0.9155	0.0973	0.0973	0.0936	0.0942
	50	0.9520	0.9520	0.9400	0.9385	0.0752	0.0752	0.0729	0.0732
	100	0.9485	0.9460	0.9365	0.9360	0.0529	0.0529	0.0519	0.0521
	200	0.9545	0.9525	0.9445	0.9500	0.0374	0.0374	0.0368	0.0369
0.30	10	0.9505	0.9495	0.8930	0.8920	0.2490	0.2479	0.2197	0.2319
	20	0.9530	0.9535	0.9215	0.9310	0.1717	0.1714	0.1606	0.1632
	30	0.9490	0.9440	0.9345	0.9305	0.1398	0.1397	0.1341	0.1357
	50	0.9485	0.9490	0.9355	0.9325	0.1077	0.1076	0.1052	0.1058
	100	0.9410	0.9435	0.9360	0.9350	0.0763	0.0762	0.0747	0.0751
	200	0.9555	0.9545	0.9465	0.9470	0.0538	0.0538	0.0531	0.0533
0.50	10	0.9480	0.9570	0.8860	0.8950	0.4118	0.4049	0.3521	0.3928
	20	0.9525	0.9535	0.9110	0.9155	0.2811	0.2790	0.2603	0.2710
	30	0.9455	0.9460	0.9245	0.9325	0.2263	0.2252	0.2146	0.2203
	50	0.9540	0.9570	0.9385	0.9430	0.1744	0.1739	0.1687	0.1716
	100	0.9555	0.9590	0.9480	0.9485	0.1225	0.1223	0.1205	0.1217
	200	0.9540	0.9555	0.9505	0.9450	0.0865	0.0864	0.0855	0.0861
0.75	10	0.9385	0.9520	0.8955	0.8915	0.7074	0.6842	0.5882	0.6931
	20	0.9440	0.9520	0.9085	0.9160	0.4662	0.4581	0.4206	0.4531
	30	0.9550	0.9585	0.9305	0.9335	0.3695	0.3652	0.3439	0.3605
	50	0.9425	0.9440	0.9295	0.9315	0.2816	0.2796	0.2700	0.2780
	100	0.9465	0.9520	0.9395	0.9415	0.1965	0.1958	0.1916	0.1946
	200	0.9510	0.9510	0.9455	0.9465	0.1376	0.1373	0.1356	0.1371
1.00	10	0.9475	0.9490	0.8930	0.8840	1.0843	1.0517	0.9026	1.0514
	20	0.9505	0.9570	0.9230	0.9155	0.7112	0.6978	0.6411	0.6863
	30	0.9470	0.9550	0.9305	0.9275	0.5622	0.5545	0.5227	0.5469
	50	0.9460	0.9505	0.9280	0.9305	0.4259	0.4222	0.4073	0.4189
	100	0.9445	0.9490	0.9365	0.9405	0.2956	0.2942	0.2872	0.2921
	200	0.9505	0.9535	0.9430	0.9425	0.2076	0.2072	0.2047	0.2070
1.50	10	0.9470	0.9465	0.8975	0.8830	1.8961	1.8741	1.6492	1.8381
	20	0.9485	0.9505	0.9165	0.9055	1.2639	1.2570	1.1681	1.2029
	30	0.9535	0.9570	0.9385	0.9285	1.0018	0.9979	0.9523	0.9659
	50	0.9505	0.9480	0.9310	0.9335	0.7654	0.7636	0.7438	0.7484
	100	0.9435	0.9475	0.9370	0.9405	0.5311	0.5304	0.5211	0.5232
	200	0.9530	0.9560	0.9480	0.9495	0.3752	0.3750	0.3699	0.3720
2.00	10	0.9495	0.9375	0.8915	0.8875	2.6205	2.6064	2.2883	2.5190
	20	0.9470	0.9535	0.9195	0.9085	1.7674	1.7660	1.6602	1.6939
	30	0.9505	0.9470	0.9315	0.9280	1.4236	1.4235	1.3650	1.3699
	50	0.9545	0.9490	0.9380	0.9400	1.0880	1.0883	1.0608	1.0591
	100	0.9470	0.9455	0.9375	0.9380	0.7594	0.7597	0.7483	0.7486
	200	0.9480	0.9475	0.9430	0.9455	0.5395	0.5397	0.5360	0.5367
2.50	10	0.9400	0.9500	0.8935	0.8865	3.4087	3.3907	2.9752	3.2834
	20	0.9430	0.9470	0.9135	0.9155	2.2831	2.2802	2.1321	2.1834
	30	0.9560	0.9525	0.9315	0.9270	1.8347	1.8335	1.7537	1.7665
	50	0.9475	0.9545	0.9315	0.9300	1.3984	1.3980	1.3509	1.3583
	100	0.9415	0.9430	0.9375	0.9340	0.9754	0.9754	0.9583	0.9603
	200	0.9515	0.9505	0.9425	0.9415	0.6847	0.6847	0.6741	0.6763

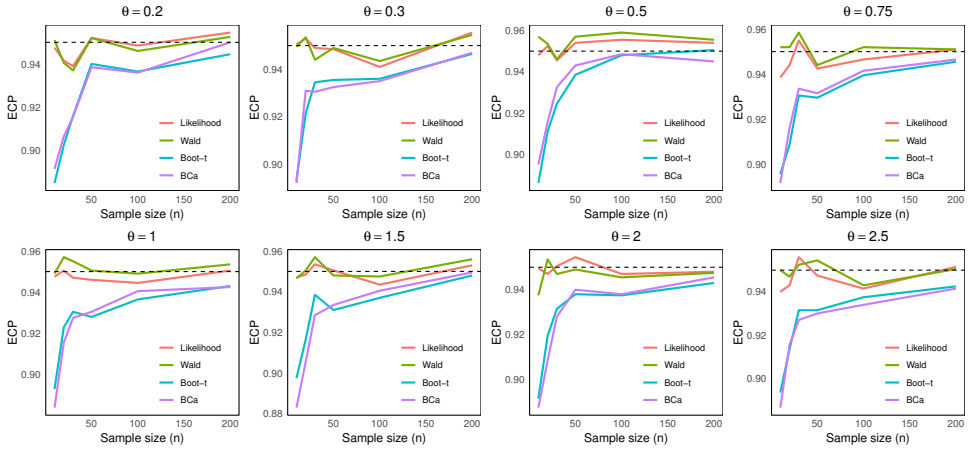

Figure 4. Plots of the ECPs of the CIs for the parameter of the Double XShanker distribution

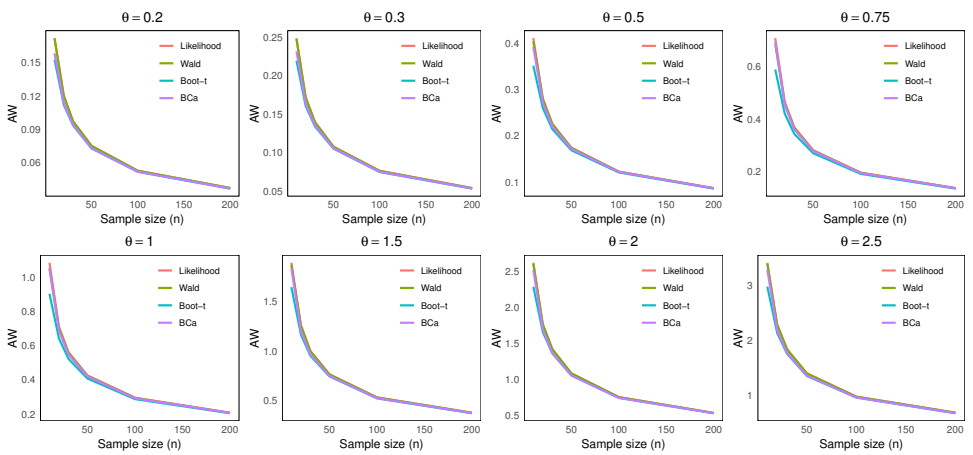

Figure 5. Plots of the AWs of the CIs for the parameter of the Double XShanker distribution

Table 2. Sensitivity of bootstrap-based confidence intervals to the number of bootstrap resamples

			Bootstrap-t		BCa bootstrap	
n	ϑ	B	ECP	AW	ECP	AW
10	0.5	1000	0.8790	0.3530	0.8925	0.3926
		2000	0.8850	0.3539	0.8940	0.3935
		5000	0.8845	0.3545	0.8940	0.3944
	2.5	1000	0.8950	2.9629	0.8825	3.2431
		2000	0.8960	2.9693	0.8840	3.2543
		5000	0.8950	2.9724	0.8835	3.2602
50	0.5	1000	0.9400	0.1689	0.9370	0.1706
		2000	0.9405	0.1693	0.9365	0.1712
		5000	0.9400	0.1695	0.9400	0.1714
	2.5	1000	0.9455	1.3570	0.9425	1.3571
		2000	0.9460	1.3595	0.9460	1.3598
		5000	0.9440	1.3618	0.9455	1.3611
200	0.5	1000	0.9385	0.0853	0.9390	0.0855
		2000	0.9380	0.0855	0.9390	0.0857
		5000	0.9395	0.0857	0.9385	0.0859
	2.5	1000	0.9410	0.6785	0.9405	0.6774
		2000	0.9385	0.6804	0.9420	0.6794
		5000	0.9390	0.6813	0.9415	0.6804

Note. The $B = 1000$ entries provide the baseline within this separate sensitivity analysis. For each Monte Carlo sample, the results for $B = 1000$, $B = 2000$, and $B = 5000$ were obtained from nested bootstrap resamples.

4. Real data application

We apply the fitted Double XShanker model and the four confidence intervals to a rainfall dataset. For model-comparison purposes, the Double XShanker distribution was compared with several one-parameter lifetime distributions commonly used for positively skewed data. The competing models are listed below.

1. Akshaya distribution (Shanker, 2017a):

$$f(x;\theta) = \frac{\theta^4}{\theta^3+3\theta^2+6\theta+6}(1+x)^3 e^{-\theta x},$$

2. Amarendra distribution (Shanker, 2016a):

$$f(x;\theta) = \frac{\theta^4}{\theta^3+\theta^2+2\theta+6}\left(1+x+x^2+x^3\right) e^{-\theta x},$$

3. Devya distribution (Shanker, 2016b):

$$f(x;\theta) = \frac{\theta^5}{\theta^4+\theta^3+2\theta^2+6\theta+24}\left(1+x+x^2+x^3+x^4\right)e^{-\theta x},$$

4. Iwueze distribution (Elechi et al., 2022):

$$f(x;\theta) = \frac{\theta^5}{\theta^4+2\theta^3+6\theta^2+12\theta+24}(1+x+x^2)^2 e^{-\theta x},$$

5. Prakaamy distribution (Shukla, 2018a):

$$f(x;\theta) = \frac{\theta^6}{\theta^5+120}(1+x^5)e^{-\theta x},$$

6. Suja distribution (Shanker, 2017b):

$$f(x;\theta) = \frac{\theta^5}{\theta^4+24}(1+x^4)e^{-\theta x},$$

7. Iwok-Nwikpe distribution (Iwok and Nwikpe, 2021):

$$f(x;\theta) = \frac{\theta^3}{\theta+2}(x^2+x)e^{-\theta x},$$

8. Nwikpe distribution (Nwikpe et al., 2021):

$$f(x;\theta) = \frac{\theta^3}{12(\theta^2+10)}(\theta^3x^5+12)e^{-\theta x},$$

9. Ola distribution (Al-Ta'ani and Gharaibeh, 2023):

$$f(x;\theta) = \frac{\theta^8}{\theta^7+6\theta^4+5040}(x^7+x^3+1)e^{-\theta x},$$

10. Ram Awadh distribution (Shukla, 2018b):

$$f(x;\theta) = \frac{\theta^6}{\theta^6+120}(\theta+x^5)e^{-\theta x}.$$

The dataset consists of 72 March rainfall measurements (in inches) recorded at the Los Angeles Civic Center during the period 1943–2018:

4.55, 2.47, 3.43, 3.66, 0.79, 3.07, 1.40, 0.87, 0.44, 6.14, 0.48, 2.99, 0.56, 1.02, 5.30, 0.31, 0.57, 1.10, 2.78, 1.79, 2.49, 0.53, 2.50, 3.34, 1.49, 2.36, 0.53, 2.70, 3.78, 4.83, 1.81, 1.89, 8.02, 5.85, 4.79, 4.10, 3.54, 8.37, 0.28, 1.29, 5.27, 0.95, 0.26, 0.81, 0.17, 5.92, 7.12, 2.74, 1.86, 6.98, 2.16, 4.06, 1.24, 2.82, 1.17, 0.32, 4.32, 1.47, 2.14, 2.87, 0.05, 0.01, 0.35, 0.48, 3.96, 1.75, 0.54, 1.18, 0.87, 1.60, 0.09, 2.69.

Descriptive statistics for this dataset are summarized in Table 3, while Figure 6 provides visual representations, including a histogram, a box-and-whisker plot, a kernel density plot, and a violin plot, highlighting the positive skewness of the data.

Table 3. Descriptive statistics for the rainfall data at the Los Angeles Civic Center

Sample Size	Minimum	Q1	Median	Mean	Q3	Maximum	St.Dev
72	0.010	0.805	1.875	2.450	3.570	8.370	2.055

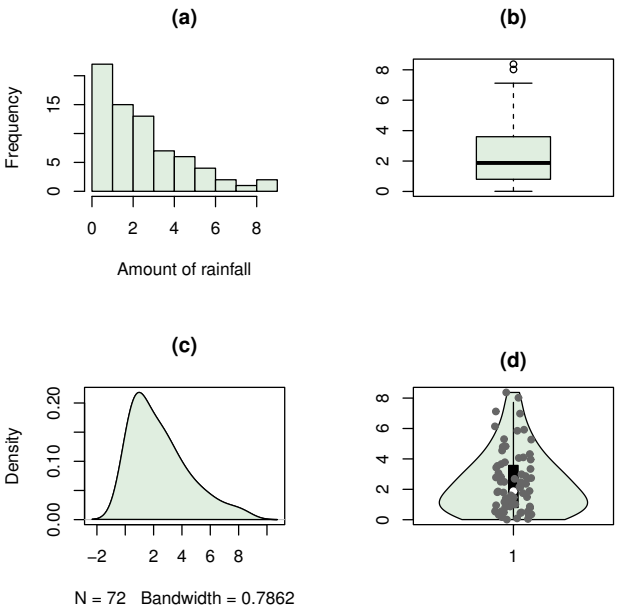

Figure 6. Visual representations of the rainfall data: (a) Histogram, (b) Box-and-whisker plot, (c) Kernel density plot, and (d) Violin plot

A comprehensive evaluation of model performance was conducted using multiple statistical criteria, namely the maximized log-likelihood $(\log(\hat{L}))$, the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC), also known as the Schwarz Information Criterion. These metrics provide a basis for assessing model fit of competing distributions. Specifically, AIC and BIC offer adjusted likelihood-based measures that balance model fit against model complexity, thereby supporting more reliable model selection. The AIC and BIC are formally defined as follows:

$$\text{AIC} = 2k - 2\log(\hat{L}) \quad \text{and} \quad \text{BIC} = k\log(n) - 2\log(\hat{L}),$$

where k represents the number of estimated parameters in the model and $\hat{L}$ denotes the maximized value of the likelihood function for the model. A distribution with lower AIC

and BIC values is generally preferred, as these statistics indicate a better balance between model fit and complexity. Table 4 provides the parameter estimates and goodness-of-fit measures for the dataset under investigation.

Table 4. Comparative analysis of model fit statistics for different distributions applied to rainfall data at the Los Angeles Civic Center

Distribution	Estimate	LogLik	AIC	BIC
Double XShanker	0.5780	-135.681	**273.362**	**275.639**
Akshaya	1.1897	-141.282	284.564	286.841
Amarendra	1.2446	-139.365	280.730	283.006
Devya	1.5727	-142.642	287.283	289.560
Iwueze	1.5155	-144.230	290.460	292.737
Prakaamy	2.0103	-147.667	297.333	299.610
Suja	1.6531	-142.474	286.949	289.225
Iwok-Nwikpe	1.0811	-149.447	300.894	303.170
Nwikpe	1.8141	-143.623	289.245	291.522
Ola	2.7082	-158.132	318.263	320.540
Ram Awadh	1.9007	-143.549	289.097	291.374

Note. Bold values indicate the best-fitting model based on the lowest AIC and BIC values.

Table 4 presents the log-likelihood values, AIC (Akaike, 1974), and BIC (Schwarz, 1978) for the Double XShanker distribution and ten alternative distributions fitted to the rainfall data from the Los Angeles Civic Center. Among these distributions, the Double XShanker distribution provided the best overall fit, indicated by the lowest AIC and BIC values, suggesting strong compatibility with the empirical data. The maximum likelihood estimate of the Double XShanker parameter was 0.5780.

To further assess uncertainty associated with the parameter estimate, Table 5 presents the 95% two-sided CIs for the Double XShanker distribution parameter, constructed using four methods: likelihood-based, Wald-type, bootstrap-t, and BCa bootstrap. The likelihood-based method produced a CI of (0.4989, 0.6664), with a width of 0.1675. The Wald-type method yielded a similar interval (0.4945, 0.6615) with a slightly narrower width of 0.1670. In contrast, the bootstrap-t and BCa bootstrap methods produced narrower intervals—(0.5066, 0.6610) and (0.5045, 0.6598), respectively—in agreement with the simulation results. Compared to a similar simulation scenario with $\theta = 0.5$ and $n = 50$ (close to our dataset size of $n = 72$), the likelihood-based and Wald-type intervals had widths of 0.1744 and 0.1739, respectively, whereas the bootstrap-t and BCa bootstrap intervals were narrower at 0.1687 and 0.1716. These results support the use of the Double XShanker distribution for the rainfall data and show consistency between the simulation and empirical findings in a moderate sample-size setting.

Table 5. The 95% two-sided CIs and corresponding widths based on rainfall data at the Los Angeles Civic Center

CI Method	Confidence Interval	Width
Likelihood-based	(0.4989, 0.6664)	0.1675
Wald-type	(0.4945, 0.6615)	0.1670
bootstrap-t	(0.5066, 0.6610)	0.1544
BCa bootstrap	(0.5045, 0.6598)	0.1553

5. Conclusions and discussions

We compared four 95% two-sided confidence intervals for the Double XShanker parameter θ: likelihood-based, Wald-type, bootstrap-t, and BCa bootstrap intervals. An explicit expression for the observed Fisher information was derived, allowing direct computation of the Wald interval once the maximum likelihood estimate is obtained. Performance was evaluated through Monte Carlo simulations over a range of sample sizes and parameter values, using empirical coverage probability (ECP) and average width (AW) as criteria.

The simulation results show that the likelihood-based and Wald-type intervals remain close to the nominal 0.95 level across most settings considered. In contrast, the bootstrap-t and BCa bootstrap intervals are typically narrower but exhibit undercoverage in small samples. Their coverage improves as the sample size increases, approaching the nominal level in moderate to large samples. The additional sensitivity analysis using $B = 1000, 2000$, and 5000 bootstrap resamples showed only small changes in ECP and AW, indicating that the main conclusions are stable with respect to the number of bootstrap resamples considered.

The methods were further illustrated using rainfall data from the Los Angeles Civic Center. Based on the maximized log-likelihood, AIC (Akaike, 1974), and BIC (Schwarz, 1978), the Double XShanker distribution provided the best fit among the competing one-parameter lifetime models considered. The interval estimates from the data analysis were consistent with the simulation findings: bootstrap-based intervals were slightly narrower, whereas likelihood-based and Wald-type intervals aligned more closely with nominal coverage behavior.

Bootstrap procedures, particularly the bootstrap-t and BCa methods, require repeated estimation and, for BCa bootstrap, additional jackknife calculations. This increases computational cost in large-scale resampling settings. In practice, parallel computation and a careful choice of the number of resamples can reduce this burden. In R, packages such as `boot` (Canty & Ripley, 2025) and `bootstrap` (Kostyshak, 2019) provide standard implementations.

Overall, for small to moderate samples, the likelihood-ratio and Wald intervals are preferable when coverage accuracy is the primary concern. For larger samples, bootstrap-t and BCa bootstrap intervals may be considered when narrower intervals are desired and computational resources are available. When computation is limited, the Wald-type interval remains attractive due to its simple form and low cost.

Acknowledgements

We thank the reviewers for helpful comments and suggestions. This study was supported by the Thammasat University Research Unit in Mathematical Sciences and Applications.

References

Akaike, H., (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), pp. 716–723.

Al-Ta'ani, O., Gharaibeh, M. M., (2023). Ola distribution: A new one parameter model with applications to engineering and Covid-19 data. *Applied Mathematics & Information Sciences*, 17(2), pp. 243–252.

Bittmann, F., (2021). *Bootstrapping: An integrated approach with Python and Stata*. Munich: De Gruyter Oldenbourg. Retrieved from `https://doi.org/10.1515/9783110693348`.

Canty, A., Ripley, B., (2025). *boot: Bootstrap Functions* (R package version 1.3-32) [Computer software]. Retrieved from `https://cran.r-project.org/package=boot`. `https://doi.org/10.32614/CRAN.package.boot`.

Casella, G., Berger, R. L., (2002). *Statistical Inference* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Davison, A. C., Hinkley, D. V., (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Efron, B., (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), pp. 171–185.

Efron, B., Tibshirani, R. J., (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC.

Elechi, O., Okereke, E., Chukwudi, I., Chizoba, K. and Wale, O., (2022). Iwueze's distribution and its application. *Journal of Applied Mathematics and Physics*, 10, pp. 3783–3803.

Etaga, H. O., Nwankwo, M. P., Oramulu, D. O. and Obulezi, O. J., (2023). An improved XShanker distribution with applications to rainfall and vinyl chloride data. *Scholars Journal of Engineering and Technology*, 11(9), pp. 212–224. `https://doi.org/10.36347/sjet.2023.v11i09.005`.

Henningsen, A., Toomet, O., (2011). MaxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3), pp. 443–458.

Iwok, I. A., Nwikpe, B. J., (2021). The Iwok-Nwikpe distribution: statistical properties and its application. *Asian Journal of Probability and Statistics*, 15(1), pp. 35–45. https://doi.org/10.9734/ajpas/2021/v15i130347.

Kostyshak, S., (2019). *bootstrap: Functions for the Book "An Introduction to the Bootstrap"* (R package version 2019.6) [Computer software]. Retrieved from https://cran.r-project.org/package=bootstrap. https://doi.org/10.32614/CRAN.package.bootstrap.

Kummaraka, U., Srisuradetchai, P. (2023). Interval estimation of the dependence parameter in bivariate Clayton copulas. *Emerging Science Journal*, 7(5), pp. 1478–1490.

Messaadia, H., Zeghdoudi, H., (2018). Zeghdoudi distribution and its applications. *International Journal of Computing Science and Mathematics*, 9(1), pp. 58–65. https://doi.org/10.1504/IJCSM.2018.090722.

Myung, I. J., (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1), pp. 90–100.

Nadarajah, S., Kotz, S., (2006). The beta exponential distribution. *Reliability Engineering & System Safety*, 91(6), pp. 689–697. https://doi.org/10.1016/j.ress.2005.05.008.

Niyomdecha, A., Srisuradetchai, P. and Tulyanitikul, B., (2023). Gamma zero-truncated Poisson distribution with the minimum compounded function. *Thailand Statistician*, 21(4), pp. 863–886.

Nwikpe, B. J., Isaac, D. E. and Amos, E., (2021). Nwikpe probability distribution: statistical properties and goodness of fit. *Asian Journal of Probability and Statistics*, 11(1), pp. 52–61.

Nwry, A. W., Kareem, H. M., Ibrahim, R. B. and Mohammed, S. M., (2021). Comparison between bisection, Newton, and secant methods for determining the root of the non-linear equation using MATLAB. *Turkish Journal of Computer and Mathematics Education*, 12(14), pp. 1115–1122.

Panichkitkosolkul, W., Srisuradetchai, P., (2022). Bootstrap confidence intervals for the parameter of zero-truncated Poisson-Ishita distribution. *Thailand Statistician*, 20(4), pp. 918–927.

Pawitan, Y., (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Oxford University Press.

Robert, C. P., Casella, G. (2010). *Introducing Monte Carlo Methods with R*. New York: Springer.

Schwarz, G., (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), pp. 461–464.

Severini, T. A., (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Shanker, R., (2015). Shanker distribution and its applications. *International Journal of Statistics and Applications*, 5(6), pp. 338–348.

Shanker, R., (2016a). Amarendra distribution and its applications. *American Journal of Mathematics and Statistics*, 6(1), pp. 44–56.

Shanker, R., (2016b). Devya distribution and its applications. *International Journal of Statistics and Applications*, 6(4), pp. 189–202.

Shanker, R., (2017a). Akshaya distribution and its application. *American Journal of Mathematics and Statistics*, 7(2), pp. 51–59.

Shanker, R., (2017b). Suja distribution and its application. *International Journal of Probability and Statistics*, 6(2), pp. 11–19.

Shukla, K. K., (2018a). Prakaamy distribution with properties and applications. *Journal of Applied Quantitative Methods*, 13(3), pp. 30–38.

Shukla, K. K., (2018b). Ram Awadh distribution with properties and applications. *Biometrics & Biostatistics International Journal*, 7(6), pp. 515–523.

Srisuradetchai, P., Dangsupa, K., (2022). On interval estimation of the geometric parameter in a zero-inflated geometric distribution. *Thailand Statistician*, 21(1), pp. 93–109.

Srisuradetchai, P., Tonprasongrat, K., (2022). On interval estimation of the Poisson parameter in a zero-inflated Poisson distribution. *Thailand Statistician*, 20(2), pp. 357–371.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 69–86, https://doi.org/10.59139/stattrans-2026-028
Received – 06.11.2024; accepted – 19.02.2026

Inferential aspects of a M/M/1 queuing system using bivariate prior

Namrata Pareek[1], **Amit Choudhury**[2]

Abstract

This paper presents a Bayesian framework for estimating the performance measures of a single-server Markovian queue. A bivariate prior is employed to naturally incorporate the system's ergodicity condition along with the traditional likelihood function, which fundamentally differentiates this work from that of the existing literature. The Bayesian estimators of the performance measures are presented in the study, and we investigate standard errors and credible intervals of the Bayesian estimates. In addition, the predictive distributions for waiting times and the number of customers in the system at departure epoch were derived, offering deeper insights into the system dynamics. We also conducted a simulation study to assess the efficacy and reliability of the proposed methodology. Finally, a real-world scenario was presented to demonstrate the practical applicability of the developed algorithms.

Key words: Bayesian inference, bivariate prior, credible region, Markovian queuing system, performance measures, single server queuing model, standard error, traffic intensity.

Mathematics Subject Classification: 90B22, 60K25.

1. Introduction and motivation

Researchers developing probabilistic models for a wide range of queuing scenarios and deriving precise formulae to describe system performance in classical settings remark on the evolution of queueing theory over the years. However, a real-world application of theoretical models requires proper parameter estimation and data alignment to account for system complexity. These issues highlight fundamental concerns about whether statistical estimators can maintain their reliability and effectiveness in the face of real-world uncertainty and variability.

[1] Department of Statistics, Gauhati University, Guwahati, Assam-781014, India & Department of Statistics, Handique Girls' College, Guwahati, Assam-781001, India. E-mail: imp.info.namrata@gmail.com. ORCID: https://orcid.org/0009-0002-4217-8328.

[2] Department of Statistics, Gauhati University, Guwahati, Assam-781014, India. E-mail: achoudhury@rediffmail.com. ORCID: https://orcid.org/0000-0001-8777-8883.

The area of statistical inference for queueing systems features two main methodological frameworks: the frequentist approach and the Bayesian approach. Early research predominantly employed frequentist methods. Pioneering works by Clarke (1957) and Lilliefors (1966) introduced estimation procedures for the M/M/1 model. Subsequent contributions by Scruben and Kulkarni (1982) and Basawa and Prabhu (1988) revealed notable challenges, including estimator instability and the lack of standard errors for certain performance measures. To address these limitations, researchers such as Zheng and Seila (2000) and Dutta and Choudhury (2020) presented bounding strategies. Choudhury and Basak (2018a) derived the Maximum Likelihood Estimates (MLEs) of traffic intensity (ρ) for an M/M/1 queue. Subsequently, Choudhury and Basak (2018b) obtained MLEs of the average number of customers in the system (L) and the average number of customers in the queue (L_q). Despite these methodological advances, the frequentists often face difficulties in incorporating prior knowledge and accounting for model uncertainty, leading to a growing interest in Bayesian alternatives.

Bayesian inference provides more flexibility by treating parameters as random variables and integrating prior information. McGrath *et al.* (1987a, 1987b) were among the first to apply Bayesian techniques to queueing models, highlighting their advantages over their classical counterparts. Building on this foundation, Armero and Bayarri (1994a, 1994b) introduced informative priors, such as the Gauss-Hypergeometric prior, to overcome issues like the non-existence of moments and to enhance predictive accuracy. Sharma and Kumar (1999) discussed the issue of statistical inference both from a frequentist and a Bayesian perspective. Their work consisted of constructing the UMP critical region for testing the hypothesis on performance measures using a randomized testing procedure. Choudhury and Borthakur (2008) further demonstrated the relative robustness of Bayesian estimators compared to frequentist counterparts, while Choudhury and Mukherjee (2013) evaluated different priors and loss functions to refine the estimation process. Jose and Manoharan (2014) incorporated the ergodicity condition (arrival rate is less than service rate, that is, $\lambda < \mu$) through a bivariate prior, improving model fidelity. Cruz *et al.* (2017) used Sampling/Importance Resampling (SIR) to estimate ρ in M/M/s and M/M/s/K queues. Almeida and Cruz (2017) worked with Jeffreys prior in an M/M/1 queue. Later, Jose and Manoharan (2018) proposed an alternate bivariate prior to further strengthen estimation stability. Suyama *et al.* (2018) suggested estimators for λ, μ, and ρ in M/M/s queues using samples of system size and sojourn times at arbitrary points. Cruz *et al.* (2018) investigated ρ in M/M/1/K queues using maximum likelihood, Bayesian, and bootstrap techniques. Basak and Choudhury (2021) developed Bayesian estimators based on observed queue lengths. Bura and Sharma (2024) addressed queueing models with customer balking. Basak and Choudhury (2024) addressed estimation in an M/M/1/K system under known traffic intensity. Das *et al.* (2025) investigated a novel M/M/1 queueing model with balking, and consequently estimated traffic intensity and performance measures using both

classical and Bayesian methods. Recently, Kushvaha *et al.* (2025) refined Bayesian inference under a variety of prior structures and loss functions, emphasizing the benefits of posterior risk analysis. Basak and Choudhury (2025) estimated λ and μ in an M/M/1 queue using McKay's bivariate gamma prior and compared them with the corresponding MLEs.

Motivated by these advancements, the present article adopts a Bayesian approach for parameter estimation in the M/M/1 queueing model, employing the notion of stochasticity in parameters. In an M/M/1 queue, time-variability refers to the evolution of parameters, such as the arrival rate or service rate, over time, and this evolution may be either deterministic or stochastic. On the other hand, stochasticity arises within Bayesian inference, where parameters are modelled as random variables. This is not because they are inherently random in reality. It is done to allow the incorporation of prior information and uncertainty about their true but fixed values. Consequently, Bayesian inference remains valid for both fixed and time-varying parameters. However, the current work adopts a stochastic treatment through the use of a bivariate prior, which serves to capture uncertainty as well as potential dependence between the arrival and service rates, without presuming that the parameters themselves are physically random processes.

The novelty and main contributions of this article are the use of a bivariate prior along with a simple and analytically convenient likelihood function, which assumes exponential distributions for both interarrival and service times. This bivariate prior explicitly incorporates the natural ergodicity condition ($\lambda < \mu$). It is worth noting that the use of such priors has been relatively underexplored in the existing literature. This is the principal research gap that the current work seeks to address. Also, explicit expressions for the predictive distributions of waiting times have not yet been thoroughly explored. Owing to the need in the existing literature, we address this limitation by formulating such predictive distributions within the M/M/1 framework. This deepens the theoretical understanding of the system dynamics and extends the scope of Bayesian analysis in queuing models.

The paper has been structured as follows. Section 2 includes a brief description of the model. Section 3 formulates Bayesian estimators for the performance measures and derives the predictive distributions for the departure epoch and waiting times. Section 4 presents a simulation study to exemplify the proposed methodology. Section 5 contains an illustrative numerical example. Section 6 outlines the concluding remarks.

2. Model description

In this study, we consider an M/M/1 queuing model, where customer arrivals follow a Poisson process with mean $\frac{1}{\lambda}$, and the server provides exponential service with

mean $\frac{1}{\mu}$. The system operates on a First-Come First-Served (FCFS) basis, and the queue is assumed to be in equilibrium. This implies that the traffic intensity ($\rho = \frac{\lambda}{\mu}$) must be less than one ($\rho < 1$), ensuring system stability. If this condition is not initially met, operational adjustments must be made to avoid an unstable system where the queue size would grow infinitely.

For an M/M/1 system, the number of customers in the system at departure epoch (M) is given by (Medhi, 2003):

$$P(M = m) = \begin{cases} \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^m, m = 0,1,2,\dots \\ 0, otherwise \end{cases}. \quad (1)$$

The key performance measures for the $M/M/1$ queue, in terms of λ and μ, are:

- Average number of customers in the system, $L = \frac{\lambda}{\mu-\lambda}$;
- Average number of customers in the queue, $L_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$;
- Average waiting time of a customer in the system is $W = \frac{1}{(\mu-\lambda)}$;
- Average waiting time of a customer in the queue is $W_q = \frac{\lambda}{\mu(\mu-\lambda)}$;
- The probability that the number of customers in the system is greater than or equal to s, that is, $P(N \geq s) = \left(\frac{\lambda}{\mu}\right)^s$.

The distribution of waiting time in the system, W, (Gross and Harris, 2008), is given by

$$f(t) = (\mu - \lambda)e^{-(\mu-\lambda)t}, t > 0. \quad (2)$$

The distribution of waiting time in the queue, W_q, (Gross and Harris, 2008), is

$$f(t) = \{1 - \left(\frac{\lambda}{\mu}\right), \ t = 0 \ \frac{\lambda}{\mu}(\mu - \lambda)e^{-\lambda\mu(\mu-\lambda)t}, t > 0. \quad (3)$$

These expressions form the basis for evaluating system performances, and they will be used in the subsequent sections to derive Bayes estimators for key performance measures and the predictive distributions.

3. Bayesian inference and predictive distributions

This section applies Bayesian inferential technique to estimate the performance measures of an M/M/1 queueing system. We proceed by adopting a joint prior of λ and μ (Jose and Manoharan, 2014, p. 72) defined as

$$\pi(\lambda, \mu) = \alpha\beta e^{-\{(\alpha-\beta)\lambda+\beta\mu\}} \quad (4)$$

where $0 < \lambda < \mu < \infty$ and $\alpha, \beta > 0$.

Unlike independent priors commonly used in the literature, this bivariate form inherently satisfies the equilibrium constraint, making it more appropriate for modelling queueing systems and ensuring valid posterior inference.

The likelihood function is constructed from the observed data of the M/M/1 queueing system. Let $x_1, x_2, \dots, x_{n_s}$ be a random sample of service times of size n_s and $y_1, y_2, \dots, y_{n_a}$ be a random sample of interarrival times of size n_a. Since interarrival and service times follow exponential distributions with rate parameters λ and μ, respectively, the likelihood function for (λ, μ) is given by:

$$L(\lambda, \mu) = \lambda^{n_a} e^{-\lambda t_a} \mu^{n_s} e^{-\mu t_s} \tag{5}$$

where $t_a = \sum y_j$ and $t_s = \sum x_i$ are the total observed interarrival and service times, respectively. This form is computationally convenient and has been utilized in similar Bayesian analyses to ensure mathematical feasibility (e.g. Armero and Bayarri, 1994b).

Combining the prior and the likelihood, the posterior joint distribution of λ and μ is then given by

$$\Pi(\lambda, \mu | data) = \frac{\lambda^{n_a} e^{-(\alpha-\beta+t_a)\lambda} \mu^{n_s} e^{-(\beta+t_s)\mu}}{\left[\left(\frac{n_s!}{(\beta+t_s)^{n_s+1}}\right)\left(\frac{n_a!}{(\alpha+t_a+t_s)^{n_a+1}}\right)\right]\left[\sum_{i=0}^{n_s} \binom{n_a+i}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^i\right]}. \tag{6}$$

This posterior distribution (6) will serve as the basis for deriving Bayes estimators of the performance measures considered in this study.

Before proceeding with Bayesian estimation, we perform a sensitivity analysis with respect to the hyperparameters α and β of the bivariate prior in (4) and evaluate it against independent Gamma priors for $\lambda = 2$ and $\mu = 3$. Table 1 presents posterior means and mean squared errors (MSEs) from one hundred simulated samples. The posterior means of both λ and μ consistently approximate their respective true values across all prior specifications. Furthermore, the MSEs remain stable, showing minimal sensitivity to the choice of the prior. Together, these findings support that the inferential conclusions are robust and not overly sensitive to reasonable changes in prior specification, hence validating the proposed Bayesian framework.

Table 1. Sensitivity analysis results with posterior summaries under different prior parameters

Prior	Average posterior mean of (λ)	MSE (λ)	Average posterior mean of (μ)	MSE (μ)
Bivar ($\alpha = 1, \beta = 0.1$)	2.029	0.0397	3.033	0.0753
Bivar ($\alpha = 1, \beta = 1$)	2.064	0.0457	2.956	0.0672
Bivar ($\alpha = 5, \beta = 1$)	2.063	0.0458	2.953	0.0661
Gamma (1,1)	1.982	0.0366	2.952	0.0685
Gamma (2,1)	2.005	0.0367	2.981	0.0690

3.1. Derivation of Bayes estimators

For Bayesian estimation, we employ the Squared Error Loss Function (SELF). Under SELF, the Bayes estimator for a parameter is its posterior mean. We thus formulate a unified expression for the estimators as follows:

$$E\left\{\frac{\lambda^u}{\mu^v(\mu-\lambda)^w}\right\} = \int_0^\infty \int_\lambda^\infty \frac{\lambda^u}{\mu^v(\mu-\lambda)^w} \Pi\,(\lambda,\mu|data)d\mu d\lambda = \int_0^\infty \int_\lambda^\infty \frac{\lambda^u}{\mu^{v+w}\left(1-\frac{\lambda}{\mu}\right)^w} \Pi\,(\lambda,\mu|data)d\mu d\lambda \tag{7}$$

where u, v, and w are non-negative integers.

Applying the negative binomial series expansion and performing algebraic simplification results in the simplified expression for the estimators.

$$E\left\{\frac{\lambda^u}{\mu^v(\mu-\lambda)^w}\right\} = \frac{\sum_{k=0}^{\infty}\left[\binom{w+k-1}{k}\int_0^\infty \lambda^{n_a+u+k} e^{-(\alpha-\beta+t_a)\lambda}\left\{\int_\lambda^\infty \mu^{n_s-v-w-k} e^{-(\beta+t_s)\mu} d\mu\right\} d\lambda\right]}{\left[\left(\frac{n_s!}{(\beta+t_s)^{n_s+1}}\right)\left(\frac{n_a!}{(\alpha+t_a+t_s)^{n_a+1}}\right)\right]\left[\sum_{i=0}^{n_s}\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^i\right]}. \tag{8}$$

Assuming n_s and, consequently, $n_s - v - w$ as sufficiently large, the inner integral $\int_\lambda^\infty \mu^{n_s-v-w-k} e^{-(\beta+t_s)\mu} d\mu$ can be approximated using the incomplete gamma function, provided $k < n_s - v - w - 1$. This leads to the following generalized formula:

$$E\left\{\frac{\lambda^u}{\mu^v(\mu-\lambda)^w}\right\} \approx \frac{\sum_{k=0}^{n_s-v-w-1}\left[\binom{w+k-1}{k}\left\{\frac{\binom{n_a+u+k}{u+k}}{\binom{n_s}{v+w+k}}\frac{(u+k)!}{(v+w+k)!}\right\}\left\{\sum_{i=0}^{n_s-v-w-k}\binom{n_a+u+k+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k}\right\}\right]}{\left[\sum_{i=0}^{n_s}\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^i\right]}. \tag{9}$$

Assigning particular values to u, v, and w in (9) yields explicit Bayes estimators for key queueing characteristics, as demonstrated below.

Bayes Estimator of the Traffic intensity (ρ) *and its standard error*

The Bayes estimator ρ^* of ρ is obtained by putting $u = 1, v = 0, w = 0$ in (9).

$$\rho^* = \frac{(n_a+1)}{n_s}\frac{\sum_0^{n_s-1}\binom{n_a+i+1}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+1}}{\sum_0^{n_s}\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^i}. \tag{10}$$

The corresponding standard error is obtained by substituting $u = 2$, $v = 0$, $w = 0$ in (9).

$$S.E.(\rho^*) = \sqrt{\left[\frac{(n_a+2)(n_a+1)}{n_s(n_s-1)}\frac{\sum_0^{n_s-2}\binom{n_a+i+2}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+2}}{\sum_0^{n_s}\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^i}\right] - (\rho^*)^2}. \tag{11}$$

Bayes Estimator of the average number of customers in the system (L) and its standard error.

Putting $u = 1, v = 0, w = 1$ in (9), the Bayes estimator L^* of L is

$$L^* = \frac{\sum_{k=0}^{n_s-2} \frac{\binom{n_a+k+1}{k+1}}{\binom{n_s}{k+1}} \sum_{i=0}^{n_s-k-1} \binom{n_a+i+k+1}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k+1}}{\sum_0^{n_s} \binom{n_a+i}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}}. \quad (12)$$

The standard error of this estimator is obtained by substituting $u = 2, v = 0, w = 2$ in (9).

$$S.E.(L^*) = \sqrt{\left[\frac{\sum_{k=0}^{n_s-3} (k+1) \frac{\binom{n_a+k+2}{k+2}}{\binom{n_s}{k+2}} \sum_{i=0}^{n_s-k-2} \binom{n_a+i+k+2}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k+2}}{\sum_0^{n_s} \binom{n_a+i}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}}\right] - (L^*)^2}. \quad (13)$$

Bayes Estimator of the average number of customers in the queue (L_q)

Putting $u = 2, v = 1, w = 1$ in (9), the Bayes estimator L^* of L is

$$L_q^* = \frac{\sum_{k=0}^{n_s-3} \frac{\binom{n_a+k+2}{k+2}}{\binom{n_s}{k+2}} \sum_{i=0}^{n_s-k-2} \binom{n_a+i+k+2}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k+2}}{\sum_0^{n_s} \binom{n_a+i}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}}. \quad (14)$$

The standard error of this estimator is obtained by substituting $u = 4, v = 2, w = 2$ in (9).

$$\text{S.E.}\left(L_q^*\right) = \sqrt{\left[\frac{\sum_{k=0}^{n_s-5} (k+1) \frac{\binom{n_a+k+4}{k+4}}{\binom{n_s}{k+4}} \sum_{i=0}^{n_s-k-4} \binom{n_a+i+k+4}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k+4}}{\sum_0^{n_s} \binom{n_a+i}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}}\right] - (L_q^*)^2}. \quad (15)$$

Bayes Estimator of the average waiting time of a customer in the system (W)

Putting $u = 0, v = 0, w = 1$ in (9), the Bayes estimator W^* of W is

$$W^* = \frac{\sum_{k=0}^{n_s-2} \frac{\binom{n_a+k}{k}}{\binom{n_s}{k+1}} \left(\frac{\beta+t_s}{k+1}\right) \sum_{i=0}^{n_s-k-1} \binom{n_a+i+k}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k}}{\sum_0^{n_s} \binom{n_a+i}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}}. \quad (16)$$

The standard error of this estimator is obtained by substituting $u = 0$, $v = 0$, $w = 2$ in (9).

$$S.E.(W^*) = \sqrt{\left[\frac{\sum_{k=0}^{n_s-3} (k+1) \frac{\binom{n_a+k}{k}}{\binom{n_s}{k+2}} \frac{(\beta+t_s)^2}{(k+1)(k+2)} \sum_{i=0}^{n_s-k-2} \binom{n_a+i+k}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k}}{\sum_0^{n_s} \binom{n_a+i}{i} \left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}}\right] - (W^*)^2}. \quad (17)$$

Bayes Estimator of the average waiting time of a customer in the queue (W_q)

Putting $u = 1, v = 1, w = 1$ in (9), the Bayes estimator W_q^* of W_q is

$$W_q^* = \frac{\sum_{k=0}^{n_S-3} \frac{\binom{n_a+k+1}{k+1}}{\binom{n_S}{k+2}} \left(\frac{\beta+t_S}{k+2}\right) \sum_{i=0}^{n_S-k-2} \binom{n_a+i+k+1}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i+k+1}}{\sum_0^{n_S} \binom{n_a+i}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i}}. \tag{18}$$

The standard error of this estimator is obtained by substituting $u = 2, v = 2, w = 2$ in (9).

$$S.E.\left(W_q^*\right) = \sqrt{\left[\frac{\sum_{k=0}^{n_S-5}(k+1)\frac{\binom{n_a+k+2}{k+2}}{\binom{n_S}{k+4}} \frac{(\beta+t_S)^2}{(k+3)(k+4)} \sum_{i=0}^{n_S-k-4} \binom{n_a+i+k+2}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i+k}}{\sum_0^{n_S} \binom{n_a+i}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i}}\right] - (W_q^*)^2}. \tag{19}$$

Bayes Estimator of Probability of the system state being greater than s

We have, $p = Pr(N \geq s) = \left(\frac{\lambda}{\mu}\right)^s$, where N is the state of the system.

Putting $u = s, v = s, w = 0$ in (9), the Bayes estimator p^* of p is-

$$p^* = \frac{\binom{n_a+s}{s}}{\binom{n_S}{s}} \frac{\sum_{i=0}^{n_S-s} \binom{n_a+s+k+1}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i+k+s}}{\sum_{i=0}^{n_S} \binom{n_a+i}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i}}. \tag{20}$$

The standard error of this estimator is obtained by substituting $u = 2s$, $v = 2s$, $w = 0$ in (9).

$$S.E.\left(p^*\right) = \sqrt{\left[\frac{\binom{n_a+2s}{2s}}{\binom{n_S}{2s}} \frac{\sum_{i=0}^{n_S-2s} \binom{n_a+2s+k+1}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i+k+2s}}{\sum_{i=0}^{n_S} \binom{n_a+i}{i} \left(\frac{\beta+t_S}{\alpha+t_a+t_S}\right)^{i}}\right] - (p^*)^2}. \tag{21}$$

3.2. Predictive distributions and their moments

Predictive distributions in the Bayesian paradigm are crucial for forecasting queuing system metrics, including queue lengths and waiting times. By quantifying uncertainty, these distributions optimize resource allocation and staffing, thereby enhancing system efficiency. This subsection outlines the methodology for deriving predictive distributions for the number of customers at departure epoch, waiting time in the system, and waiting time in the queue.

Let $x = (x_1, x_2, x_3, \ldots, x_n)$ be a random sample drawn from a distribution characterized by the probability mass function (p.m.f.) or probability density function (p.d.f.) $f(x|\theta)$, where θ belongs to the parameter space Ω. Given a prior distribution $\pi(\theta)$ on the parameter θ, and the corresponding posterior distribution $\Pi(\theta|x)$, the posterior predictive distribution for a future observation m, conditioned on the observed data x, is defined as (Chen *et al.*, 2012):

$$p(m|x) = \int_{\Omega} f(x|\theta)\, \Pi(\theta|x), d\theta. \tag{22}$$

3.2.1. Predictive distribution of the number in the system at departure epoch

Building on the distribution of M given in equation (1), we now derive the predictive distribution for the number of customers in the system at a departure epoch.

$$P(M = m|data) = \int_0^\infty \int_\lambda^\infty \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^m \Pi\,(\lambda, \mu|data)d\mu d\lambda$$

This integral simplifies to the following expression:

$$P(M = m|data) = \frac{\left[\frac{\binom{n_a+m}{m}}{\binom{n_s}{m}} \Sigma_{i=0}^{n_s-m}\left\{\binom{n_a+m+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+m}\right\}\right] - \left[\frac{\binom{n_a+m+1}{m+1}}{\binom{n_s}{m+1}} \Sigma_{i=0}^{n_s-m-1}\left\{\binom{n_a+m+1+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+m+1}\right\}\right]}{\left[\Sigma_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}. \tag{23}$$

Both raw and central factorial moments of this distribution exist and can be calculated as given below.

$$E[M(M-1)(M-2)\ldots(M-r+1)|data, \lambda, \mu] = \int_0^\infty \int_\lambda^\infty r! \frac{\lambda^r}{(\mu-\lambda)^r} \Pi(\lambda, \mu|data)d\mu d\lambda$$

which simplifies to

$$\mu'_{(r)} = r! \frac{\Sigma_{k=0}^{n_S-r-1}\left[\frac{\binom{r+k-1}{k}\binom{n_a+r+k}{r+k}}{\binom{n_s}{r+k}} \Sigma_{i=0}^{n_S-r-k}\left\{\binom{n_a+r+k+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k}\right\}\right]}{\left[\Sigma_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}. \tag{24}$$

Thus, the mean number of customers in the system at departure epoch is

$$\mu'_{(1)} = \mu'_1 = \frac{\Sigma_{k=0}^{n_s-2}\left[\frac{\binom{n_a+k+1}{k+1}}{\binom{n_s}{r+k}} \Sigma_{i=0}^{n_s-k-1}\left\{\binom{n_a+k+i+1}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k}\right\}\right]}{\left[\Sigma_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}$$

3.2.2. Predictive distribution of waiting time in the system

Proceeding from the distribution of the system waiting time, as given in (2), we now derive its predictive distribution as follows:

$$P(T = t|data) = \int_0^\infty \int_\lambda^\infty (\mu - \lambda)\exp[-(\mu - \lambda)t]\,\Pi\,(\lambda, \mu|data)d\mu d\lambda$$

Similarly, by integrating and using the incomplete gamma integral, the predictive distribution for waiting time is obtained as

$$P(T = t|data) = \frac{\left[\left(\frac{n_s+1}{\beta+t+t_s}\right)\Sigma_{i=0}^{n_s-m}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s+t}{\alpha+t_a+t_s}\right)^{i}\right\}\right] - \left[\left(\frac{n_a+1}{\alpha+t_a+t_s}\right)\Sigma_{i=0}^{n_s}\left\{\binom{n_a+1+i}{i}\left(\frac{\beta+t_s+t}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}{\left[\Sigma_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}. \tag{25}$$

The r[th] factorial moment is obtained as

$$E[T(T-1)\ \ldots(T-r+1)\,|data,\rho,\mu]\ =\int_0^\infty\int_\lambda^\infty \frac{r!}{(\mu-\lambda)^r}\Pi(\lambda,\mu|data)d\mu d\lambda$$

which, on simplification, gives

$$\mu'_{(r)}=r!\frac{\left[\sum_{k=0}^{n_s-r-1}\left\{\frac{\binom{r+k-1}{k}\binom{n_a+k}{k}k!(\beta+t_s)^r}{\binom{n_s}{r+k}(r+k)!}\sum_{i=0}^{n_s-r-k}\left\{\binom{n_a+k+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k}\right\}\right\}\right]}{\left[\sum_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}. \tag{26}$$

3.2.3. Predictive distribution of waiting time in the queue

Similarly, using (3), the predictive distribution for waiting time in the queue is obtained for two cases.

Case I: For $t=0$, the predictive distribution of waiting time in the queue is given by

$$P(T=t|data)=\int_0^\infty\int_\lambda^\infty\left(1-\frac{\lambda}{\mu}\right)\Pi\,(\lambda,\mu|data)d\mu d\lambda$$

On simplification, we have

$$P(T=t|data)=1-\left(\frac{n_a+1}{n_s}\right)\left[\frac{\sum_{i=0}^{n_s-1}\left\{\binom{n_a+i+1}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+1}\right\}}{\sum_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}}\right]. \tag{27}$$

Case II: For $t>0$, the predictive distribution of waiting time in the queue is given by

$$P(T=t|data)\ =\int_0^\infty\int_\lambda^\infty\frac{\lambda}{\mu}(\mu-\lambda)\,\Pi\,(\lambda,\mu|data)d\mu d\lambda$$

which simplifies to

$$P(T=t|data)=\frac{\left[\left(\frac{n_a+1}{\alpha+t_a+t_s}\right)\sum_{i=0}^{n_s}\left\{\binom{n_a+i+1}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]-\left[\frac{(n_a+1)(n_a+2)}{2n_s}\frac{(\beta+t_s)}{(\alpha+t_a+t_s)^2}\sum_{i=0}^{n_s-1}\left\{\binom{n_a+i+2}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}{\left[\sum_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]}. \tag{28}$$

The r[th] factorial moment of this distribution is

$$E[T(T-1)\ldots(T-r+1)|data,\lambda,\mu]=\int_0^\infty\int_\lambda^\infty\frac{r!\lambda}{\mu(\mu-\lambda)^r}\Pi(\lambda,\mu|data)d\mu d\lambda \tag{29}$$

Or,

$$\mu'_{(r)}=r!\frac{\left[\sum_{k=0}^{n_s-r-2}\left\{\frac{\binom{r+k-1}{k}\binom{n_a+k+1}{k+1}}{\binom{n_s}{r+k+1}}\frac{(k+1)!}{(r+k+1)!}\frac{(\beta+t_s)^{r+1}}{(\alpha+t_a+t_s)}\sum_{i=0}^{n_s-r-k-1}\left\{\binom{n_a+k+i+1}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i+k}\right\}\right\}\right]}{\left[\sum_{i=0}^{n_s}\left\{\binom{n_a+i}{i}\left(\frac{\beta+t_s}{\alpha+t_a+t_s}\right)^{i}\right\}\right]} \tag{30}$$

The formulae (27) and (28) provide the predictive distributions for the waiting time in the queue for two scenarios: $t\ =\ 0$ and $t\ >\ 0$.

4. Simulation

This section presents a simulation study to assess the numerical performance of the Bayes estimators derived in Section 3, with particular emphasis on ρ, a critical parameter that influences system stability and efficiency. Also, the predictive distribution of the number of customers at the departure epoch is evaluated.

To examine the reliability of the estimators, simulations are conducted across varying sample sizes $n = (50, 100, 200, 500)$. Based on these settings, individual interarrival and service times are generated using inverse transform sampling from exponential distributions for three parameter configurations: $(\lambda, \mu) = \{(12,15), (18, 20), (70, 100)\}$. From these, the aggregated interarrival and service times are obtained. Finally, the simulated data produce the Bayes estimates and other performance metrics. The results are detailed in the subsequent tables.

For the different sample sizes and varying pairs of (λ, μ), Bayes estimates of ρ and their MSEs are reported in Table 2. Table 3 presents the predictive distribution of the number of customers in the system at departure epoch. The MSEs serve as frequentist evaluation metrics in the simulation context. Given the Bayesian nature of our analysis, we additionally examine the standard errors of the estimates reported in Table 4. The 95% and 99% credible regions of ρ are presented in Table 5 and Table 6 respectively.

Table 2. Bayes estimates of ρ along with their MSEs for different pairs of arrival and service rates and for different sample sizes of interarrival and service times

Sample size	Pair 1 (12, 15)		Pair 2 (18, 20)		Pair 3 (70, 100)	
	Estimate of ρ	MSE	Estimate of ρ	MSE	Estimate of ρ	MSE
$n_a = 50;\ n_s = 50$	0.80901	0.00541	0.92465	0.00670	0.69730	0.00893
$n_a = 100;\ n_s = 100$	0.87290	0.00508	0.90795	0.00415	0.76627	0.02101
$n_a = 200;\ n_s = 200$	0.80164	0.00425	0.90175	0.00261	0.76858	0.01272
$n_a = 500;\ n_s = 500$	0.84318	0.00427	0.85272	0.00169	0.76319	0.00391

From Table 2, we can observe that the estimated values are very close to the true values and the MSEs are minimal. Pair 2 (18, 20) yields the most substantial gain, where MSE decreases from 0.00670 to 0.00261 as the sample size increases from 50 to 200. This represents a 61% reduction in the estimation error. Similarly, Pair 1 (12, 15) demonstrates a steady reduction in MSE from 0.00541 to 0.00425.

Table 3. Predictive distributions of number of customers at departure epoch for different pairs of arrival and service rates and for different sample sizes of interarrival and service times

Sample size	m	Predictive distribution		
		Pair 1 (12, 15)	Pair 2 (18, 20)	Pair 3 (70, 100)
$n_a = 100;\ n_s = 100$	0	0.11692	0.25864	0.12216
	1	0.09707	0.18197	0.10076
	2	0.08130	0.13003	0.07044
	3	0.06867	0.09434	0.05964
	4	0.05846	0.06945	0.05088
	5	0.04329	0.05187	0.04373
$n_a = 200;\ n_s = 200$	0	0.19628	0.21151	0.25163
	1	0.15191	0.16096	0.18281
	2	0.11851	0.12350	0.13398
	3	0.09317	0.09552	0.09905
	4	0.07282	0.07448	0.07385
	5	0.05893	0.05853	0.05553
$n_a = 500;\ n_s = 500$	0	0.22895	0.09320	0.39798
	1	0.17415	0.08211	0.23814
	2	0.13297	0.07532	0.14306
	3	0.10190	0.06425	0.08628
	4	0.07839	0.05706	0.05224
	5	0.06053	0.05080	0.03175

The predictive distributions in Table 3 exhibit theoretically consistent behavior with probability mass concentrating at lower queue states and following the expected exponential decay pattern of M/M/1 systems. Notably, Pair 3 (70, 100) achieves convergence with the probability of an empty system increasing from 0.12216 to 0.39798 as the sample size grows from 100 to 500. Pair 1 (12, 15) demonstrates steady improvement with $P(m = 0)$ rising from 0.11692 to 0.22895. Pair 2 (18, 20) exhibits high-utilization behavior, with the distribution of probabilities across queue states aligning with expected steady-state characteristics. The results indicate the effectiveness of this predictive mechanism in predicting queue performance under varying configurations.

Since different methods of estimating ρ have been previously studied in the literature, here a comparison is carried out against some of the earlier established techniques and our approach. Singh and Acharya (2019) estimated ρ using both Maximum Likelihood (ML) and Bayesian procedures. Their estimates yield higher Mean Squared Errors (MSEs) of ρ using MLE (0.0112 for a sample size of 100 and 0.0413 for a sample size of 50). The corresponding MSEs for the Bayes estimates are 0.0053 and 0.0167, respectively. In contrast, this article reports an MSE of ρ for 100 samples at 0.00425 and that for a sample size 50 at 0.00541. This illustrates that both ML and Bayesian estimates result in higher MSEs compared to the findings of the current paper. Singh *et al.* (2021)

reported Root Mean Squared Errors (RMSEs) using ML and Bayesian estimation. For $n = 50$, the RMSEs are 0.1407 and 0.1133, respectively, indicating higher errors. For $n = 100$, the RMSE using MLE was 0.1004, also demonstrating higher errors. Several simulation designs by Sohn (1996) show the MSEs of Bayesian estimates of ρ. For a sample size 100, the reported MSEs of various estimators for ρ exceed those obtained in this study: ad-hoc Bayes estimator using lognormal prior with covariates (0.088, 0.0109), gamma prior with covariates (0.0900, 0.0099), model-based regression estimator (0.6280, 0.0137), and data-based raw estimator (0.1070, 0.0400) all demonstrate higher MSEs compared to the results presented here for equivalent sample size.

Table 4 shows that the standard errors of these Bayes estimates are low and decrease with an increasing sample size, thereby confirming the precision and reliability improvements inherent in Bayesian estimation.

Table 4. Standard error (SE) of the estimates of traffic intensity (ρ) for different pairs of arrival and service rates and for different sample sizes of interarrival and service times

Sample Size	**Pair of (λ, μ)**		
	(12, 15)	**(18, 20)**	**(70, 100)**
$n_a = 50, n_s = 50$	0.1070	0.0864	0.0874
$n_a = 100, n_s = 100$	0.0742	0.0752	0.0735
$n_a = 200, n_s = 200$	0.0727	0.0709	0.0646
$n_a = 500, n_s = 500$	0.0486	0.0510	0.0490

Table 5. 95% credible intervals of traffic intensity (ρ) for different pairs of arrival and service rates (λ, μ)

Sample Size	**Pair of (λ, μ)**		
	(12, 15)	**(18, 20)**	**(70, 100)**
$n_a = 50, n_s = 50$	(0.5040, 0.9704)	(0.5565, 0.9847)	(0.5549, 0.9844)
$n_a = 100, n_s = 100$	(0.6567, 0.9883)	(0.7147, 0.9939)	(0.6461, 0.9865)
$n_a = 200, n_s = 200$	(0.6611, 0.9589)	(0.7357, 0.9903)	(0.6366, 0.9345)
$n_a = 500, n_s = 500$	(0.6816, 0.8733)	(0.7653, 0.9703)	(0.6268, 0.8032)

Table 6. 99% Credible intervals of traffic intensity (ρ) for different pairs of arrival and service rates (λ, μ)

Sample Size	**Pair of (λ, μ)**		
	(12, 15)	**(18, 20)**	**(70, 100)**
$n_a = 50, n_s = 50$	(0.4457, 0.9927)	(0.4931, 0.9960)	(0.4916, 0.9959)
$n_a = 100, n_s = 100$	(0.6030, 0.9968)	(0.6586, 0.9979)	(0.5930, 0.9964)
$n_a = 200, n_s = 200$	(0.6216, 0.9878)	(0.6926, 0.9972)	(0.5930, 0.9964)
$n_a = 500, n_s = 500$	(0.6554, 0.9079)	(0.7261, 0.9906)	(0.6029, 0.8352)

In both Table 5 and Table 6, a notable decrease in the width of the credible intervals is observed as the sample size increases from 50 to 500, thereby enhancing the precision of the posterior estimates. Under $(\lambda, \mu) = (18, 20)$, the 95% credible interval narrows from (0.5565, 0.9847) at $n = 50$ to (0.7653, 0.9703) at $n = 500$. Similarly, the 99% credible interval under the same rate pair reduces from (0.4931, 0.9960) to (0.7261, 0.9906). This pattern remains consistent across all three rate pairs, confirming that the posterior distribution concentrates around the true value of ρ as the sample size increases. Despite the differing nominal values of ρ, all three rate pairs produce credible intervals that broadly cover the true value of traffic intensity. For $(\lambda, \mu) = (70, 100)$, the intervals are generally narrower and converge more quickly. This indicates enhanced estimation performance at lower traffic intensity levels.

Comparing the 95% (Table 5) and 99% (Table 6) credible intervals reveals the expected widening of intervals at higher confidence levels, reflecting the inverse relationship between precision and certainty. The difference in bounds between the two tables is evident at lower sample sizes, showing greater uncertainty. At $n = 50$, the 99% interval for the pair (12, 15) is (0.4457, 0.9927), compared to the 95% interval (0.5040, 0.9704). At $n = 500$, this expands only slightly from (0.6816, 0.8733) at 95% to (0.6554, 0.9079) at 99%. This behavior reaffirms that Bayesian credible intervals appropriately reflect posterior uncertainty, with interval width decreasing as data volume increases, and interval coverage increasing as the confidence level rises.

Implementation of the proposed estimators:

The following steps outline the simulation-based procedure used to implement the Bayesian estimators for arrival and service rates, as well as associated performance measures:

Step 1: Select appropriate values for the hyperparameters α and β based on prior knowledge.

Step 2: Choose multiple pairs of (λ, μ) such that $\lambda < \mu$ to ensure system stability.

Step 3: Fix the sample sizes for the interarrival and service times.

Step 4: Generate random samples of interarrival and service times from exponential distributions with rates λ and μ.

Step 5: Calculate the total interarrival time $(t_a)$and the total service time (t_s) from the simulated data.

Step 6: Substitute the values of α, β, n_a, n_s, t_a , and t_s into the Bayesian estimators derived in Section 3. The output yields the Bayes estimates along with their associated standard errors for the desired performance measures.

5. Real-life queueing problem

To demonstrate the practical applicability of the proposed Bayesian framework, we apply the derived estimators to empirical data obtained from (Czech University of Life Sciences Prague, n.d., § 9.6.6)

The empirical analysis is based on observational data from a service system characterized by stochastic customer arrivals following a Markovian process. The dataset, encompasses 30 customer transactions recorded over an 8-hour operational period. The data collection includes timestamped measurements of arrival times, service initiation times, service completion times, and derived performance metrics, including queue length, interarrival intervals, service durations, and customer waiting times.

From the given data, interarrival times were calculated as the time difference between consecutive customer arrivals. The dataset yields:

- **Interarrival times**: $[t_1, t_2, \ldots, t_{30}]$ where t_{i} represents the time between arrivals $i-1$and i, $i = 0,1,2, \ldots, 30$
- **Service times**: $[s_1, s_2, \ldots, s_{30}]$ representing individual service durations.

These values are subsequently used to compute: $t_a = \Sigma t_{\mathrm{i}} = 387$ and $t_s = \Sigma s_{\mathrm{i}} = 362$

The processed data serve as input for the Bayesian procedure to obtain posterior estimates of the traffic intensity ρ and its standard error. The results obtained are $\rho^* = 0.80503$ and $S.E.(\rho^*) = 0.10253$

6. Conclusion

This study presents a Bayesian approach to parameter estimation of an M/M/1 queuing model using a bivariate prior and a simple Markovian likelihood of arrival and service rate. Bayes estimates of ρ and predictive distributions at departure epoch are simulated using inverse transform sampling. The results demonstrate good accuracy with low standard errors and MSEs as low as 0.00169 for $n = 500$. Also, the predictive distributions exhibit strong convergence, with probability mass concentrating around key values. Moreover, the credible intervals computed encompass true parameter values. As the sample size increases from 50 to 500, the credible intervals become narrower, thereby indicating improved precision. The 99% intervals are wider than the 95% intervals, particularly at smaller sample sizes, confirming the statistical validity of the Bayesian procedure. This enables users to assess uncertainty at different confidence levels depending on the context of the application. Therefore, the findings establish that the developed methodology is reliable for uncertainty assessment of traffic intensity, with improved performance at larger sample sizes and lower utilization levels.

While the proposed Bayesian model demonstrates strong predictive capabilities and practical relevance, several considerations may guide future research. First, the use of the M/M/1 queueing model, although widely studied, may not fully reflect the dynamics of more complex or non-Markovian service systems encountered in real-world scenarios. Second, it introduces a specific bivariate prior distribution for parameter inference. Exploring alternative bivariate priors and higher-order Markovian and non-Markovian queues could provide further insights into estimation accuracy and model robustness.

Acknowledgement

The authors thank the anonymous reviewers for their detailed and thoughtful comments, which have significantly contributed to improving the quality of this paper.

References

Armero, C., Bayarri, M. J., (1994a). Bayesian prediction in M/M/1 queues. *Queueing Systems*, 15(1), pp. 401–417.

Armero, C., Bayarri, M. J., (1994b). Prior assessments for prediction in queues. *Journal of the Royal Statistical Society Series D: The Statistician*, 43(1), pp. 139–153.

Almeida, M. A., Cruz, F. R., (2018). A note on Bayesian estimation of traffic intensity in single-server Markovian queues. *Communications in Statistics-Simulation and Computation*, 47(9), pp. 2577–2586.

Basak, A., Choudhury, A., (2021). Bayesian inference and prediction in single server M/M/1 queuing model based on queue length. *Communications in Statistics-Simulation and Computation*, 50(6), pp. 1576–1588.

Basak, A., Choudhury, A., (2024). Bayesian estimation of finite buffer size in single server Markovian queuing system. *International Journal of System Assurance Engineering and Management*, 15(6), pp. 2366–2373.

Basak, A., Choudhury, A., (2025). Estimating arrival and service rates from queue length observations. *Journal of Computational and Applied Mathematics*, 455, p. 116196.

Basawa IV, Prabhu N. U., (1988). Large sample inference from single server queues. *Queuing Systems*, Vol. 3, pp. 289–304.

Bura, G. S., Sharma, H., (2024). Maximum likelihood and Bayesian estimation on M/M/1 queueing model with balking. *Communications in Statistics-Theory and Methods*, 53(14), pp. 5117–5145.

Chen, M. H., Shao, Q. M. and Ibrahim, J. G., (2012). *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media.

Choudhury, C., Borthakur, A. C., (2008). Bayesian inference and prediction in the single-server Markovian queue. *Metrika*, Vol. 67, pp. 371–383.

Choudhury, A., Basak, A., (2018a). Statistical inference on traffic intensity in an M/M/1 queueing system. *International Journal of Management Science and Engineering Management*, Vol. 13, pp. 274–279.

Choudhury, A., Basak, A., (2018b) Inference About Mean Number of Customers in the stable M/M/1 queue. *J. Assam Sc. Soc.*, Vol. 59, pp. 100–108.

Chowdhury, S., Mukherjee, S. P., (2013). Estimation of traffic intensity based on queue length in a single M/M/1 queue", Communications in Statistics – Theory and Methods, Vol. 42, pp. 2376–2390.

Clarke A. B., (1957). Maximum likelihood estimates in a simple queue. *Annals of Mathematical Statistics*, Vol. 28, pp. 1036–1040.

Cruz, F. R., Quinino, R. D. C. and Ho, L. L., (2017). Bayesian estimation of traffic intensity based on queue length in a multi-server M/M/s queue. *Communications in Statistics-Simulation and Computation*, 46(9), pp. 7319–7331.

Cruz, F. R., Almeida, M. A., D'Angelo, M. F. and van Woensel, T., (2018). Traffic intensity estimation in finite Markovian queueing systems. *Mathematical Problems in Engineering*, 2018(1), p. 3018758.

Czech University of Life Sciences Prague, n.d., (2025). *Queueing Models*. Available at: https://orms.pef.czu.cz/QueuingModels.html (Accessed: 19 August 2025).

Das, D., Tamuli, A. and Choudhury, A., (2025). Classical and Bayesian estimation of performance measures in a novel M/M/1 queueing model with balking. *Communications in Statistics-Simulation and Computation*, pp. 1–18.

Dutta, K., Choudhury, A., (2020). Estimation of performance measures of M/M/1 queues-a simulation-based approach. *Int. J. Applied Management Science*, Vol. 12, pp. 265–279.

Gross D., Harris, C. M., (2008) Fundamentals of Queueing Theory (3rd ed.), *Wiley*, India (P.) Ltd., New Delhi.

Jose J. K., Manoharan, M., (2014). Bayesian Estimation of Rate Parameters of Queuing Models. *Journal of Probability and Statistical Science*, Vol. 12, pp. 69–76.

Jose, J. K., Manoharan, M., (2018). Bayesian estimation of queue parameters by bivariate prior distributions with order restriction. *Journal of the Indian Statistical Association*, Vol. 56, pp. 29–46.

Kushvaha, B., Das, D. and Tamuli, A., (2025). Bayesian inference of traffic intensity in M/M/1 queue under symmetric and asymmetric loss functions. *Monte Carlo Methods and Applications*, 31(2), pp.109–118.

Lilliefors, H. W., (1966). Some confidence intervals for queues. *Oper. Res.*, Vol. 14, pp. 723–727.

McGrath, M. F., Gross, D. and Singpurwalla, N. D., (1987a). A subjective Bayesian approach to the theory of queues: I – Modeling. *Queuing Systems*, Vol. 1, pp. 317–333.

McGrath, M. F., Gross, D. and Singpurwalla, N. D., (1987b). A subjective Bayesian approach to the theory of queues: II – Inference and information in M/M/1 queues. *Queuing Systems*, Vol. 1, pp. 335–353.

Medhi, J., (2003). *Stochastic models in queuing theory (2nd ed.)*. New York: Academic Press.

Scruben, L., Kulkarni, R., (1982). Some consequences of estimating parameters for the M/M/1 queue. *Operations Research Letters*, Vol. 1, pp. 75–78.

Sharma, K. K., Kumar, V., (1999) Inference on M/M/1: (∞: FIFO) queue systems. *OPSEARCH*, Vol. 36, pp. 26–34.

Singh, S.K., Acharya, S.K. (2019) Bayesian change point problem for traffic intensity in $M/Er/1$ queueing model. *Japanese Journal of Statistics and Data Science,* 2, pp. 49–70 (2019). https://doi.org/10.1007/s42081-018-0026-2

Singh, S. K. et.al., (2021). Estimation of traffic intensity from queue length data in a deterministic single server queueing system, Journal of Computational and Applied Mathematics, Vol. 398, 113693, ISSN 0377-0427, https://doi.org/10.1016/j.cam.2021.113693.

Sohn, S. Y., (1996). Influence of a prior distribution on traffic intensity estimation with covariates. *Journal of Statistical Computation and Simulation*, *55*(3), pp. 169–180. https://doi.org/10.1080/00949659608811759.

Suyama, E., Quinino, R. C. and Cruz, F. R., (2018). Simple and yet efficient estimators for Markovian multiserver queues. *Mathematical Problems in Engineering*, 2018(1), p. 3280846.

Zheng, S., Seila, A. F., (2000) Some well-behaved estimators for the M/M/1 queue, *Operations Research Letters*, Vol. 26, pp. 231–235.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 87–111, https://doi.org/10.59139/stattrans-2026-029
Received – 05.06.2025; accepted – 07.05.2026

The bias of the estimation of income distribution parameters in non-probability surveys: an experiment based on EU-SILC microdata for Poland

Arkadiusz Kozłowski[1]

Abstract

The selection bias is a major issue when using non-probability samples for inference about finite populations. Minimizing it requires sufficient auxiliary power from external sources and proper modelling. This article examines the potential bias of estimates of income distribution parameters using real-life data from a household survey (European Union Statistics on Income and Living Conditions, EU-SILC) and computer simulations. It investigates the change in the performance of the estimates depending on data defectiveness, the sample size, the pool of auxiliary variables, sources of auxiliary information, and the estimation method. The non-probability samples had non-ignorable bias intentionally introduced by designing selection mechanisms directly tied to income or the access to a computer and the Internet. The results indicated that the bias of the estimates behaved predictably for simple parameters, such as the median and mean. However, for more complex parameters, the bias patterns were less stable. The use of sociodemographic variables for weighting did not consistently reduce the bias; on the contrary, in certain cases, it led to the aggravation of the bias. The number of auxiliary variables and the sample size were as important for the bias reduction as the actual estimation methods, of which the doubly robust estimation performed best.

Key words: selection bias, non-probability sampling, non-ignorable mechanism, income inequality, data weighting.

1. Introduction

Survey research is typically conducted on samples of elements with the aim of estimating the parameters of the whole finite population from which a sample was drawn. The method of sample selection is the key element in determining the possibility and validity of subsequent inference about the population parameters. Over the past

[1] Faculty of Management, Department of Statistics, University of Gdańsk, Gdańsk, Poland.
E-mail: arkadiusz.kozlowski@ug.edu.pl. ORCID: https://orcid.org/0000-0001-6282-1494.

70 years, the predominant paradigm in survey research methodology has been probability (random) sampling for sample selection and a design-based (randomisation) approach for inference about finite population parameters. The foundation of this approach was established by Neyman (1934) and subsequently developed by Hansen and Hurwitz (1943) and Horvitz and Thompson (1952), among others. The classic works on survey sampling fall within this framework; see, for example, Cochran (1963), Kish (1995), and Särndal et al. (1997).

However, random sampling is costly and not always feasible, primarily due to the need to have a list of all units belonging to the study population (sampling frame) and the need to reach randomly selected (inconvenient) units. In the practice of social research, a sample designed to be random is rarely truly random in the strict sense because of the incorrect mapping of units in the sampling frame (coverage error) and the presence of non-responses. Another common obstacle in list-based samples is the lack of anonymity of the respondents (the entities in a sampling frame must be identifiable so that a researcher can reach them if they are sampled), which makes it difficult to research sensitive topics.

In contrast, non-probability sampling (e.g. quota sampling, haphazard sampling, purposive sampling, see Baker et al., 2013) offers many practical advantages for the researcher. These include convenience, low costs, and the speed of the attainment of results. Recent trends pertaining to the conditions in which surveys are conducted, such as declining response rates in traditional probability surveys, the increased burden on respondents, the ever-growing demand for timely statistics, and the abundance of non-probability data sources (including *big data*), have prompted questions about the possible replacement of traditional research based on probability samples with the non-probability equivalents, even for official statistics (Beaumont, 2020; Beaumont and Rao, 2021; Kalton, 2023).

For non-probability samples to be a valid source for inference about finite populations, their two main deficiencies must be properly handled. The first is the lack of a widely acceptable framework for the assessment of estimation error. In probability sampling, the inference is based on the framework of possible outcomes from repeated draws of a sample with known inclusion probabilities. In non-probability sampling, the inclusion probabilities are not known (the sheer concept of inclusion probability in the non-probability sample and the possible repetition of the selection process are not clear, see, e.g. Lohr 2022, p. 526). The next section provides a short overview of the most important approaches to inference from non-probability samples.

The second weakness is the selection bias - systematic differences between the composition of a sample and the true population characteristics caused by the selection mechanism. Selection bias arises when the inclusion into the sample is somehow correlated with the studied variables. For example, to estimate how frequently people with

disabilities participate in public events, the sample may be selected as a convenience sample. People who go out more often, attend concerts, go to restaurants, etc., will be easier (more convenient) to recruit; therefore, their participation in the sample will be greater than their proportion in the general population. Such a sample will be biased and will generate systematically erroneous results (Szreder and Kozłowski, 2024, pp. 73–74).

Selection bias can be reduced if the researcher knows the selection mechanism well and can properly model it using auxiliary variables and additional reliable data sources, such as a probability sample or administrative records. These auxiliary variables and the right model are crucial for bias reduction. If the variables at hand have weak explanatory power for the selection mechanism, the bias could still be eliminated, provided they have strong explanatory power for the target variables. If a model captures all the forces that influence entry into a non-probability sample and/or the variability of the study variable, the units in and out of the non-probability sample are then exchangeable given the variables in the model (Mercer et al., 2017; i.e. conditionally exchangeable; see Li, 2024). Being so is the crucial and usually unmet assumption for the validity of inference from non-probability samples.

In most social surveys, it is reasonable to assume that the right model for the selection mechanism or the study variables could never be found due to the complexity of phenomena such as opinions, behaviors, attitudes, etc. In addition, the issue of finding the right model is complicated by the fact that the selection bias depends on the study variable; thus, the model properly capturing the selection forces may be different for different variables. Most surveys based on probability samples in practice use one set of demographic variables to calibrate weights assigned to sampled units to known population totals in order to, inter alia, reduce bias stemming from non-responses and other non-random errors. The question is how useful such variables could be in mitigating the effect of selection bias. Several researchers have noted that demographics and simple weighting are not sufficient to reduce the bias; see the review in Cornesse et al. (2020), as well as studies by Lavrakas et al. (2022) and Mercer and Lau (2023). They uncovered this by comparing probability and non-probability surveys with similar objectives with a benchmark. In this article, we wanted to check to what extent demographics and some weighting procedures can mitigate the selection bias, but instead of having a couple of real non-probability samples, we turned to a Monte Carlo simulation to have a broader scope of possibilities. Still, our experiment leveraged natural relationships between variables in a social survey, not artificially generated.

Using computer simulations on real-life data for generating non-probability samples is not obvious. When testing the performance of the proposed estimation methods for non-random samples, artificially generated data are often used (e.g. Chen, Li and Wu, 2020; Yang, Kim and Song, 2020; Kim and Morikawa, 2023). The use of real data

(from random and non-random samples) is often limited to a one-off application where, rather than assessing replicative properties, results are compared to known population values (e.g. Wang et al., 2015; Chen, Valliant and Elliott, 2019) or reliable probability survey estimates (e.g. Beaumont et al., 2024), or the estimates themselves and some aspects of estimation procedures are evaluated (e.g. Chen, Li and Wu, 2020; Yang, Kim and Song, 2020; Kim et al., 2021).

To assess the replicative properties of estimation from non-probability samples using real data, probability sampling is used, but with different selection probabilities, in a way that mimics the possible non-probability mechanism. Valliant (2020) and Buelens et al. (2018) used cut-off sampling (planned undercoverage), which deliberately excluded certain elements from the frame. Also, Valliant (2020) and Lee & Valliant (2009) used stratified sampling with disproportional allocation. Marella (2023) used Poisson sampling. Liu et al. (2023) used unequal probability random systematic sampling without replacement.

Most authors in simulation studies (on artificial or real non-probability data) test the ignorable mechanism. Testing a non-ignorable mechanism, such as where the probability of getting into a non-random sample depends on the target variable, is done by Kim & Morikawa (2023) and Marella (2023).

Virtually all simulations on non-probability samples focus on estimating means (including proportions) and totals. A notable exception is the study by Wiśniowski et al. (2020), where regression parameters were estimated.

The simulation in this study aimed to test the extent to which the bias caused by non-ignorable selection mechanisms can be minimized using socio-demographic auxiliary variables when estimating parameters of household income distribution. Contrary to most studies, where only means are tested, the parameters estimated include non-linear functions of the target variable. The estimation methods cover various techniques, but only the variants where a single vector of weights can be derived have been used. This was to mimic the production of estimates in most probability surveys. The estimation methods also do not take into account the non-ignorability of the selection process. In reality, it is usually not known whether the selection process is ignorable or not. If the conditions are ignorable, one can expect unbiased estimates with manageable precision. The interesting part is what happens when the process is non-ignorable, and the performance of estimators depends on the natural relationship between income and socio-demographic characteristics.

The scenario tested in the simulation is that the target variable is only observed in a non-probability sample, and the estimation is strengthened by additional information on auxiliary variables from an independent probability sample and/or the entire popu-

lation. The data from a large-scale probability survey is used for the experiment. Although the experiment is based on one specific survey, it aims to introduce a methodological framework for such an assessment and identify certain regularities.

The rest of the article is structured as follows. The next section is devoted to the primary approaches to estimation in non-probability sample surveys. Our method of investigating estimation bias is then presented, starting with the real data description, followed by methods of generating non-probability samples, and ending with the estimation procedures. The results of the experiment are presented in Section 4, and some intriguing outcomes are discussed in Section 5. The article ends with a conclusion section.

2. Estimation from non-probability samples

We are interested in the finite population of elements indexed by $U = \{1, 2, \dots, N\}$. The elements may be observed on a study variable y and a vector of covariates $\boldsymbol{x}$. The population is considered constant, meaning that it is assumed that the surveying element i ($i = 1, 2, \dots, N$) would always yield the same values y_i and $\boldsymbol{x}_i$. A survey aims to estimate population parameters (functions of y) based on a sample of elements and possibly some auxiliary information. The unknown part of this setup is which elements are included in the sample. Following Meng (2018), we introduce the sample membership indicator R_i ($R_i = 1$ if element i is in the sample, $R_i = 0$ otherwise). The indicator is the result of all forces that determine whether an individual enters the sample. These include the selection mechanism (probabilistic or not), non-response, coverage errors, and others. Most of them are not controlled by the researcher. In probability sampling and pure conditions (without non-random errors), the distribution of R_i is known; notably, $E(R_i) = P(R_i = 1) = \pi_{P,i}$ (first-order inclusion probabilities) are known for each population element prior to sampling. These probabilities, along with second-order inclusion probabilities, $\pi_{P,ij} = P(R_i R_j = 1)$, play a central role in the generalization of the sample outcomes to the entire population (Särndal, Swensson and Wretman, 1997). In practice, the pure conditions rarely occur, and the selection mechanism is often a non-probability one.

Consider the non-probability selection mechanism leading to a sample s_{NP}. The inclusion probabilities are not known; thus, the usual randomization approach for inference about the population parameters cannot be applied. Nevertheless, the body of statistical theory offers some possibilities of inference from non-probability samples. Two primary approaches are quasi-randomization and superpopulation modelling (Zhang, 2019; Valliant, 2020). Other solutions can be regarded as variations or combinations of these two approaches. Every method requires some additional data sources

aside from a non-probability one, some auxiliary variables besides the ones studied, and explicit or implicit modelling.

To effectively use non-probability data for the production of estimates, a single vector of weights is desirable. The weights should be independent from any study variable and constructed in such a manner that a linear combination of the values of the study variable with the weights as coefficients yields an estimate of the total:

$$\hat{y} = \sum_{i \in s_{NP}} w_{NP,i} \cdot y_i \tag{1}$$

where $w_{NP,i}$ is the weight for unit i in the non-probability sample. Using a single vector of weights is the usual practice in probability surveys. It is convenient because the weights could be used for different y variables and parameters. In this study, only estimation methods that allow the construction of a single weight vector will be used.

The quasi-randomization approach (also: inverse probability weighting or propensity score adjustment, PS) is conceptually equivalent to the method introduced earlier to address nonresponse bias (see Valliant et al. 2013, pp. 321–338). In the nonresponse case, the mechanism leading to nonresponse is treated as the second phase of sampling; in this phase, unlike in the usual sampling design treated here as the first phase, the inclusion probabilities are not known and must be modelled after the survey is conducted based on information regarding both respondents and non-respondents. In the non-probability sample case, the aim is to estimate the (pseudo)probabilities of being included in s_{NP}. The estimation is performed after the sample is selected and is based on covariates that are believed to affect the inclusion in the non-probability sample. After the pseudo-inclusion probabilities are estimated, the inverses of the estimated inclusion probabilities can be used as weights in the production of estimates:

$$w_{NP,\,i}^{(PS)} = \frac{1}{\hat{\pi}_{NP,\,i}} \tag{2}$$

where $\hat{\pi}_{NP,\,i}$ is the estimated pseudo-inclusion probability for unit i in the s_{NP}. There are other propensity-score-based methods (see, e.g. Rivers, 2006), but they are beyond the scope of this article.

Superpopulation modelling (SM; model-based approach) is a prediction-based paradigm in which the values of variables of interest for unsampled units are predicted using specified models, and population estimates are generated as functions of these predictions (Valliant, Dorfman and Royall, 2000). This approach is predicated upon the idea that a finite population is merely a sample from a superpopulation. This superpopulation is defined by the model binding the variable of interest with specific covariates. In other words, it is assumed that finite population elements are generated by the model with certain fixed parameters and a random error. The model parameters are estimated from the sample elements, and the estimator of the total is then constructed

simply as a sum of y-values over the sample elements plus the sum of predictions over the non-sampled elements (Elliott and Valliant, 2017). There are several variants of the model-based approach, such as multilevel regression and poststratification method (Gelman and Little, 1997), mass imputation (Kim et al., 2021), using penalized estimation (Liu, Wang and Pan, 2025) or machine learning techniques (Ferri-García, Castro-Martín and Rueda, 2021).

The model-based estimators are not typically presented as a linear combination of sample elements, but they could be written in this form for simple models such as the linear model below:

$$E_M(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad V_M(y_i|\boldsymbol{x}_i) = v \tag{3}$$

where E_M and V_M are expectation and variance with respect to the superpopulation model. Estimating parameter vector $\boldsymbol{\beta}$ from the non-probability sample using the ordinary least squares method allows the weights to be presented in the following form:

$$w_{NP,\,i}^{(SM)} = \widehat{\boldsymbol{X}}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i \tag{4}$$

where $\widehat{\boldsymbol{X}}$ is the vector of population totals for auxiliary variables (these totals could be known or estimated from the probability reference sample), and $\boldsymbol{X}$ is a matrix of values for auxiliary variables from the non-probability sample.

A combination of quasi-randomization and superpopulation approaches yields a doubly robust estimation (DR). In this method, the PS weights are involved in making predictions based on the model. The doubly robust estimation is robust in the sense that it protects against the misspecification of either the inclusion or the superpopulation model (Chen, Li and Wu, 2020; Yang, Kim and Song, 2020).

In the doubly robust estimation, adopting the same linear model (3) translates into the following formula for the weights (see Valliant 2020, p. 243):

$$w_{NP,\,i}^{(DR)} = w_{NP,\,i}^{(PS)}\left[1 + \left(\widehat{\boldsymbol{X}} - \widehat{\boldsymbol{x}}_{PS}\right)^T(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{x}_i\right] \tag{5}$$

where $\widehat{\boldsymbol{x}}_{PS} = \sum_{i\in s_{NP}} w_{NP,\,i}^{(PS)} \cdot \boldsymbol{x}_i$ is the vector of estimated population totals for x-variables using the PS method and $\boldsymbol{W} = diag\left\{w_{NP,\,i}^{(PS)}\right\}$. The weights in (5) are the same as the linear calibration weights, where the initial weights, rather than design weights, are the weights (2) from the PS method (see Deville and Särndal, 1992).

3. Methodology of the experiment

3.1. Data

The data for the experiment come from the European Union Statistics on Income and Living Conditions (EU-SILC) survey. The aim of the survey is to obtain estimates

of income, poverty, social exclusion, and other aspects of living conditions. The survey is harmonized across 32 European countries, and its implementation is overseen by Eurostat. In each country, the survey "must be based on a nationally representative probability sample of the population residing in private households within the country" (European Commission 2022, p. 59). The data used in the experiment are microdata obtained from Statistics Poland pertaining to the 2022 edition conducted in Poland (Eurostat, 2024). The sample in this survey was selected using a two-stage sampling scheme, with enumeration census areas as primary and dwellings as secondary sampling units (all the households in the selected dwellings are surveyed). Stratified sampling with disproportional allocation was used at the first stage and simple random sampling without replacement in the second stage. Stratifying variables were NUTS 2 regions, city size classes and share of richest dwellings in the primary sampling unit. The last feature was calculated using information obtained from tax records (GUS, 2023).

The variables considered in the simulation concerned the household as a whole or the person completing the questionnaire. As the main objectives of the EU-SILC survey relate to incomes and income is also one of the variables in social surveys that might be most heavily affected by non-probabilistic sample selection (in the same way as it is affected by non-random errors, concerning measurement, coverage and non-response, especially for the top income units; see e.g. Brzeziński, Sałach and Wroński, 2020; Carranza, Morgan and Nolan, 2023), the main parameters of interest in our experiment are various parameters of income distribution. The main study variable is the equivalized annual disposable income (y_i), which is not directly observed but is calculated as follows:

$$y_i = \frac{TDHI_i}{EHS_i} \tag{6}$$

where: $TDHI_i$ is the total disposable household income for a household i, and EHS_i is the equivalized household size, calculated as a sum of weights: 1 for the first person aged 14 and over, 0.5 for each additional person at this age, and 0.3 for each person aged 13 or less. For brevity, the equivalized annual disposable income will hereinafter be referred to as "equivalized income".

The value of y_i is assigned to each person in the household i. The following parameters describe the population of persons, but the statistical unit in the dataset is a household. Therefore, the household data is considered as grouped data of persons, which is reflected in the way the parameters are calculated (i.e. they are weighted by the number of persons in a household). Additionally, for all the formulas that follow, except the mean, it is assumed that the households are ordered in a non-decreasing way with

respect to y. The parameters to be estimated and their population formulas are as follows:

- Mean (Mean of equivalized income for a person):

$$Y_{Mean} = \frac{\sum_{i=1}^{N} y_i \cdot z_i}{\sum_{i=1}^{N} z_i} \tag{7}$$

where: z_i is the household size (total number of household members).

- Median (Half of the persons in the population have equivalized income not higher than the median):

$$Y_{Median} = min\left\{ y_i \middle| \sum_{i'=1}^{i} p_{i'} \geq 0.5 \right\} \tag{8}$$

where: $p_{i'} = \frac{z_{i'}}{\sum_{i=1}^{N} z_i}$.

- Poverty (At-risk-of-poverty rate - the percentage of persons with an equivalized income below the at-risk-of-poverty threshold set at 60% of the median):

$$Y_{Poverty} = \frac{\sum_{i=1}^{N} I(y_i < 0.6 \cdot Y_{Median}) \cdot z_i}{\sum_{i=1}^{N} z_i} \tag{9}$$

where: $I(u)$ is the indicator function, $I(u) = 1$ if u is true, and $I(u) = 0$ otherwise.

- Depth (Relative median at-risk-of-poverty gap - the difference between the median equivalized income of persons below the at-risk-of-poverty threshold and this threshold, expressed as a percentage of the at-risk-of-poverty threshold):

$$Y_{Depth} = \frac{0.6 \cdot Y_{Median} - \tilde{Y}_{Median}}{0.6 \cdot Y_{Median}} \tag{10}$$

where $\tilde{Y}_{Median} = min\{y_i \mid \sum_{i'=1}^{i} p'_{i'} \geq 0.5 \ \wedge \ y_i < 0.6 \cdot Y_{Median}\}$ is the median equivalized income of persons below the at-risk-of-poverty threshold, and $p'_{i'} = \frac{z_{i'}}{\sum_{i:\, y_i < 0.6 \cdot Y_{Median}} z_i}$.

- Inequality (Income quintile share ratio - the ratio of total income received by the 20% of the population with the highest income to that received by the 20% of the population with the lowest income):

$$Y_{Inequality} = \frac{\sum_{i \in top20\%} TDHI_i}{\sum_{i \in bottom20\%} TDHI_i} \tag{11}$$

where $i \in top20\%$ and $i \in bottom20\%$ mean such households (i) for which $\sum_{i'=1}^{i} \frac{z_{i'}}{\sum_{i=1}^{N} z_i} > 0.8$ and $\sum_{i'=1}^{i} \frac{z_{i'}}{\sum_{i=1}^{N} z_i} < 0.2$, respectively (assuming units are sorted in non-decreasing order with respect to y, not the $TDHI$).

- Gini (Gini coefficient; For the household data, the Gini coefficient is calculated using the trapezoid rule, because the household data are treated here as grouped data of individuals):

$$Y_{Gini} = 1 - \sum_{i=1}^{N} \left(\sum_{i'=0}^{i} \frac{z_{i'} y_{i'}}{Y_{total}} + \sum_{i'=0}^{i-1} \frac{z_{i'} y_{i'}}{Y_{total}} \right) \cdot p_i \tag{12}$$

where $Y_{total} = \sum_{i=1}^{N} z_i y_i$, $p_i = \frac{z_i}{\sum_{i=1}^{N} z_i}$, $z_0 = 0, y_0 = 0$.

The sociodemographic variables utilized as auxiliaries concerned the household as a whole or the person completing the questionnaire. These variables are as follows (numbers in parentheses indicate the number of categories): region (7), degree of urbanization (3), tenure status (2), educational attainment level (5), self-defined current economic status (3), household type (10), number of rooms available to household (6), sex (2), marital status (4), age at the end of the income reference period (7). All auxiliary variables are categorical (or categorized), which reflects the usual way these variables are used for weight adjustment in probability sampling. To use these variables in the models, dummy variables were created representing all but one variant from each original variable. Only additive effects were considered in the models. This yielded a total of 39 explanatory variables. The experiment examined two variants: using all variables (39) or only selected ones, most correlated with income (first 5 variables; a total of 15 binary variables).

In the original sample from EU-SILC 2022, 19,757 households were surveyed. 19,655 household left after removing the units with missing values on auxiliary variables. Since it was a probability sample, each household in the dataset came with a weight that reflected different inclusion probabilities and subsequent adjustments. These weights were used to construct the probable population of 12,633,327 households by copying each entry as many times as its weight, which was rounded to the nearest integer. This dataset was too big to run the simulations described below, so a simple random sample with replacement was drawn of size $N = 100{,}000$. This sample served as the target population in further analysis (the parameters of this population were practically the same as the weighted statistics from the original EU-SILC sample). It was assumed that this sample was representative of the general population of households in Poland in terms of relationships between variables that were exploited in the experiment.

3.2. Sampling

When selecting a non-probability sample, there is no defined randomness involved, as in probability sampling, but one could imagine the accidental nature of various activities related to the selection of respondents (e.g. the respondent visiting a website where an invitation to participate in the survey will be displayed). If it were possible to

select a non-probability sample again, with the same relevant research conditions, a different sample composition would be obtained. Therefore, when writing an algorithmic selection of a non-probability sample for a computer simulation, a probabilistic mechanism can be used. However, in such a mechanism, some kind of selection bias should be incorporated. The experiment tested two non-ignorable mechanisms: The first one was dependent directly on income (hereinafter "income-dependent"), and the second one was dependent on having a computer and Internet access (hereinafter "Internet-dependent").

In the first mechanism, first-order inclusion probabilities for all members of the target population were generated using the logistic function as follows:

$$\pi_{NP,i}^{(1)} = \frac{1}{1 + e^{-(b_0 + b_1 \cdot y_i)}} \tag{13}$$

where b_0 is an intercept chosen in such a manner that $\sum_{i=1}^{N} \pi_{NP,i}^{(1)} = n_{NP}$ for a chosen non-probability sample size n_{NP}, and b_1 is the parameter affecting the strength of the selection bias. Two values of b_1 have been tested, -1 and -0.2 (see Table 1). Both values are negative, which means that households with higher income were assigned smaller inclusion probabilities. Similar to Liu et al. (2023), the sampling scheme was based on random systematic sampling (Tillé 2006, pp. 127–128), which allowed for a fixed sample size in a sampling without replacement with unequal probabilities. The population was randomly sorted before each selection.

The second mechanism depends on owning a computer and having access to the Internet. The sampling was done as planned undercoverage - the samples were drawn only from the portion of the households that met both criteria. 79.6% of households in the target population own a computer and have Internet access. From this part of the population, the samples were drawn using simple random sampling without replacement.

Following Meng (2018), we can measure the selection bias (or data defect) for the estimation of the mean as $E_R\left[\rho_{R,y}^2\right]$, where E_R is the expected value over an R-mechanism, and $\rho_{R,y}$ is the correlation coefficient between the sample membership indicator R and the study variable y, calculated as follows:

$$\rho_{R,y} = \frac{Cov_K(R,y)}{D_K(R) \cdot D_K(y)} = \frac{\frac{1}{N}\sum_{i=1}^{N} R_i y_i - f_{NP} \cdot \bar{Y}}{\sqrt{f_{NP}(1 - f_{NP})} \cdot \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N} y_i^2 - \bar{Y}^2\right)}} \tag{14}$$

where Cov_K and D_K denote covariance and standard deviation with respect to the empirical distribution of a variable (i.e. population parameters), and $f_{NP} = n_{NP}/N$, $\bar{Y} = 1/N \sum_{i=1}^{N} y_i$. Since the sampling unit in our experiment was a household and the pa-

rameters of interest pertaining to income are always weighted by the number of household members, the coefficient (14) should also be weighted. The formula is then as follows:

$$\rho_{R,y_w} = \frac{\sum_{i=1}^{N} p_i R_i y_i - \bar{R}_w \cdot \bar{Y}_w}{\sqrt{(\bar{R}_w - \bar{R}_w^2)} \cdot \sqrt{(\sum_{i=1}^{N} p_i Y_i^2 - \bar{Y}_w^2)}} \tag{15}$$

where $p_i = z_i / \sum_{i=1}^{N} z_i$ is the weight of an element i, z_i denotes the number of household members, $\bar{R}_w = \sum_{i=1}^{N} p_i R_i$, and $\bar{Y}_w = \sum_{i=1}^{N} p_i Y_i = Y_{Mean}$.

Three selection mechanisms and two non-probability sample sizes ($n_{NP} = 1000$ or $n_{NP} = 2000$) were tested, yielding six designs with different defectiveness. For each design $M = 1000$ samples were generated, and for each sample the correlation coefficient (15) was calculated. Table 1 presents the means of (square) correlation coefficients for each sampling design. Negative mean correlation coefficients in the first two R-mechanisms were induced by negative coefficients b_1 in (13). The positive correlation for the third mechanism arose naturally as the effect of households with a computer and Internet access being more affluent. Each subsequent mechanism is by one order of magnitude less defective in terms of mean ρ_{R,y_w}^2. What is important, the defectiveness of a sampling mechanism increases with increasing sample size. For comparison, if the R-mechanism were simple random sampling (SRS), the value of $E_R[\rho_{R,y_w}^2]$ would be $1/(N-1) = 0.00001$ and $E_R[\rho_{R,y_w}] = 0$. The expected squared correlation in SRS is thus about seven times smaller than in the least defective non-probability mechanism introduced in the experiment.

Table 1. Estimated sampling design defects measured by the mean (square) correlation coefficients between sample membership indicator R and equivalized income, by R-mechanism and sample size

Selection mechanism dependent on	n_{NP}	Mean over M samples of	
		ρ_{R,y_w}	ρ_{R,y_w}^2
Income (strongly), with $b_1 = -1$	1000	-0.05269	0.00278
	2000	-0.07438	0.00554
Income (weakly), with $b_1 = -0.2$	1000	-0.01549	0.00025
	2000	-0.02185	0.00049
Having a computer and internet access	1000	0.00741	0.00007
	2000	0.01046	0.00012

Source: own calculations.

As Table 1 shows, both non-probability sampling mechanisms introduced in the experiment led to data defects. In both cases, the mechanisms depended on variables

that were not used at the estimation stage. Therefore, they were non-ignorable – the assumption for propensity score modelling was not met. In such a situation, the effectiveness of the estimation methods in reducing bias depends on how well the auxiliary variables can explain the variation in the variable responsible for selection bias.

Probability reference samples, needed for auxiliary information at the estimation stage, were drawn independently for each non-probability sample using stratified sampling with proportional allocation. The strata were formed by cross-classifying *Region* and *Degree of urbanization*, which resulted in 21 strata. The size of these samples was $n_P = 1000$.

3.3. Estimation

It was assumed that the estimation methods tested in the experiment should have similar practical properties to the estimation methods used in most probability sample surveys. There were, therefore, two criteria for selecting the methods: no need to create a separate model for various study variables and the possibility of creating a single weight vector, thus allowing for the estimation of multiple parameters for distinct target variables. An additional assumption was to use different estimation approaches discussed in Section 2. Therefore, three solutions were used in the experiment: the propensity-score adjustment method, the model-based method, and doubly robust estimation. In each method, the essential part was the formula for weights.

Several propensity-score adjustment methods have been proposed in the literature (see, e.g. Valliant and Dever, 2011; Wang, Valliant and Li, 2021; Kim and Kwon, 2024), including methods based on machine learning techniques (Rueda et al., 2023; Ferri-García et al., 2024). In this experiment, the pseudo-inclusion probabilities were estimated using the logistic regression model:

$$\hat{\pi}_{NP,\,i} = \frac{e^{\boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{x}_i{}^T\widehat{\boldsymbol{\beta}}}} \tag{16}$$

where $\boldsymbol{x}_i$ is the vector of values for auxiliary variables for unit i, and $\widehat{\boldsymbol{\beta}}$ is the vector of estimated model parameters. The model parameters were estimated using the maximum pseudo-likelihood estimation method as proposed by Chen et al. (2020). In this method, the estimates $\widehat{\boldsymbol{\beta}}$ maximise the pseudo-log-likelihood function:

$$l^*(\boldsymbol{\beta}) = \sum_{i \in s_{NP}} \boldsymbol{x}_i{}^T\boldsymbol{\beta} - \sum_{i \in s_P} \frac{1}{\pi_{P,i}} \log\left(1 + e^{\boldsymbol{x}_i{}^T\boldsymbol{\beta}}\right) \tag{17}$$

where $\pi_{P,i}$ is the inclusion probability for element i in the reference sample. The maximum is obtained by solving the score equation $\frac{\partial}{\partial \boldsymbol{\theta}} l^*(\boldsymbol{\beta}) = \sum_{i \in s_{NP}} \boldsymbol{x}_i - \sum_{i \in s_P} \frac{1}{\pi_{P,i}} \cdot \pi_{NP,i} \cdot \boldsymbol{x}_i = 0$ by means of the Newton-Raphson iterative procedure. The weights were then calculated using formula (2).

The weights for the model-based and the doubly robust methods were calculated as described in Section 2, formulas (4) and (5), respectively. The linear model assumed in the derivation of weights in those methods may not be the best choice when all auxiliary variables are of a categorical nature, as was the case in the experiment, but it allows the single weight vector to be derived. This practical advantage makes these estimation procedures resemble the production of estimates in probability surveys.

In the practice of sample surveys, the distribution of weights is also examined, and possibly some trimming is performed. The use of categorical auxiliary variables and additive models in our experiment should keep the weights in a reasonable range. Nevertheless, the effect of trimming extreme weights was also examined. The trimming procedure employed consists in setting negative or large weights to zero or to an upper bound, respectively, and spreading the total weight trimmed away among the other sample units. The upper bound was arbitrarily chosen as three times the median weight.

The estimators of parameters described in Section 3.1. were then constructed straightforwardly by plugging the weights into the parameter formulas. The differences in the formulas for estimators compared to formulas for parameters are as follows:

- Summation is done over households in a non-probability sample.
- The expanded values of household size, $\check{z}_i = z_i \cdot w_{NP,i}$, and total disposable household income, $\widetilde{TDHI}_i = TDHI_i \cdot w_{NP,i}$, replaced the observed ones.

The conditions that were controlled in the experiment yielded 144 experimental setups. They comprised 6 sampling procedures simulating non-probability sample selection (3 selection mechanisms and 2 sample sizes) and 24 estimation procedures (2 sets of auxiliary variables, 2 sources of auxiliary information, 3 methods of weights construction, and 2 options for weight trimming). Despite using methods from different approaches to estimation, all solutions were tested using the randomization paradigm (i.e. assuming constant population values and averaging over possible samples). The performance of each setup was assessed by the relative bias of estimates of a finite population parameter θ, using estimator $\hat{\theta}$, which was calculated as follows (for $M = 1000$ iterations):

$$RB(\hat{\theta}) = \frac{\frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m - \theta}{\theta} 100\% \tag{18}$$

4. Results

To assess the sampling bias induced by non-ignorable selection mechanisms, we examined naïve estimates of income distribution parameters (i.e. without applying any weighting). Table 2 presents the relative biases of naïve estimators. The bias was highest

in the case of a selection mechanism strongly dependent on income. The weaker dependence on income in most cases resulted in considerably lower bias. The lowest bias was typically observed for the mechanism dependent on possession of a computer and Internet access.

Table 2. Relative bias (%) of naïve estimates of parameters of income distribution by non-probability sampling mechanisms

Selection mechanism dependent on	n_{NP}	Relative bias (%) for estimation of parameters:					
		mean	median	poverty	depth	inequality	Gini
Income (strongly)	1000	-27.8	-22.9	22.1	29.0	-1.0	-4.7
	2000	-27.6	-22.6	21.8	27.7	-1.7	-4.8
Income (weakly)	1000	-8.0	-5.4	2.6	10.7	-4.8	-4.1
	2000	-8.0	-5.4	3.1	10.9	-4.9	-4.1
Having a computer and internet access	1000	3.7	4.3	-3.9	3.9	6.7	-1.6
	2000	3.6	4.2	-4.4	3.4	6.2	-1.7

Source: own calculations.

Notably, estimations of different parameters uncovered different patterns of bias. The sample mean and median behaved similarly. They had a strong negative bias for mechanisms dependent on income because inclusion probabilities were negatively correlated with income, and there was a positive bias for the mechanism dependent on having a computer and Internet access because this mechanism favored more affluent households. The poverty estimator (i.e. sample at-risk-of-poverty rate) showed the opposite pattern to median and mean in terms of signs. In income-dependent mechanisms, where poorer households appeared in the sample more frequently than in the population, the poverty rate was overestimated, and in the other mechanism, due to the opposite effect, the rate was underestimated. The depth estimator (i.e. sample relative median at-risk-of-poverty gap) differed from the previous ones in that its bias was always positive and the greatest of all estimators for the mechanism related to income. The inequality and Gini estimators appeared to be most resilient to the strong income-dependent selection mechanism. The Gini parameter was always underestimated.

As for the sample size, one cannot observe any distinctive differences in estimation bias between smaller and larger samples. This stands in contrast to the data defect indexes presented in Table 1, which were always higher for a larger sample size. This results from the specific nature of $\rho_{R,y}$, which is the correlation between binary variable R and study variable y. The variance of a binary variable depends on its mean, and the mean of R is the sampling fraction $f_{NP} = n_{NP}/N$. Changing the number of 1s in R changes the correlation, even though the underlying mechanism stays the same. Thus, the quantity $E_R[\rho_{R,y}^2]$ should not be interpreted in separation from the sample size

(additional information from a larger sample size is not reflected in $E_R\left[\rho_{R,y}^2\right]$ but in other components of the mean squared error; see Meng, 2018).

For each of the three selection mechanisms simulating non-probability sampling, various estimation strategies were examined. The relative bias of these strategies for each estimated parameter is presented in Figure 1. For each parameter and selection mechanism, there are 48 circles in the same color representing complex estimation methods (24 different estimation strategies for each of the two sample sizes) and, for comparison, two triangles representing naïve estimation, one for each tested sample size. The points are mapped exactly on the Y-axis but are slightly jittered on the X-axis to better visualize their position. The graphic aims to illustrate the general effect of using auxiliary information and different weighting procedures on bias. If the circles are closer to the horizontal line, which represents unbiasedness (bias = 0%), than the triangles, then using sociodemographic variables and various weighting procedures generally reduces the bias of naïve estimation.

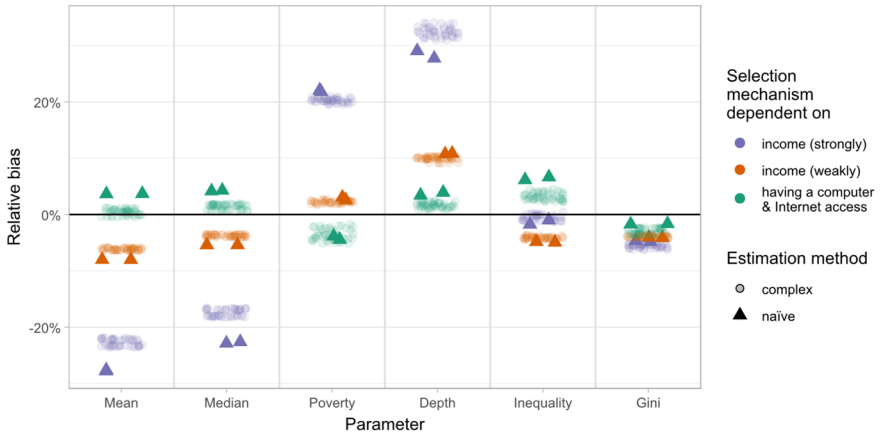

Figure 1. Relative bias (%) of naïve and complex estimates of parameters of income distribution by non-probability sampling mechanisms

Source: own calculations.

The pattern of the reduction (or increase) of the relative bias again hinged on the type of parameter estimated. For median and mean, there was always a visible improvement, irrespective of the selection mechanism. The effect was satisfactory for the Internet-dependent mechanism, sometimes reducing the bias almost to zero, but insufficient for income-dependent mechanisms. For poverty and inequality estimators, the bias was also reduced on average, but to a lesser extent. The strangest pattern was exhibited by the depth estimates. The least biased naïve estimation in the Internet-related mechanism was slightly corrected by weighting; the same applied to the estimation in the

weaker form of the income-dependent mechanism. However, in the stronger form of the income-dependent mechanism, weighting always increased the already large bias of naïve estimates for the depth parameter. The adverse effect of weighting was also present in the estimation of the Gini coefficient.

The 144 estimation strategies examined were the result of the combination of all variants of six controlled factors. To assess the impact of a particular factor, the differences in the absolute value of relative bias between complex and corresponding naïve estimations were calculated and averaged separately over specific factors. The results of these operations are presented in Table 3. Negative values mean improvement, and positive values mean deterioration in terms of estimation bias after the introduction of weights.

Table 3. The average difference in the absolute value of relative bias (in percentage points) between the complex and corresponding naïve estimations of parameters of income distribution by experimental conditions

Experimental condition		**The average difference in relative bias between the complex and naïve estimation of parameters:**					
name	**value**	**mean**	**median**	**poverty**	**depth**	**inequality**	**Gini**
Non-probability sample size	1000	-3.26	-3.23	-1.01	0.58	-1.64	0.67
	2000	-3.35	-3.32	-1.08	-0.16	-1.42	0.69
Number of auxiliary variables	small	-3.04	-2.91	-0.63	0.27	-1.65	0.75
	large	-3.57	-3.63	-1.46	0.15	-1.41	0.61
Source of auxiliary information	population	-3.30	-3.27	-1.06	0.10	-1.56	0.68
	reference sample	-3.31	-3.27	-1.03	0.33	-1.50	0.68
Estimation method	PS	-3.16	-3.10	-0.89	0.04	-1.51	0.67
	SM	-3.32	-3.36	-1.23	0.44	-1.48	0.74
	DR	-3.43	-3.36	-1.01	0.16	-1.59	0.63
Weight trimming	no	-3.32	-3.29	-1.03	0.21	-1.57	0.69
	yes	-3.29	-3.25	-1.06	0.22	-1.48	0.67

Source: own calculations.

The type of estimated parameter differentiated the effect of changing a factor on the reduction (or increase) of estimation bias compared to naïve estimation. Increasing the non-probability sample size typically resulted in a greater reduction in bias or a smaller increase in it. The exceptions were for the estimation of inequality and Gini, but the differences are small. The larger set of auxiliary variables usually had a positive effect. The only exception was for the inequality estimation. The source of auxiliary

information (population totals known exactly or estimated from a probability reference sample) exhibited no significant effect on bias, except for the depth estimator.

The effect of the estimation method was noticeable but could be outweighed by a larger set of auxiliary variables. Nevertheless, the largest bias reduction (or smallest increase) was typically observed when the doubly robust (DR) estimation method was used. It was either the best (in four cases) or the second-best (in two cases) solution for estimation. In the case of depth estimation, the best was PS weighting, which increased the bias the least, and in the case of poverty estimation, the best was SM weighting, which decreased the bias the most.

Weight trimming had a very small and uneven effect on the bias. This is because only a few weights fell outside the specified bounds. As expected, the distributions of weights were moderately skewed. There were no negative weights for the PS method, almost none for the DR method, and only a few for the SM method. The incidence of extremely large weights was 0.5% on average.

Generally, the effects of various changes in complex estimation methods were in line with the expectations (better performance for larger sample size, larger set of auxiliary variables, and DR estimator) for the estimation of simple parameters, namely, median and mean. For other, more complex parameters, the effects were not as consistent.

5. Discussion

The use of socio-demographic variables for weighting sampled units in a non-ignorable selection mechanism mostly reduced the bias, at least to some extent, in our experiment. However, in some cases, the bias increased after weighting, regardless of the method used to construct the weights. The deterioration in estimator performance following the introduction of weights was particularly pronounced for the depth estimator under the strong income-dependent mechanism, and for the Gini estimator under the Internet-dependent and strong income-dependent mechanisms.

In the depth case, the cause of this behavior may lie in the fact that the measure is a ratio of two related quantities. Changes in both the numerator and the denominator influence its value, and even if both components move in the correct direction, the difference in relative change may cause the whole measure to shift in the other direction. In the strong income-dependent mechanism, the naïve median was heavily underestimated because the selection probabilities were negatively correlated with income. Weighting corrected this bias only partially, so the weighted median remained far below the true population median. The median equivalized income of persons below 60% of the overall median was therefore estimated using an unduly small fraction of

poor households. As a result, the median income of poor households was adjusted upward disproportionally less than the overall median. This increased the gap between medians (the numerator) more than the overall median (the denominator). In the selection mechanism weakly dependent on income, these rates of adjustment were favorable because the naïve median was much less underestimated, and the weighting increased it relatively less as well.

In the Gini case, the unfavorable effects of weighting may stem from the grouped nature of the household data and from certain relationships between variables. Regarding the former, in the EU-SILC survey, we had information on household income rather than individual income, which implies that household composition affected the value of the Gini coefficient. The same individuals grouped into larger households would always yield a smaller Gini coefficient (the trapezoid rule used for grouped data inflates the area under the Lorenz curve, and this inflation increases as the number of groups/trapezes decreases). As for the latter, there was a nonlinear relationship in the population between the number of household members and the equivalized income: starting from one-person households, income increased, peaked for three-person households, and then declined. The overall linear correlation coefficient was very close to zero. However, this relationship differed between sampled and non-sampled units. In the selection mechanism strongly dependent on income, the sample correlation between household size and equivalized income was always positive and not negligible (0.11 on average). Under the Internet-dependent mechanism, the sample correlation was always negative, because in the frame population (households having a computer and Internet access), this correlation was negative (-0.13), while in the remainder of the population it was positive (0.21).

The interplay of these two factors was likely responsible for the downward bias of the weighted Gini estimator. In samples drawn under the income-dependent mechanisms, the weights were positively correlated with income (hence the upward correction of mean and median estimators), but also wealthier households tended to be slightly larger in these samples. Larger weights assigned to wealthier households increased the estimated income dispersion, but they also increased the contribution of larger households, thereby reducing the granularity of the weighted data. This household-composition effect slightly outweighed the benefits of increased dispersion, deepening the downward bias of the weighted Gini estimator. In samples from the Internet-dependent mechanism, on the other hand, the weights were negatively correlated with income (leading to reductions in the weighted mean, variance, and, more often than not, the coefficient of variation), and wealthier households tended to be smaller. Although smaller households in these samples ultimately received larger weights on average, the reduction in variability appeared to dominate, causing the estimator to decrease further on average.

6. Conclusions

Inference about finite population parameters from non-probability samples is a challenging task. The researcher must account for the possible dependencies between the selection mechanism and the variables studied, which cause selection bias. Not having the comfort of controlling the probabilities of population elements being included in the sample compels us to rely on modelling. The models should explain the selection mechanism or the studied variables, or both. The keys to the successful usage of these models, aside from knowledge about the studied phenomenon, are the reliable sources of auxiliary information about the population and a set of explanatory variables, measured in both the non-probability sample and the additional source that captures the variability of at least one modelled element. Demographics and social variables are commonly used as auxiliary variables because they are relatively easy to obtain and reliable.

As the conducted experiment showed, the usage of these variables for data weighting reduces the bias for income distribution parameters, but the effect is only certain for simple parameters such as median and mean. For more complex ones, the weighting may well induce additional bias. One of the reasons this happened in the simulation was that the weighting tended to change sample composition in terms of household size, which affected the Gini estimator. Even for simple parameters, the bias was reduced to near zero only for the Internet-related selection mechanism. For mechanisms depending directly on income, the reduction was far from satisfactory. Generally, the bias of estimates of the mean and median predictably reacted to changes in estimation strategy, showing better performance with a larger sample size, a larger set of auxiliary variables, and the use of the doubly robust estimator. However, for other, more complex parameters, the effects were less consistent.

The results of the experiment provide valuable insight into how the bias of different non-probability selection mechanisms and for different parameters is affected by various estimation practices. The experiment was carried out on a large representative sample of a real population. The computer simulation tested realistic scenarios in which sampled observations were not exchangeable, and the estimation of many parameters was done using a single vector of weights. Nevertheless, the experiment has limitations that should be mentioned. The study examined the effect of the sample selection and the estimation methods, but did not account for the change in the mode of the survey – non-probability sample surveys are often carried out online, while data from the EU-SILC 2022 survey was collected face to face or by phone. Furthermore, the EU-SILC data were treated as real observations, but in practice, a significant proportion of income values were imputed.

References

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. and Tourangeau, R., (2013). *Report of the AAPOR Task Force on Non-probability Sampling. Technical report*. Deerfield: The American Association for Public Opinion Research.

Beaumont, J.-F., (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46(1), pp. 1–29.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K., (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50(1), pp. 77–106.

Beaumont, J.-F., Rao, J. N. K., (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies. *The Survey Statistician*, 83, pp. 11–22.

Brzeziński, M., Sałach, K. and Wroński, M., (2020). Wealth inequality in Central and Eastern Europe: Evidence from household survey and rich lists' data combined. *Economics of Transition and Institutional Change*, 28(4), pp. 637–660. Available at: https://doi.org/10.1111/ecot.12257.

Buelens, B., Burger, J. and Brakel, J. A. van den, (2018). Comparing Inference Methods for Non-probability Samples. *International Statistical Review*, 86(2), pp. 322–343. Available at: https://doi.org/10.1111/insr.12253.

Carranza, R., Morgan, M. and Nolan, B., (2023). Top Income Adjustments and Inequality: An Investigation of the EU-SILC. *Review of Income and Wealth*, 69(3), pp. 725–754. Available at: https://doi.org/10.1111/roiw.12591.

Chen, J. K. T., Valliant, R. L. and Elliott, M. R., (2019). Calibrating non-probability surveys to estimated control totals using LASSO, with an application to political polling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(3), pp. 657–681. Available at: https://doi.org/10.1111/rssc.12327.

Chen, Y., Li, P. and Wu, C., (2020). Doubly Robust Inference With Nonprobability Survey Samples. *Journal of the American Statistical Association*, 115(532), pp. 2011–2021. Available at: https://doi.org/10.1080/01621459.2019.1677241.

Cochran, W.G. (1963) *Sampling techniques*. New York: Wiley.

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., Leeuw, E. D. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B. and Wenz, A., (2020). A review of conceptual approaches and empirical evidence on probability

and nonprobability sample survey research. *Journal of Survey Statistics and Methodology*, 8(1), pp. 4–36. Available at: https://doi.org/10.1093/jssam/smz041.

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

Elliott, M. R., Valliant, R., (2017) Inference for Nonprobability Samples. *Statistical Science*, 32(2), pp. 249–264.

European Commission, (2022). *Methodological guidelines and description of EU-SILC target variables*. Eurostat. Available at: https://ec.europa.eu/eurostat/documents/203647/16993001/Methodological+guidelines+2022+operation+v7.pdf/ec6bc779-6462-34aa-a8bd-256f8af34d31?t=1703153300474 (Accessed: 16 July 2024).

Eurostat, (2024). *EU-SILC Microdata*. Available at: https://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions.

Ferri-García, R., Castro-Martín, L. and Rueda, M. del M., (2021). Evaluating Machine Learning methods for estimation in online surveys with superpopulation modeling. *Mathematics and Computers in Simulation*, 186, pp. 19–28. Available at: https://doi.org/10.1016/j.matcom.2020.03.005.

Ferri-García, R., Rueda-Sánchez, J. L., Rueda, M. del M. and Cobo, B., (2024). Estimating response propensities in nonprobability surveys using machine learning weighted models. *Mathematics and Computers in Simulation*, 225, pp. 779–793. Available at: https://doi.org/10.1016/j.matcom.2024.06.012.

Gelman, A., Little, T. C., (1997). Poststratification into Many Categories Using Hierarchical Logistic Regression. *Survey Methodology*, 23(2), pp. 127–135.

GUS, (2023). *Incomes and living conditions of the population of Poland – report from the EU-SILC survey of 2021*. Warszawa: GUS.

Hansen, M.H. and Hurwitz, W.N. (1943) On the theory of sampling from finite populations. *Annals of Mathematical Statistics*, 14(4), pp. 333–362. Available at: https://doi.org/10.2307/2235923.

Horvitz, D. G., Thompson, D. J., (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), pp. 663–685.

Kalton, G., (2023). Probability vs. Nonprobability Sampling: From the Birth of Survey Sampling to the Present Day. *Statistics in Transition new series*, 24(3). Available at: https://doi.org/10.59170/stattrans-2023-029.

Kim, J. K., Kwon, Y., (2024). Comments on "Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples". *Survey Methodology*, 50(1), pp. 57–63.

Kim, J. K., Morikawa, K., (2023). An empirical likelihood approach to reduce selection bias in voluntary samples, arXiv Available at: https://doi.org/10.48550/arXiv.2211.02998.

Kim, J. K., Park, S., Chen, Y. and Wu, C., (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, 184(3), pp. 941–963. Available at: https://doi.org/10.1111/rssa.12696.

Kish, L., (1995). *Survey sampling*. New York: John Wiley & Sons.

Lavrakas, P. J., Pennay, D., Neiger, D. and Phillips, B., (2022). Comparing Probability-Based Surveys and Nonprobability Online Panel Surveys in Australia: A Total Survey Error Perspective. *Survey Research Methods*, 16(2), pp. 241–266. Available at: https://doi.org/10.18148/srm/2022.v16i2.7907.

Lee, S., Valliant, R., (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. *Sociological Methods & Research*, 37(3), pp. 319–343. Available at: https://doi.org/10.1177/0049124108329643.

Li, Y., (2024). Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples. *Survey Methodology*, 50(1), pp. 37–55.

Liu, A.-C., Scholtus, S. and De Waal, T., (2023). Correcting Selection Bias in Big Data by Pseudo-Weighting. *Journal of Survey Statistics and Methodology*, 11(5), pp. 1181–1203. Available at: https://doi.org/10.1093/jssam/smac029.

Liu, Z., Wang, D. and Pan, Y., (2025). Superpopulation model inference for non probability samples under informative sampling with high-dimensional data. *Communications in Statistics – Theory and Methods*, 54(5), pp. 1370–1390. Available at: https://doi.org/10.1080/03610926.2024.2335543.

Lohr, S. L., (2022). *Sampling. Design and analysis*. Third edition. London, New York: CRC Press.

Marella, D., (2023). Adjusting for Selection Bias in Nonprobability Samples by Empirical Likelihood Approach. *Journal of Official Statistics*, 39(2), pp. 151–172. Available at: https://doi.org/10.2478/jos-2023-0008.

Meng, X. L., (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 us presidential election. *Annals of Applied Statistics*, 12(2), pp. 685–726. Available at: https://doi.org/10.1214/18-AOAS1161SF.

Mercer, A. W., Kreuter, F., Keeter, S. and Stuart, E. A., (2017). Theory and Practice in Nonprobability Surveys: Parallels between Causal Inference and Survey Inference. *Public Opinion Quarterly*, 81(S1), pp. 250–271. Available at: https://doi.org/10.1093/poq/nfw060.

Mercer, A. W., Lau, A., (2023). Comparing Two Types of Online Survey Samples, *Pew Research Center Methods*, 7 September. Available at: https://www.pewresearch.org/methods/2023/09/07/comparing-two-types-of-online-survey-samples/ (Accessed: 15 August 2024).

Neyman, J., (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), pp. 558–625. Available at: https://doi.org/10.2307/2342192.

Rivers, D., (2006). *Understanding People. Sample Matching*. YouGovPolimetrix. Available at: http://www.websm.org/db/12/16528/Web%20Survey%20Bibliography/Understanding_people_Sample_matching/ (Accessed: 26 April 2023).

Rueda, M. del M., Pasadas-del-Amo, S., Rodríguez, B. C., Castro-Martín, L. and Ferri-García, R., (2023). Enhancing estimation methods for integrating probability and nonprobability survey samples with machine-learning techniques. An application to a Survey on the impact of the COVID-19 pandemic in Spain. *Biometrical Journal*, 65(2), p. 2200035. Available at: https://doi.org/10.1002/bimj.202200035.

Särndal, C.-E., Swensson, B. and Wretman, J. H., (1997). *Model assisted survey sampling*. New York: Springer.

Szreder, M., Kozłowski, A., (2024). *Wnioskowanie na podstawie prób losowych i nielosowych*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.

Tillé, Y., (2006). *Sampling algorithms*. New York: Springer.

Valliant, R., (2020). Comparing Alternatives for Estimation from Nonprobability Samples. *Journal of Survey Statistics and Methodology*, 8(2), pp. 231–263. Available at: https://doi.org/10.1093/jssam/smz003.

Valliant, R., Dever, J. A., (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. *Sociological Methods & Research*, 40(1), pp. 105–137. Available at: https://doi.org/10.1177/0049124110392533.

Valliant, R., Dever, J. A. and Kreuter, F., (2013). *Practical tools for designing and weighting survey samples*. New York: Springer.

Valliant, R., Dorfman, A. H. and Royall, R. M., (2000). *Finite population sampling and inference: A prediction approach*. New York: John Wiley & Sons.

Wang, L., Valliant, R. and Li, Y., (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40(24), pp. 5237–5250. Available at: https://doi.org/10.1002/sim.9122.

Wang, W., Rothschild, D., Goel, S. and Gelman, A., (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3), pp. 980–991. Available at: https://doi.org/10.1016/j.ijforecast.2014.06.001.

Wiśniowski, A., Sakshaug, J. W., Ruiz, D. A. P. and Blom, A. G., (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8(1), pp. 120–147. Available at: https://doi.org/10.1093/jssam/smz051.

Yang, S., Kim, J. K. and Song, R., (2020). Doubly Robust Inference when Combining Probability and Non-Probability Samples with High Dimensional Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(2), pp. 445–465. Available at: https://doi.org/10.1111/rssb.12354.

Zhang, L. C., (2019). On valid descriptive inference from non-probability sample, *Statistical Theory and Related Fields*, 3(2), pp. 103–113.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 113–134, https://doi.org/10.59139/stattrans-2026-030
Received – 17.02.2025; accepted – 19.01.2026

Kumaraswamy family of generalized Type II Exponentiated Half Logistic-G distribution: properties and statistical inference

Fastel Chipepa[1], **Wilbert Nkomo**[2], **Nyika Mtemeri**[3], **Takesure Nyakuamba**[4]

Abstract

We present a new family of distributions (FDs) named the Kumaraswamy type II exponentiated half logistic-G (K-TIIEHL-G). Notably, this FDs can be represented as an infinite linear combination of exponentiated-G densities enabling the derivation of important statistical properties. The density and hazard rate geometries were investigated for some special cases. The model parameters were determined using the maximum likelihood technique and Monte Carlo simulation experiments were carried out to validate the consistency of the model parameters. The Kumaraswamy type II exponentiated half logistic-Weibull (K-TIIEHL-W), a member of the K-TIIEHL-G family, was applied to two sets of real data. The results indicated that the K-TIIEHL-W model significantly outperformed several established contending equi-parameter models in terms of data fitting.

Key words: Kumaraswamy-G, Type II Exponentiated Half logistic-G, Exponentiated-G, Maximum Likelihood Estimation, Monte Carlo Simulations.

1. Introduction

Not all real-world lifespan data can be adequately modeled using existing distributions. To enhance flexibility, classical distributions often require modifications through the inclusion of one or more parameters. These parameters help address kurtosis and skewness, leading to what are termed generalized or modified distributions, which offer improved flexibility in data modeling. Cordeiro and de Castro (2011) developed the Kumaraswamy-G FDs with the cumulative distribution function (cdf) given by

$$F(x) = 1 - [1 - G^c(x;\eta)]^\lambda \tag{1}$$

and probability distribution function (pdf)

$$f(x) = c\lambda g(x;\eta) G^{c-1}(x;\eta) [1 - G^c(x;\eta)]^{\lambda-1}, \tag{2}$$

[1]Botswana International University of Science and Technology, Botswana. E-mail: chipepaf@biust.ac.bw. ORCID: https://orcid.org/0000-0001-6854-8740.
[2]Manicaland State University of Applied Sciences, Zimbabwe & Botswana International University of Science and Technology, Botswana. E-mail: wilbert.nkomo@staff.msuas.ac.zw. ORCID: https://orcid.org/0009-0006-0277-3981.
[3]Avance Clinical, Firle South Australia, Australia. E-mail: nmtemeri@gmail.com. ORCID: https://orcid.org/0000-0003-4792-8757.
[4]Manicaland State University of Applied Sciences, Zimbabwe. E-mail: takesure.nyakuamba@staff.msuas.ac.zw. ORCID: https://orcid.org/0009-0009-1682-0306.

respectively, where $G(x;\eta)$ is the parent distribution, $g(x;\eta) = \frac{dG(x;\eta)}{dx}$, η is the parent distribution vector of parameters, and c and λ are two positive shape parameters. It is worthy noting that, when $c = \lambda = 1$, the parent distribution is obtained. The supplementary parameters c and λ are intended to control the skewness and tail weight of the resulting distribution.

The Kumaraswamy FDs is a flexible and valuable tool for modeling data bounded on the interval [0,1]. It features closed-form expressions for its cumulative and probability density functions, facilitating straightforward computations and statistical inference. Additionally, the Kumaraswamy distribution is robust, effectively handling various types of skewness and kurtosis, and is widely used in fields like finance, hydrology, and environmental studies, particularly in Bayesian statistics. These advantages have inspired numerous researchers to extend the Kumaraswamy generator, leading to significant contributions such as the Kumaraswamy half logistic introduced by Usman, Ahsan-ul-Haq and Talib (2017), the Kumaraswamy Weibull by Cordeiro, Ortega and Nadarajah (2010), and the Topp-Leone Kumaraswamy-G distribution proposed by Ibrahim et al. (2020). Additionally, notable advancements include the Topp-Leone Kumaraswamy Fréchet by Bashiru (2024), Kumaraswamy odd Lindley-G by Chipepa, Oluyede and Makubate (2019) and the Kumaraswamy generalized power Lomax distribution by Nagarjuna, Vardhan and Vardhan (2021), among others. These advancements significantly enhance distribution theory by providing more flexible models that can accommodate diverse data behaviors, thereby improving the accuracy of statistical inference in various applications.

In this article, we propose a novel FDs called the K-TIIEHL-G. Our motivations for developing this distribution include:

- developing a new generalized distribution that, effectively accommodates variations in skewness, and kurtosis across a wide range of data types.
- generating various hazard rate geometries.
- fostering new insights in statistical modeling, paving way for future research and applications that require robust and adaptable methodologies.

We will structure our paper as follows: Section 2 introduces the new FDs, including a discussion on notable special cases. In Section 3, we will examine the statistical properties of the proposed FDs. Section 4 focuses on estimation of the unknown parameters using the maximum likelihood technique. In Section 5, we present the results obtained from Monte Carlo simulation experiments. Section 6 will showcase the practical applications of the proposed model using empirical data. Finally, Section 7 will summarize our findings and discuss their broader implications.

2. The new family of distributions

We develop the Kumaraswamy type II exponentiated half logistic-G (K-TIIEHL-G) FDs in this section, including an analysis of some specific special cases.

2.1. The K-TIIEHL-G distribution

Al-Mofleh et al. (2020) proposed the type II exponentiated half logistic-G (TIIEHL-G) FDs with cdf

$$F(x) = 1 - \left[\frac{1 - G^{\delta}(x,\boldsymbol{\varpi})}{1 + G^{\delta}(x,\boldsymbol{\varpi})}\right]^{\alpha} \tag{3}$$

and pdf

$$f(x) = \frac{2\alpha\delta g(x;\boldsymbol{\varpi})G^{\delta-1}(x;\boldsymbol{\varpi})}{(1+G^{\delta}(x;\boldsymbol{\varpi}))^{\alpha+1}}\left(1 - G^{\delta}(x;\boldsymbol{\varpi})\right)^{\alpha-1}, \tag{4}$$

where $\alpha, \delta > 0$ are both shape parameters, and $\boldsymbol{\varpi}$ defines the vector of parameters from the parent distribution $G(.)$.

Replacing the parent cdf in Equation (1) with the TIIEHL-G distribution and setting $\delta = 1$ yields a new FDs called K-TIIEHL-G FDs with cdf given by

$$F(x) = 1 - \left[1 - \left(1 - [K_G(x,\boldsymbol{\varpi})]^{\alpha}\right)^{c}\right]^{\lambda}, \tag{5}$$

for $\alpha, c, \lambda > 0$ where $K_G(x,\boldsymbol{\varpi}) = \frac{\bar{G}(x,\boldsymbol{\varpi})}{1+G(x,\boldsymbol{\varpi})}$ and $\bar{G}(x,\boldsymbol{\varpi}) = 1 - G(x,\boldsymbol{\varpi})$. Sub-models of the K-TIIEHL-G can be obtained by setting individual parameters to unit. The pdf and hazard rate function (hrf) of the K-TIIEHL-G FDs are respectively given by

$$\begin{aligned} f(x) &= \frac{2\alpha c\lambda g(x,\boldsymbol{\varpi})\bar{G}^{\alpha-1}(x,\boldsymbol{\varpi})}{[1+G(x,\boldsymbol{\varpi})]^{\alpha+1}}\left(1 - [K_G(x,\boldsymbol{\varpi})]^{\alpha}\right)^{c-1} \\ &\times \left[1 - \left(1 - [K_G(x,\boldsymbol{\varpi})]^{\alpha}\right)^{c}\right]^{\lambda-1} \end{aligned} \tag{6}$$

and

$$\begin{aligned} h(x) &= \frac{2\alpha c\lambda g(x,\boldsymbol{\varpi})\bar{G}^{\alpha-1}(x,\boldsymbol{\varpi})}{[1+G(x,\boldsymbol{\varpi})]^{\alpha+1}}\left(1 - [K_G(x,\boldsymbol{\varpi})]^{\alpha}\right)^{c-1} \\ &\times \left[1 - \left(1 - [K_G(x,\boldsymbol{\varpi})]^{\alpha}\right)^{c}\right]^{-1}, \end{aligned} \tag{7}$$

where $\alpha, c, \lambda > 0$, and parent parameter vector $\boldsymbol{\varpi}$.

2.2. Model identifiability

Identifiability refers to the ability to uniquely determine model parameters from data, which is crucial for accurate parameter estimates and reliable inferences in statistical modeling (Cobelli and Distefano (1980)). A key challenge in this context is the redundancy of model parameters, where distinct values produce identical outputs, complicating the estimation process and potentially leading to inaccuracies and ambiguous conclusions.

We seek to evaluate the identifiability of the K-TIIEHL-G FD. To achieve this, we examine two sets of parameters, $\Delta_1 = (\alpha_1, c_1, \lambda_1, \boldsymbol{\varpi}_1)$ and $\Delta_2 = (\alpha_2, c_2, \lambda_2, \boldsymbol{\varpi}_2)$.

The likelihood ratio is defined as follows:

$$\begin{aligned} L &= \left(\frac{\alpha_1 c_1 \lambda_1}{\alpha_2 c_2 \lambda_2}\right)\left(\frac{g(x;\varpi_1)}{g(x;\varpi_1)}\right)\left(\frac{[\bar{G}(x,\varpi_1)]^{\alpha_1-1}}{[\bar{G}(x,\varpi_2)]^{\alpha_2-1}}\right)\left(\frac{[1+G(x,\varpi_1)]^{\alpha_1+1}}{[1+G(x,\varpi_2)]^{\alpha_2+1}}\right) \\ &\times \left(\frac{\left(1-[K_G(x,\varpi_1)]^{\alpha_1}\right)^{c_1-1}}{\left(1-[K_G(x,\varpi_2)]^{\alpha_2}\right)^{c_2-1}}\right)\left(\frac{\left[1-\left(1-[K_G(x,\varpi_1)]^{\alpha_1}\right)^{c_1}\right]^{\lambda_1-1}}{\left[1-\left(1-[K_G(x,\varpi_2)]^{\alpha_2}\right)^{c_2}\right]^{\lambda_2-1}}\right). \end{aligned}$$

Consider the logarithm of the likelihood, that is

$$\begin{aligned} Log(L) &= \log\left(\frac{\alpha_1 c_1 \lambda_1}{\alpha_2 c_2 \lambda_2}\right) + \log(g(x;\varpi_1)) - \log(g(x;\varpi_2)) + (\alpha_1-1)\log\left(\bar{G}(x,\varpi_1)\right) \\ &- (\alpha_2-1)\log\left(\bar{G}(x,\varpi_2)\right) + (\alpha_1+1)\log\left([1+G(x,\varpi_1)]\right) \\ &- (\alpha_2+1)\log\left([1+G(x,\varpi_2)]\right) + (c_1-1)\log\left(1-[K_G(x,\varpi_1)]^{\alpha_1}\right) \\ &- (c_2-1)\log\left(1-[K_G(x,\varpi_2)]^{\alpha_2}\right) \\ &+ (\lambda_1-1)\log\left(1-\left(1-[K_G(x,\varpi_1)]^{\alpha_1}\right)^{c_1}\right) \\ &- (\lambda_2-1)\log\left(1-\left(1-[K_G(x,\varpi_2)]^{\alpha_2}\right)^{c_2}\right). \end{aligned}$$

The derivative of the log-likelihood with respect to x is now given by

$$\begin{aligned} \frac{d\log(L)}{d(x)} &= \frac{g'(x;\varpi_1)}{g(x;\varpi_1)} - \frac{g'(x;\varpi_2)}{g(x;\varpi_2)} + \frac{(\alpha_1-1)}{\bar{G}(x,\varpi_1)}\frac{d\bar{G}(x,\varpi_1)}{dx} - \frac{(\alpha_2-1)}{\bar{G}(x,\varpi_2)}\frac{d\bar{G}(x,\varpi_2)}{dx} \\ &- \frac{(\alpha_1+1)}{1+G(x,\varpi_1)}\frac{dG(x,\varpi_1)}{dx} + \frac{(\alpha_2+1)}{1+G(x,\varpi_2)}\frac{dG(x,\varpi_2)}{dx} \\ &- (c_1-1)\frac{(1+G(x,\varpi_1))\frac{d\bar{G}(x,\varpi_1)}{dx} - \bar{G}(x,\varpi_1)\frac{dG(x,\varpi_1)}{dx}}{(1+G(x,\varpi_1))^2\left(1-[K_G(x,\varpi_1)]\right)} \\ &+ (c_2-1)\frac{(1+G(x,\varpi_2))\frac{d\bar{G}(x,\varpi_2)}{dx} - \bar{G}(x,\varpi_2)\frac{dG(x,\varpi_2)}{dx}}{(1+G(x,\varpi_2))^2\left(1-[K_G(x,\varpi_2)]\right)} \\ &- (\lambda_1-1)c_1 A + (\lambda_2-1)c_1 B, \end{aligned}$$

where $A = \frac{\left(1-[K_G(x,\varpi_1)]^{\alpha_1}\right)^{c_1-1}\alpha_1[K_G(x,\varpi_1)]^{\alpha_1-1}\frac{(1+G(x,\varpi_1))\frac{d\bar{G}(x,\varpi_1)}{dx}-\bar{G}(x,\varpi_1)\frac{dG(x,\varpi_1)}{dx}}{(1+G(x,\varpi_1))^2}}{1-\left(1-[K_G(x,\varpi_1)]^{\alpha_1}\right)^{c_1}}$ and

$B = \frac{\left(1-[K_G(x,\varpi_2)]^{\alpha_2}\right)^{c_2-1}\alpha_2[K_G(x,\varpi_2)]^{\alpha_2-1}\frac{(1+G(x,\varpi_2))\frac{d\bar{G}(x,\varpi_2)}{dx}-\bar{G}(x,\varpi_2)\frac{dG(x,\varpi_2)}{dx}}{(1+G(x,\varpi_2))^2}}{1-\left(1-[K_G(x,\varpi_2)]^{\alpha_2}\right)^{c_2}}$.

Therefore, $\frac{d\log(L)}{dx} = 0$, if $\Delta_1 = \Delta_2$, implying that the parameters can be uniquely determined.

2.3. Some Particular Models

In this subsection, specific cases of K-TIIEHL-G FDs are delineated by specifying the parent cdf and pdf in Equations (6) and (5). The Lindley, Weibull and exponential distributions serve as baseline distributions.

2.3.1 Kumaraswamy Type II Exponentiated Half logistic-Lindley (K-TIIEHL-Lin) distribution

The Lindley distribution has cdf $G(x) = 1 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}$ and pdf $g(x) = \frac{s^2}{1+s}(1 + x)e^{-sx}$, for $s, x > 0$. Utilizing the Lindley as the parent distribution, we have the K-TIIEH-Lin distribution with cdf, pdf and hrf respectively given by

$$F(x) = 1 - \left[1 - \left(1 - \left[\frac{\left(1 + \frac{sx}{1+s}\right) e^{-sx}}{2 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}}\right]^{\alpha}\right)^{c}\right]^{\lambda},$$

$$\begin{aligned} f(x) &= \frac{\frac{2\alpha c\lambda s^2}{(1+s)(1+x)} e^{-sx} \left(\left(1 + \frac{sx}{1+s}\right) e^{-sx}\right)^{\alpha-1}}{\left(2 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}\right)^{\alpha+1}} \left(1 - \left[\frac{\left(1 + \frac{sx}{1+s}\right) e^{-sx}}{2 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}}\right]^{\alpha}\right)^{c-1} \\ &\times \left[1 - \left(1 - \left[\frac{\left(1 + \frac{sx}{1+s}\right) e^{-sx}}{2 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}}\right]^{\alpha}\right)^{c}\right]^{\lambda-1}, \end{aligned}$$

and

$$\begin{aligned} h(x) &= \frac{\frac{2\alpha c\lambda s^2}{(1+s)(1+x)} e^{-sx} \left(\left(1 + \frac{sx}{1+s}\right) e^{-sx}\right)^{\alpha-1}}{\left(2 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}\right)^{\alpha+1}} \left(1 - \left[\frac{\left(1 + \frac{sx}{1+s}\right) e^{-sx}}{2 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}}\right]^{\alpha}\right)^{c-1} \\ &\times \left[1 - \left(1 - \left[\frac{\left(1 + \frac{sx}{1+s}\right) e^{-sx}}{2 - \left(1 + \frac{sx}{1+s}\right) e^{-sx}}\right]^{\alpha}\right)^{c}\right]^{-1}, \end{aligned}$$

for $\alpha, c, \lambda, s > 0$ and $x > 0$.

2.3.2 Kumaraswamy Type II Exponentiated Half logistic-Weibull (K-TIIEHL-W) distribution

Taking the one-parameter Weibull distribution whose cdf and pdf are given by $G(x) = 1 - \exp(-x^d)$, and $g(x) = dx^{d-1}\exp(-x^d)$, respectively for $x, d > 0$ as a parent distribution, we get the K-TIIEH-W distribution with cdf, pdf and hrf respectively given by

$$F(x) = 1 - \left(1 - \left(1 - [W(x;d)]^{\alpha}\right)^{c}\right)^{\lambda},$$

and pdf

$$\begin{aligned} f(x) &= 2c\alpha\lambda dx^{d-1} \left(e^{-x^d}\right)^{\alpha} \left(2 - e^{-x^d}\right)^{-\alpha-1} \left(1 - (W(x;d))^{\alpha}\right)^{c-1} \\ &\times \left(1 - \left(1 - (W(x;d))^{\alpha}\right)^{c}\right)^{\lambda-1}, \end{aligned}$$

$$\begin{aligned} h(x) &= 2c\alpha\lambda dx^{d-1} \left(e^{-x^d}\right)^{\alpha} \left(2 - e^{-x^d}\right)^{-\alpha-1} \left(1 - (W(x;d))^{\alpha}\right)^{c-1} \\ &\times \left(1 - \left(1 - (W(x;d))^{\alpha}\right)^{c}\right)^{-1}, \end{aligned}$$

where $W(x;d) = \frac{e^{-x^d}}{2-e^{-x^d}}$, for $\alpha, c, \lambda, d > 0$ and $x > 0$.

2.3.3 Kumaraswamy Type II Exponentiated Half logistic-Exponential (K-TIIEHL-E) distribution

Taking the exponential distribution with the cdf given as $G(x;a) = 1-\exp(-ax)$, and pdf $g(x;a) = a\exp(-ax)$, for $x, a > 0$ as a parent distribution, the K-TIIEH-E distribution is obtained with cdf, pdf and hrf respectively given by

$$F(x) = 1 - \left(1 - \left(1 - [E(x;a)]^{\alpha}\right)^{c}\right)^{\lambda},$$

and pdf

$$\begin{aligned} f(x) &= 2\alpha c\lambda a \left(e^{-ax}\right)^{\alpha} \left(2 - e^{-ax}\right)^{-\alpha-1} \left(1 - (E(x;a))^{\alpha}\right)^{c-1} \\ &\times \left(1 - \left(1 - (E(x;a))^{\alpha}\right)^{c}\right)^{\lambda-1}, \end{aligned}$$

$$\begin{aligned} h(x) &= 2\alpha c\lambda a \left(e^{-ax}\right)^{\alpha} \left(2 - e^{-ax}\right)^{-\alpha-1} \left(1 - (E(x;a))^{\alpha}\right)^{c-1} \\ &\times \left(1 - \left(1 - [E(x;a)]^{\alpha}\right)^{c}\right)^{-1}, \end{aligned}$$

where $E(x;a) = \frac{e^{-x^d}}{2-e^{-x^d}}$ for $\alpha, c, \lambda, a > 0$ and $x > 0$.

2.4. Geometries of the density and hazard functions

We selected the Lindley, Weibull, and exponential distributions as parent distributions for the K-TIIEHL-G model to plot the pdf and hrf of the K-TIIEHL-G model. This approach allows us to examine the diverse shapes exhibited by this FDs.

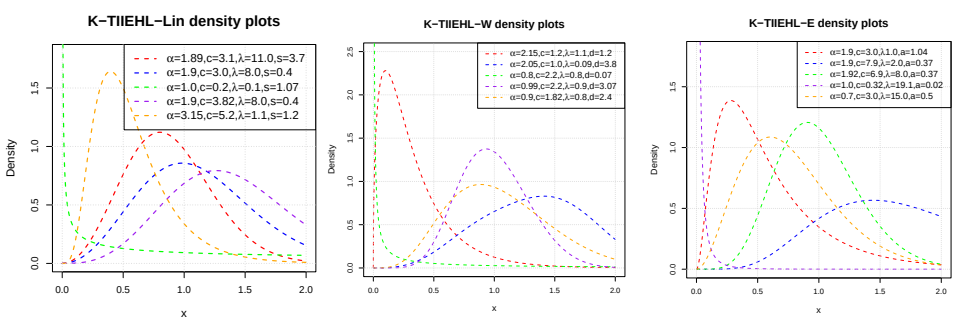

Figure 1. Density shapes for K-TIIEHL-Lin, K-TIIEHL-W and K-TIIEHL-E distributions

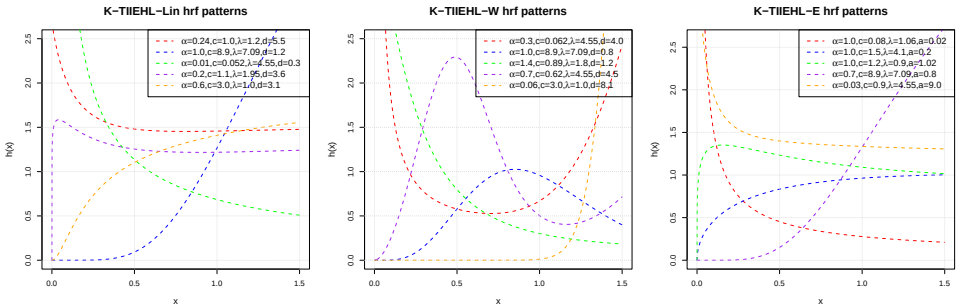

Figure 2. Hazard rate function plots of K-TIIEHL-Lin, K-TIIEHL-W and K-TIIEHL-E distributions

The plots presented in Figure [1] and Figure [2] illustrate the remarkable flexibility of this FDs. Figure [1] demonstrates a variety of density shapes, including reversed-J, positively skewed, negatively skewed, and nearly symmetric forms. Additionally, the hazard rate functions in Figure [2] encompass a range of profiles, such as bathtub, inverted bathtub, as well as both increasing and decreasing patterns. This diversity underscores the model's adaptability in capturing different statistical behaviors.

2.5. Some 3D shapes for kurtosis and skewness of special cases

This subsection showcases some 3D plots of kurtosis and skewness for the Lindley, Weibull and exponential distributions discussed Section 2.4.

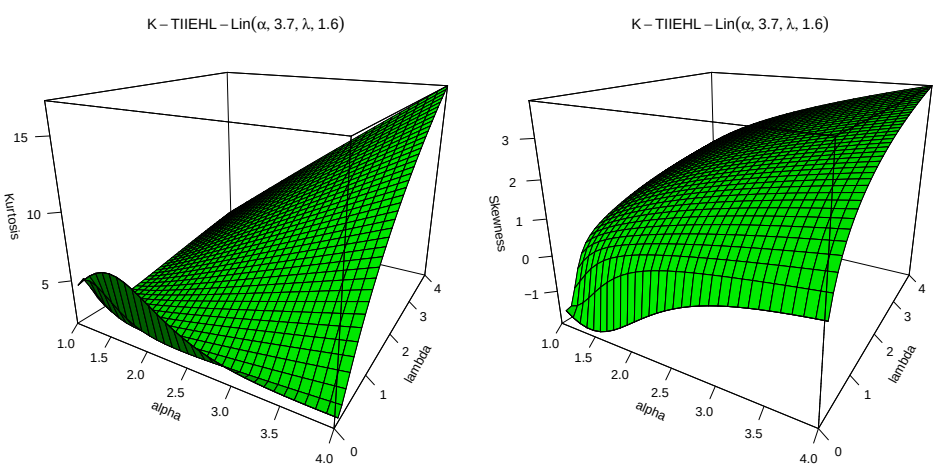

Figure 3. 3D plots of K-TIIEHL-Lin distribution

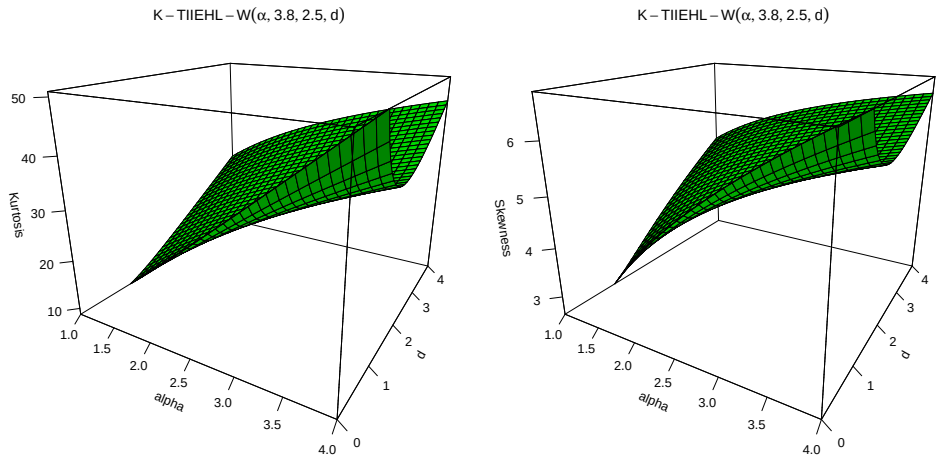

Figure 4. 3D plots of K-TIIEHL-W distribution

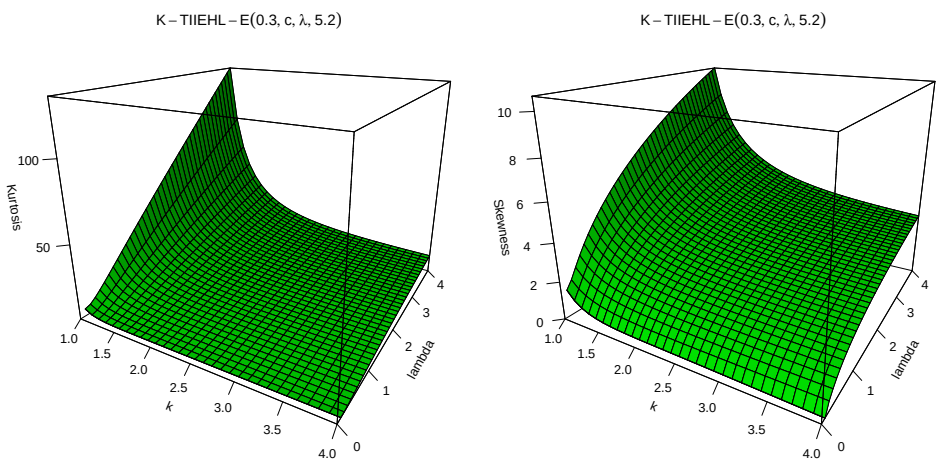

Figure 5. 3D plots of K-TIIEHL-E distribution

Figures [3], [4] and [5] shows sets of 3D plots of some kurtosis and skewness features of K-TIIEHL-G distribution for the special cases. Analyzing these plots, we observe that varying skewness and kurtosis levels are attained after fixing some selected parameter values.

3. Model properties

We seek to develop some statistical and mathematical properties associated with this novel FDs (K-TIIEHL-G) in this section.

3.1. Quantile function (Qfn)

The Qfn plays a crucial role in data transformation, underpinning techniques such as random number generation, quantile normalization, and quantile regression. Its importance lies in its ability to facilitate a variety of statistical analyses and data manipulation methods.

The Qfn of the K-TIIEHL-G FDs can be obtained by inverting the K-TIIEHL-G cdf, that is

$$u = 1 - \left[1 - \left(1 - \left[\frac{\bar{G}(Q_X(u))}{1 + G(Q_X(u))}\right]^{\alpha}\right)^{c}\right]^{\lambda}$$

for $0 \leq u \leq 1$, utilizing statistical software like R, for a specified baseline cdf $G(.)$. Consequently, we obtain the quantile values by solving the non-linear function,

$$Q_X(u) = G^{-1}\left[\frac{1 - \left[1 - (1-u)^{\frac{1}{\lambda}}\right]^{\frac{1}{c}}}{1 + \left[1 - (1-u)^{\frac{1}{\lambda}}\right]^{\frac{1}{c}}}\right]. \tag{8}$$

3.2. Density expansion

The pdf of the K-TIIEHL-G FDs can be written as

$$f(x) = \sum_{m=0}^{\infty} \Upsilon_{m+1} g_{m+1}(x;\varpi),$$

where $g_{m+1}(x;\varpi) = (m+1)g(x;\varpi)G^{m}(x;\varpi)$ is the exponentiated-G (Expo-G) distribution with power parameter $(m+1)$ and

$$\begin{aligned}\Upsilon_{m+1} &= 2\alpha c\lambda \sum_{i,j,k,l=0}^{\infty} \frac{(-1)^{i+l+j+m}}{m+1} \binom{\lambda-1}{i} \binom{c(i+1)-1}{j} \binom{k+\alpha(j+1)}{k} \\ &\times \binom{k}{l} \binom{\alpha+l-1}{m}\end{aligned} \tag{9}$$

is the linear component. Therefore, the K-TIIEHL-G FDs is expressible as an infinite linear combination of the Expo-G densities from which we can directly derive other statistical properties. Visit the *appendix* for detailed derivations.

3.3. Order statistics

One of the valuable concepts in statistical sciences is order statistics, which possess a diverse array of applications. They are used in car races, modeling auctions, and insurance policies, as well as in optimizing production processes and estimating distribution

parameters. The pdf of the i^{th} order statistics from the K-TIIEHL-G FDs is given by

$$\begin{aligned} f_{i:n}(x) &= \frac{f(x)}{B(i,n-i+1)}\sum_{j=0}^{n-i}\binom{n-i}{j}[F(x)]^{j+i-1} \\ &= \frac{1}{B(i,n-i+1)}\sum_{j=0}^{n-i}\sum_{r=0}^{\infty}\binom{n-i}{j}\Upsilon_{r+1}g_{r+1}(x;\varpi), \end{aligned}$$

where, $B(.,.)$ represents the beta function, $g_{r+1}(x,\varpi) = (r+1)g(x,\varpi)G^r(x,\varpi)$ is the Expo-G distribution, $(r+1)$ is the power parameter and

$$\begin{aligned} \Upsilon_{r+1} &= \alpha\beta\theta^2\sum_{k,l,m,q,p,r=0}^{\infty}(-1)^{k+l+m+q+r}(1-\theta)^p\binom{i+j-1}{k} \\ &\times \binom{\beta(k+1)-1}{l}\binom{p}{q}\binom{q(\theta-1)}{r}\left(\frac{1}{r+1}\right). \end{aligned}$$

In light of this, we draw the conclusion that the pdf of the i^{th} order statistic from the K-TIIEHL-G FDs is expressible as an infinite linear mixture of the Expo-G densities.

3.4. Moments and incomplete moments

Moments are vital statistical tools that reveal insights into the shape, variability, and distributional properties of data, going beyond mere central tendency measures. Incomplete moments provide essential details often missed by aggregate statistics, making them particularly relevant for analyzing complex socio-economic issues such as inequality. If Y_{r+1} is an Expo-G distributed random variable (rv), then, the q^{th} moment of the K-TIIEHL-G FDs is

$$E(Z^q) = \sum_{r=0}^{\infty}\Upsilon_{r+1}E(Y_{r+1}^q),$$

where $E(Y_{r+1}^q)$ represents the q^{th} moment of Y_{r+1} and, Υ_{r+1} is specified by Equation (9). We define the q^{th} incomplete moments as

$$I_Z(t) = \int_0^t z^q f(x)dx = \sum_{r=0}^{\infty}\Upsilon_{r+1}I_{r+1}(t),$$

where $I_{r+1}(t) = \int_0^t z^q g_{r+1}(x;\varpi)dx$. We state the moment generating function (mgf) of Z as

$$M_Z(t) = \sum_{r=0}^{\infty}\Upsilon_{r+1}E(e^{tY_{r+1}}),$$

where Υ_{r+1} is defined in Equation (9) and $E(e^{tY_{r+1}})$ is the mgf of the Expo-G distribution. Note that Lorenz and Bonferroni curves for the K-TIIEHL-G can be found from incomplete moments.

3.5. Likelihood ratio order

The likelihood ratio (LR) is essential for model selection using information criteria. By incorporating stochastic ordering, particularly in machine learning models, one can enhance predictive accuracy and generalization, especially in fields where the ordinal nature of the data is significant. Consider X and Y to be continuous rvs with pdfs $f(x)$ and $g(x)$, respectively. If the function $z \to \frac{g(x)}{f(x)}$ is increasing across the combined supports of X and Y, we say X is less than Y in the lr order, denoted as $X \leq_{lr} Y$ (Shaked and Shanthikumar (2007)).

Proposition: If $X \sim K-TIIEHL-G(\alpha,c,\lambda_1,\varpi)$ and $X \sim K-TIIEHL-G(\alpha,c,\lambda_2,\varpi)$, then $Y \leq_{lr} X$ provided $b_1 < b_2$ and $Y \leq_{lr} Z$ provided $b_2 < b_1$.
Proof: Let f_X and f_Y be densities to X and Y, respectively. From Equation (6),

$$\frac{f_X(\alpha,c,\lambda_1,\varpi)}{f_Y(\alpha,c,\lambda_2,\varpi)} = \frac{\lambda_1}{\lambda_2}\left[1-\left(1-[K_G(x,\varpi)]^{\alpha}\right)^{c}\right]^{\lambda_1-\lambda_2}. \tag{10}$$

Differentiating Equation (10) with respect to x, we have

$$\frac{\frac{f_X(\alpha,c,\lambda_1,\varpi)}{f_Y(\alpha,c,\lambda_2,\varpi)}}{\partial x} = \frac{\lambda_1(\lambda_1-\lambda_2)}{\lambda_2}\left[1-\left(1-[K_G(x,\varpi)]^{\alpha}\right)^{c}\right]^{\lambda_1-\lambda_2-1}$$

increasing in x for $\lambda_1 < \lambda_2$. Consequently, $Y \leq_{lr} X$. Likewise, it is proven that $X \leq_{lr} Y$ if $\lambda_2 < \lambda_1$. It follows therefore that $X \leq_{st} Y$, where (st) denotes stochastic order.

3.6. Rényi entropy

K-TIIEHL-G FDs' Rényi entropy (Rényi (1960)) can be written as

$$I_R(z) = \frac{\log\left[\sum_{j=0}^{\infty}\Lambda_l e^{[(1-\omega)I_{REG}]}\right]}{1-\omega},$$

for $\omega \neq 1, \omega > 0$, where

$$\begin{aligned}\Lambda_l &= (\alpha\beta\theta^2)^{\omega}\sum_{e,f,h,k,l=0}^{\infty}(-1)^{e+f+k+l}\binom{\omega(\lambda-1)}{e}\binom{\omega(c-1)+ce}{f}\\ &\times \binom{\omega([\alpha+1)]+\alpha f+h-1}{h}\binom{h}{k}\binom{\omega(\alpha-1)+k}{l}\left(\frac{l}{\omega}+1\right)^{-\omega},\end{aligned}$$

and $I_{REG} = \frac{1}{1-\omega}\int_0^{\infty}\left[\left(\left(\frac{l}{\omega}+1\right)[g(x,\varpi)G^{\frac{l}{\omega}}(x,\varpi)]\right)^{\omega}\right]dx$ represents the RE of the Expon-G distribution with power parameter $(\frac{l}{\omega}+1)$. As a result, the Rényi entropy RE of the K-TIIEHL-G FDs stems directly from Rényi entropy of the Expo-G distribution. For comprehensive derivations refer to the *appendix* .

4. Estimation

Consider a random sample of size n, where each observation, X_i, follows the K-TIIEHL-G distribution, and let $\Delta = (\alpha, c, \lambda, \Psi)^T$ represent the vector of model parameters. The log-likelihood function $\ell = \ell(\Delta)$ for this sample can be written as

$$\begin{aligned}\ell(\Delta) &= n\ln(2\alpha c\lambda) + \sum_{i=1}^{n}\ln\left(g(x_i;\varpi)\right) + (\alpha-1)\sum_{i=1}^{n}\ln\left[(\bar{G}(x_i;\varpi)\right] \\ &- (\alpha+1)\sum_{i=1}^{n}\ln\left[1+G(x_i,\varpi)\right] + (c-1)\sum_{i=1}^{n}\ln\left[1-[K_G(x,\varpi)]^{\alpha}\right] \\ &+ (\lambda-1)\sum_{i=1}^{n}\ln\left[1-\left(1-[K_G(x,\varpi)]^{\alpha}\right)^{c}\right]. \end{aligned} \tag{11}$$

The numerical solution of the nonlinear equations $\left[\frac{\partial\ell}{\partial\alpha}, \frac{\partial\ell}{\partial c}, \frac{\partial\ell}{\partial\lambda}, \frac{\partial\ell}{\partial\varpi_k}\right]^T = \mathbf{0}$ using iterative methods in R for a specific baseline cdf $G(x;\varpi)$, yields the MLEs of α, c, λ and ϖ_k. Visit the *appendix* for individual components of the score vector.

5. Simulation

We seek to weigh the efficiency of MLEs by carrying out a simulation study. Table 1 gives simulation studies results. We simulated for n= 25, 60, 100, 250, 500, 1000 and 2000 for N=3000 from the K-TIIEHL-W distribution. Average bias (AvBIAS) and root mean square error (RMSErr) for an estimated parameter, for instance ($\hat{c}$), are computed as follows:

$$AvBIAS(\hat{c}) = \frac{\sum_{i=1}^{N}\hat{c}_i}{N} - c, \quad \text{and} \quad RMSErr(\hat{c}) = \sqrt{\frac{\sum_{i=1}^{N}(\hat{c}_i - c)^2}{N}},$$

respectively. Examining the data presented in Table 1, it is evident that the average estimated parameter values converge towards the true parameter values with an increasing sample size. Moreover, both the Root Mean Square Error (RMSErr) and Average Bias (AvBIAS) show a consistent decline towards zero across all parameters with larger sample sizes. This observation highlights that the K-TIIEHL-W distribution provides consistent and efficient parameter estimates.

Table 1. Results of Monte Carlo simulations for the K-TIIEHL-W distribution: Mean, RMSErr, and AvBIAS

		$\alpha = 1.1$, $c = 1.11$,$\lambda = 0.82$, $d = 1.1$			$\alpha = 1.1$, $c = 1.1$, $\lambda = 1.1$,$d = 0.62$		
	n	Mean	RMSErr	AvBIAS	Mean	RMSErr	AvBIAS
	25	1.6801	1.3076	0.5801	1.7864	1.6681	0.6864
	60	1.5704	0.9813	0.4704	1.5948	1.1871	0.4948
	100	1.4687	0.8097	0.3687	1.3970	0.8084	0.2971
α	250	1.4275	0.7076	0.3275	1.3330	0.6859	0.2330
	500	1.3194	0.5797	0.2194	1.1870	0.3480	0.0870
	1000	1.2261	0.4213	0.1261	1.1331	0.1550	0.0331
	2000	1.1815	0.2744	0.0815	1.1283	0.0997	0.0283
	25	1.2442	0.5341	0.1442	2.9936	4.9795	1.8936
	60	1.1575	0.3272	0.0925	1.6388	1.9090	0.5388
	100	1.1234	0.2425	0.0766	1.3158	1.0450	0.2158
c	250	1.0998	0.1852	0.0602	1.2267	0.7108	0.1267
	500	1.0697	0.1534	0.0338	1.1426	0.4061	0.0426
	1000	1.0162	0.1346	0.0203	1.1361	0.2702	0.0416
	2000	1.0085	0.0126	0.0094	1.1237	0.2202	0.0337
	25	1.0492	1.3722	0.2492	1.6964	2.5982	0.5964
	60	0.8763	0.6451	0.0763	1.2657	1.1870	0.1657
	100	0.8516	0.4362	0.0516	1.2080	0.6763	0.1081
λ	250	0.8445	0.3220	0.0286	1.2051	0.4791	0.1051
	500	0.8344	0.2444	0.0195	1.1989	0.3057	0.0989
	1000	0.8286	0.1935	0.0144	1.1862	0.2183	0.0862
	2000	0.8081	0.1536	0.0081	1.1373	0.1973	0.0273
	25	2.2212	4.3650	1.1212	1.4258	3.1007	0.8258
	60	1.7284	1.1997	0.6284	1.0248	0.9991	0.4248
	100	1.5857	1.0656	0.4857	0.8108	0.6889	0.2108
d	250	1.5393	0.9524	0.4393	0.7112	0.4017	0.1112
	500	1.3852	0.7719	0.2852	0.6548	0.1919	0.0548
	1000	1.2565	0.5620	0.1565	0.6265	0.0944	0.0265
	2000	1.1972	0.3739	0.0972	0.6245	0.0713	0.0245

		$\alpha = 0.83$, $c = 1.1$, $\lambda = 1.1$, $d = 0.63$			$\alpha = 0.9$, $c = 1.0$, $\lambda = 0.8$, $d = 1.2$		
	n	Mean	RMSErr	AvBIAS	Mean	RMSErr	AvBIAS
	25	1.4503	1.6864	0.6503	1.4685	1.1435	0.5685
	60	1.2697	1.0268	0.4697	1.3473	1.0018	0.4473
	100	1.0496	0.6277	0.2496	1.2106	0.9649	0.3106
α	250	0.9591	0.4094	0.1591	1.1390	0.7247	0.2390
	500	0.8828	0.2049	0.0828	1.0796	0.4094	0.1796
	1000	0.8551	0.1012	0.0551	0.9789	0.1716	0.0789
	2000	0.8510	0.0814	0.0510	0.9176	0.0380	0.0176
	25	2.0024	2.2259	1.4024	1.1118	0.2362	0.2118
	60	1.9122	1.6588	0.9122	1.0757	0.1140	0.1757
	100	1.7680	1.5527	0.6680	1.0566	0.0965	0.1466
c	250	1.6235	1.1358	0.5235	1.0419	0.0604	0.1219
	500	1.5244	0.9048	0.4244	1.0285	0.0372	0.1085
	1000	1.3954	0.6954	0.2954	1.0227	0.0166	0.0927
	2000	1.2982	0.4760	0.1982	1.0205	0.0078	0.0605
	25	1.3164	1.5108	0.2164	1.0165	0.8932	0.2069
	60	1.2170	0.4721	0.1170	0.9508	0.5928	0.2052
	100	1.2118	0.3613	0.1118	0.9332	0.4495	0.2027
λ	250	1.2012	0.2484	0.1012	0.9235	0.3788	0.1765
	500	1.1893	0.2168	0.0893	0.8573	0.3268	0.1668
	1000	1.1691	0.1764	0.0691	0.8348	0.2596	0.1492
	2000	1.1634	0.1280	0.0634	0.8131	0.2013	0.0835
	25	1.5605	2.6207	0.9605	2.1491	3.2233	0.9491
	60	1.2535	1.2710	0.6535	1.9198	1.6246	0.7198
	100	0.9593	0.8515	0.3593	1.7307	1.2020	0.5307
d	250	0.8355	0.5619	0.2355	1.6073	1.0181	0.4073
	500	0.7250	0.2926	0.1250	1.5180	0.6697	0.3180
	1000	0.6815	0.1475	0.0815	1.3500	0.4386	0.1500
	2000	0.6737	0.1162	0.0737	1.2286	0.2114	0.0486

6. Applications

This section aims to illustrate the practicality and adaptability of the K-TIIEHL-G FDs, with a specific focus on the K-TIIEHL-W distribution, which is distinguished for its superiority over various non-nested models. To assess model performance, a range of statistical measures were employed, including: $-2\log L$, Akaike Information Criterion (AIC), Consistent Akaike Information Criterion (CAIC), Bayesian Information Criterion (BIC), Cramér-von Mises (W^*), Anderson-Darling (A^*), and Kolmogorov-Smirnov (K-S) statistics, along with its corresponding p-values. The model with the lowest statistics and the highest K-S p-value is deemed most effective in capturing the data's characteristics. Parameter estimation for the K-TIIEHL-W model was conducted using the non-linear minimization function (nlm) in R software.

In our analysis, we conducted a comparative analysis of the K-TIIEHL-W distribution against five alternative non-nested models including Kumaraswamy and exponentiated half logistic extensions. This comprehensive approach highlights the robustness of the K-TIIEHL-W distribution in capturing various underlying data patterns.

The contending models (whose pdfs are provided in the *appendix*) are: the exponentiated half logistic Weibull Poisson (EHLWP) by Chipepa et al. (2022), Kumaraswamy Weibull (KW) by Cordeiro, Ortega and Nadarajah (2010), exponentiated half logistic odd Burr III-log-logistic Poisson (EHL-OBIII-LLOG) by Oluyede et al. (2022), odd exponentiated half logistic Burr XII (OEHLBXII) distribution by Aldahlan and Afify (2018), and the Topp-Leone odd Burr III log-logistic (TL-OBIII-LLOG) introduced by Moakofi, Oluyede and Gabanakgosi (2022). Parameter estimates accompanied by their respective standard errors (SEs), in parentheses, are presented in Tables 2 and 3. To ensure the uniqueness of the parameter estimates, profile plots were provided in all datasets. Furthermore, various other visual representations such as fitted densities, empirical cumulative distribution function (ECDF), Kaplan-Meier (KM) survival curves, and scaled total time on test (TTT) plots were also included.

6.1. Blood cancer data

The initial dataset includes the lifetimes (in days) of 43 blood cancer patients from a health hospital in Saudi Arabia. This data was also analyzed by Abouammoh and Abdulghani (1994) and by Elangovan and Arunkumar (2022). The data values are: 115, 181, 255, 418, 441, 461, 516, 739, 743, 789, 807, 865, 924, 983, 1025, 1062, 1063, 1165, 1191, 1222, 1222, 1251, 1277, 1290, 1357, 1369, 1408, 1455, 1478, 1519, 1578, 1578, 1599, 1603, 1605, 1696, 1735, 1799, 1815 ,1852, 1899, 1925, 1965.

The maximum likelihood estimates along with their corresponding SEs (shown in parentheses), and GoF statisstics are displayed in Table 2. The GoF results show that the K-TIIEHTL-W model surpasses its competitors.

Table 2. Estimates and GoF statistics for K-TIIEHL-W and other distributions on blood cancer data

	Estimates				GoF statistics							
Model	α	c	λ	d	$-2\log(L)$	AIC	AICC	BIC	W^*	A^*	$K-S$	$p-value$
$K-TIIEHL-W$	0.0128	6.3945	1.18×10^{3}	0.4766	660.4967	668.4967	669.5494	675.5415	0.1584	1.0159	0.1197	0.6684
	(5.13×10^{-3})	(2.94×10^{-3})	(6.27×10^{-7})	(0.0563)								
	α	β	δ	θ								
$EHL-WP$	0.5602	0.0079	1.0316	3.7292	695.6587	703.6587	704.7114	710.7114	0.1745	1.1327	0.1888	0.1014
	(0.0.0523)	(0.0.0025)	(0.1794)	(0.6659)								
	a	b	α	β								
KW	26.8610	238.2508	8.57×10^{-3}	0.2171	664.2811	672.2811	673.3338	679.3260	0.2048	1.2840	0.1395	0.3729
	(7.49×10^{-5})	(1.89×10^{-6})	(2.54×10^{-3})	(0.0253)								
	a	b	α	c								
$EHL-OBIII-LLoG$	546.0500	1.55×10^{3}	2.4683	2.45×10^{-3}	699.6970	707.6970	708.7496	714.7417	0.6931	3.8696	0.2552	0.0738
	(2.15×10^{-3})	(8.65×10^{-12})	(5.77×10^{-9})	(4.79×10^{-5})								
	α	λ	a	b								
$OEHL-BXII$	0.8791	3.76×10^{-5}	0.8240	1.7557	700.8792	708.9792	710.0312	716.0234	0.1666	1.0261	0.1787	0.1283
	(1.05×10^{-10})	(5.41×10^{-6})	(2.53×10^{-9})	(1.19×10^{-9})								
	α	β	b	λ								
$TL-OBIII-LLoG$	1.4708	135.4324	3.08645	0.5315	697.4568	705.4568	706.5094	712.5016	0.6580	3.6943	0.2563	0.0705
	(0.3493)	(0.0326)	(1.4764)	(0.1262)								

Figure [6] clearly shows that the K-TIIEHL-W parameters on blood cancer data can be uniquely determined. The 95% asymptotic confidence intervals (ACI) for the model parameters of the K-TIIEHTL-W distribution are $\alpha \in [0.0128 \pm 0.0101]$, $c \in [6.3945 \pm 0.0058]$, $\lambda \in [1.18\times10^{3} \pm 1.23\times10^{-6}]$, and $d \in [0.4766 \pm 0.1103]$, respectively.

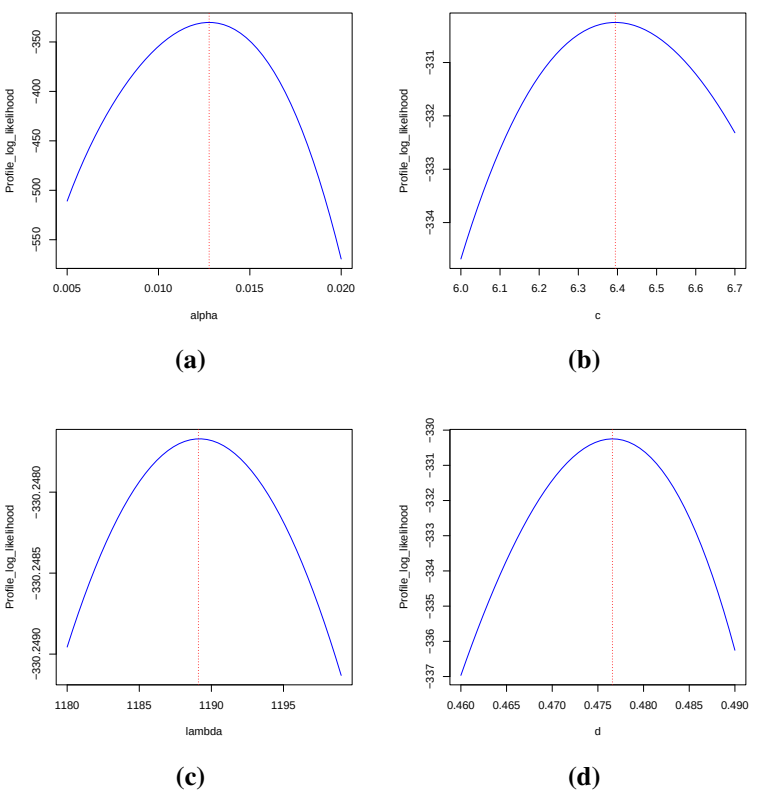

Figure 6. Profile log-likelihood plots of the K-TIIEHL-W: Blood cancer data

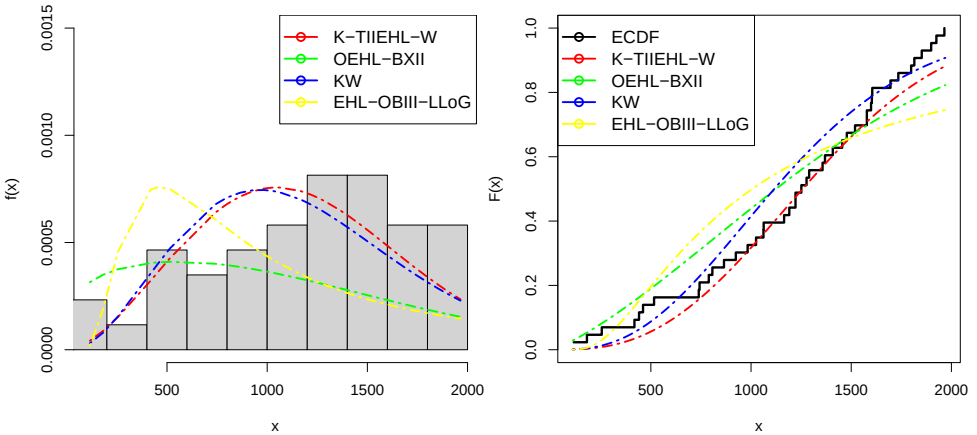

Figure 7. Plots of the observed histogram with superimposed estimated pdfs and observed ogives and estimated cdfs for blood cancer data

Figure [7] displays the histogram and ogive of the observed data, along with the fitted densities and the ECDFs for visual assessment of the fitting accuracy. These plots suggest that the K-TIIEHL-W distribution offers a better fit for the blood cancer data compared to the competing models.

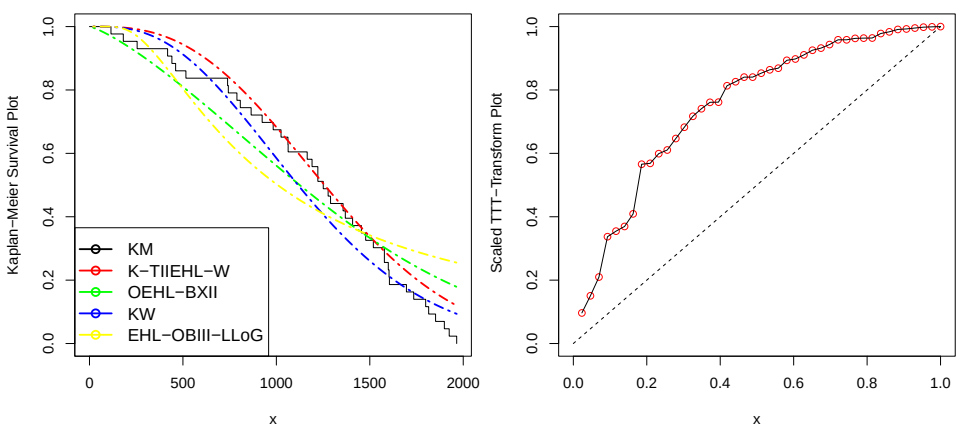

Figure 8. Plots of the KM survival curve along with the estimated KM survival curves, and the TTT-scaled plot of the blood cancer data

The KM survival curve depicted in Figure [8] demonstrates a strong alignment with the K-TIIEHL-W distribution when compared to alternative distributions on blood cancer data. Additionally, the TTT-scaled plot indicates that the hrf exhibits an increasing hazard rate pattern.

6.2. Time until failure of viscose/polyester yarn data

The following right-skewed dataset, as discussed by Arshad, Iqbal and Mutairi (2021), pertains to the temporal duration preceding the failure of a viscose/polyester yarn within a textile experiment aimed at assessing the tensile fatigue properties of the yarn. The data comprises a sample of 100 cm of yarn subjected to a strain level of 2.3%. The data was also analyzed by Shanker, Fesshaye and Selvaraj (2015). The values are:

86, 146, 251, 653, 98, 249, 400, 292, 131, 169, 175, 176, 76, 264, 15, 364, 195, 262, 88, 264, 157, 220, 42, 321, 180, 198, 38, 20, 61, 121, 282, 224, 149, 180, 325, 250, 196, 90, 229, 166, 38, 337, 65, 151, 341, 40, 40, 135, 597, 246, 211, 180, 93, 315, 353, 571, 124, 279, 81, 186, 497, 182, 423, 185, 229, 400, 338, 290, 398, 71, 246, 185, 188, 568, 55, 55, 61, 244, 20, 284, 393, 396, 203, 829, 239, 236, 286, 194, 277, 143, 198, 264, 105, 203, 124, 137, 135, 350, 193, 188.

Table 3. Estimates and GoF statistics for K-TIIEHL-W and other distributions on failure of polyester/viscose data

	Estimates				GoF statistics							
Model	α	c	λ	d	$-2\log(L)$	AIC	AICC	BIC	W^*	A^*	$K-S$	$p-value$
$K-TIIEHL-W$	9.00×10^{-3}	2.8380	18.4990	0.7042	1249.7020	1257.7020	1258.1230	1268.1230	0.0803	0.5169	0.0724	0.7706
	(2.48×10^{-3})	(8.16×10^{-3})	(1.23×10^{-7})	(0.0492)								
	α	β	δ	θ								
$EHL-WP$	0.3984	0.1083	4.7245	44.1292	1249.8360	1258.8360	1258.957	1268.9560	0.0929	0.5201	0.0813	0.5227
	(0.1121)	(0.1051)	(1.8259)	(0.0164)								
	a	b	α	β								
KW	67.1040	1.95×10^{3}	38.3790	0.0882	1249.9780	1257.9780	1258.399	1268.3980	0.1105	0.6073	0.0879	0.4222
	(4.07×10^{-7})	(1.95×10^{-9})	(1.51×10^{-7})	(6.17×10^{-4})								
	a	b	α	c								
$EHL-OBIII-LLoG$	0.5281	132.9400	1.8537	2.4061	1295.1610	1303.1610	1303.5820	1313.5820	0.8184	4.4807	0.1672	0.0949
	(0.0238)	(3.60×10^{-3})	(0.4569)	(5.21×10^{-3})								
	α	λ	a	b								
$OEHL-BXII$	0.5379	3.07×10^{-5}	0.9549	1.9233	1300.8180	1308.8180	1309.2390	1319.2390	0.1341	0.8553	0.1162	0.1340
	(1.15×10^{-10})	(3.69×10^{-6})	(1.29×10^{-9})	(6.42×10^{-10})								
	α	β	b	λ								
$TL-OBIII-LLoG$	0.1007	203.9300	0.5567	9.1446	1286.1350	1294.1350	1294.5560	1304.5560	0.7040	3.8564	0.1680	0.0714
	(3.89×10^{-3})	(2.34×10^{-4})	(0.1191)	(4.30×10^{-5})								

Table 3 presents the maximum likelihood estimates along with their corresponding SEs (in parentheses). The GoF results indicate that the K-TIIEHTL-W model outperforms its competitors.

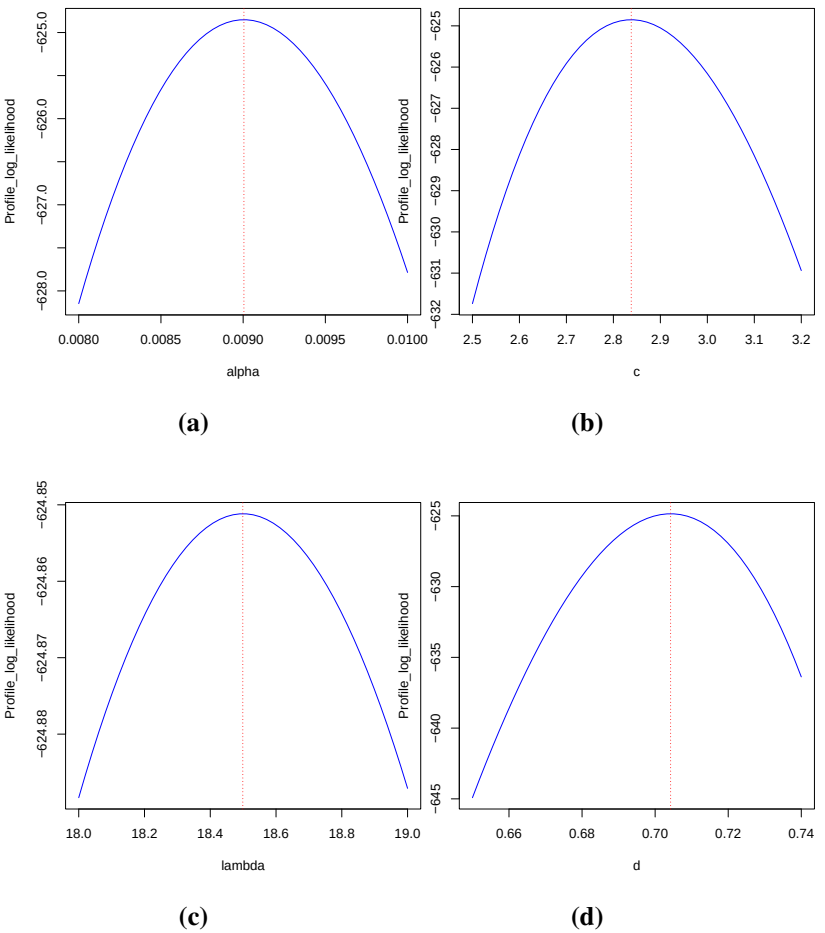

Figure 9. Profile log-likelihood plots of the K-TIIEHL-W: Viscose yarn data

Figure [9] clearly shows that the K-TIIEHL-W parameters on viscose yarn data can be uniquely determined. The 95% ACI for the parameters of the K-TIIEHTL-W distribution are as follows: $\alpha \in [9.00\times10^{-3} \pm 0.0049]$, $c \in [2.8380 \pm 0.0160]$, $\lambda \in [18.4990 \pm 0.0002]$, and $d \in [0.7042 \pm 0.0965]$.

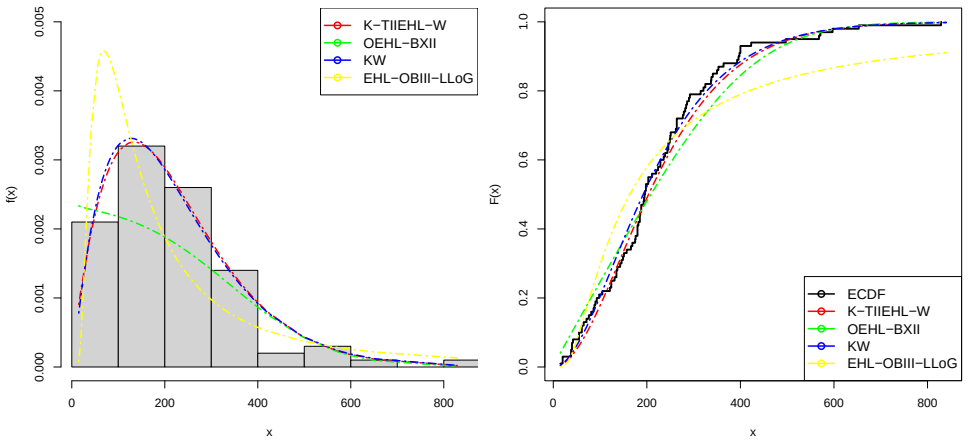

Figure 10. Plots of the observed KM survival curve alongside the estimated KM survival curves, and the TTT scaled plot for polyester/viscose yarn dataa

Figure [10] present the histogram and ogive of the observed data, accompanied by the fitted density functions and ECDF curves for an evaluative comparison of the fitting quality on failure of polyester/viscose data. The visual evidence indicates that the K-TIIEHL-W yields the most accurate representation of the data.

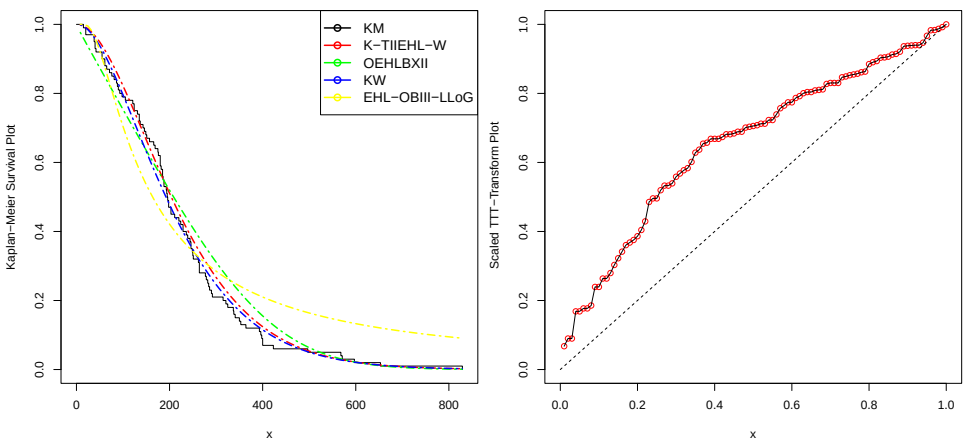

Figure 11. Plots of the observed Kaplan-Meier survival curve and the estimated KM survival curves, and the TTT-scaled plot for blood cancer data

The KM survival curve illustrated in Figure [11] reveals that the K-TIIEHL-W distribution closely aligns with the polyester/viscose yarn data in comparison to other competing

distributions. Furthermore, the TTT-scaled plot suggests that the hrf demonstrates a pattern of increasing hazard rates.

7. Results and discussions

In this study, we introduced the K-TIIEHL-G FDs and explored their mathematical and statistical properties, along with specific cases. By employing maximum likelihood estimation techniques, we effectively estimated the unknown parameters and assessed their consistency through simulation studies. The results underscored the efficacy of the K-TIIEHL-W distribution, a member of the K-TIIEHL-G family. Furthermore, our analysis of the flexibility and practical applications of the K-TIIEHL-G FDs using empirical datasets demonstrated that the K-TIIEHL-W distribution significantly outperformed several existing equi-parameter models.

As we look to the future, we propose two exciting avenues for further research: first, integrating this new family of distributions into bivariate extensions through the copula approach; and second, investigating Bayesian estimation techniques to analyze these distributions, leveraging their adaptability to improve modeling accuracy and facilitate informed decision-making. Exploring these directions could greatly enhance our understanding of the theoretical properties and practical applications of the K-TIIEHL-G distributions, thereby expanding their relevance in engineering reliability and lifetime analysis, with meaningful implications for the design, maintenance, and optimization of complex engineering systems and components.

To access the appendix, please click on the link:
`https://drive.google.com/file/d/1QfzFva31__XXGN0UZrQ4IQcTvafiQjTW/view?usp=sharing`

References

Abouammoh, A. M., Abdulghani, S. A., (1994). On Partial Orderings and Testing of New Better Than Renewal Used Classes . *Reliability Engineering and System Safety*, 43(2), pp. 37–41.

Aldahlan, M., Afify, A. Z., (2018). The Odd Exponentiated Half-Logistic Burr XII Distribution. *Pakistan Journal of Statistics and Operation Research*, 14(2), pp. 305–317.

Al-Mofleh, H., Elgarhy, M., Afify, A. Z., Zannon, M., (2021). Type II Exponentiated Half Logistic Generated Family of Distributions with Applications. *Electronic Journal of Applied Statistical Analysis*, 13(2), pp. 536–561.

Arshad, M. Z. and Iqbal, M., Mutairi, A., (2021. A Comprehensive Review of Datasets for Statistical Research in Probability and Quality Control. *Journal of Mathematics and Computer Science*, 11(3), pp. 663–638.

Bashiru, S. O., (2024). A Study on Topp-Leone Kumaraswamy Fréchet Distribution with Applications: Methodological Study. *Turkiye Klinikleri Journal of Biostatistics*, 16(1), pp. 1–15.

Chipepa, F., Oluyede, B., Chamunorwa S., Makubate B., Zidana, C., (2022). The Exponentiated Half Logistic-Generalized-G Power Series Class of Distributions: Properties and Applications. *Journal of Probability and Statistical Science*, 20(1), pp. 21–40.

Chipepa, F., Oluyede, B., 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.

Cobelli, C., Distefano 3rd, J. J., (1980). Parameter and Structural Identifiability Concepts and Ambiguities: A Critical Review and Analysis. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 239(1), pp. 7-–24.

Cordeiro, G. M., de Castro, J., (2011). A New Family of Generalized Distributions. *Journal of Statistical Computation and Simulations*, 81, pp. 883-–893.

Cordeiro, G. M., Ortega, E. M., Nadarajah, S., (2010). The Kumaraswamy Weibull Distribution with Application to Failure Data. *Journal of the Franklin Institute*, 7(8), pp. 1399-–1429.

Elangovan, R., Arunkumar, S., (2022). A New Extension of Quasi Sujatha Distribution with Properties and Applications using Time Dependent Covariates. *Indian Journal of Natural Sciences*, 13(74), pp. 0976–0997.

Ibrahim, S., Doguwa, S. I., Isah, A., Haruna, J. M., (2020). The Topp Leone Kumaraswamy-G Family of Distributions with Applications to Cancer Disease Data. *Journal of Biostatistics and Epidemiology*, 6(1), pp. 40–51.

Moakofi, T., Oluyede, B., Gabanakgosi, M., (2022). The Topp-Leone Odd Burr III-G Family of Distributions: Model, Properties and Applications. *Statistics, Optimization and Information Computing* 1, pp. 1–27.

Nagarjunai, V. B. V., Vardhan, R. V. and Vardhan, C. (2021). Kumaraswamy generalized power Lomax distribution and its applications. *Stats* 4(1), pp. 28–45.

Oluyede, B., Peter, O. P., Nkumbuludzi, N., Huybrechts, B., (2022). The Exponentiated Half-Logistic Odd Burr III-G: Model, Properties and Applications. *Pakistan Journal of Statistics and Operation Research*, 18(1), pp. 33–57.

Shaked, M., Santhikumar, J. G., (2007). Stochastic Orders. *Springer Series in Statistics, Springer New York, NY*.

Shanker, R., Fesshaye, S., Selvaraj, S., (2015). On Modeling of Lifetimes Data Using Exponential and Lindley Distributions. *Biometrics & Biostatistics International Journal*, 2(5), pp. 883–893.

Usman R. M., Ahsan-ul-Haq, M., Talib, J., (2017). Kumaraswamy Half-Logistic Distribution: Properties and Applications. *Journal of Statistics Applications and Probability*, 6(7), pp. 597–609.

Rényi, A., (1960). On Measures of Entropy and Information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 25, pp. 547–561.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 135–155, https://doi.org/10.59139/stattrans-2026-031
Received – 30.04.2025; accepted – 17.04.2026

Modeling exchange rate volatility using EM algorithm with hybrid GARCH models

Ali Salman Habeeb[1]

Abstract

The most important basic tools for modeling and analyzing econometric models are linear and nonlinear time series analyses, with the aim of formulating an accurate probabilistic model that expresses the behavior of the phenomenon over a specific time period. In this study, we construct a three-step hybrid probability model that includes model identification, model estimation, and diagnosis checking using the Box-Jenkins approach. We explored the performance of two important tools for estimating unknown parameters of the proposed model. The first is Maximum Likelihood (ML), which has a wide reputation in the statistical literature, and is provided in *Eviews*10 software. Furthermore, we compared it with the second proposed numerical method, namely Expectation-Maximization (EM algorithm), which requires creating a special MATLAB code to implement its performance. Moreover, we compared two methods using information criteria (AIC, BIC, HQ). In order to show the effectiveness of the hybrid ARCH-GARCH models, we used them to measure the volatility of the weekly exchange rate of the Central Bank of Iraq's sales of the US dollar against the Iraqi dinar from the first week of 2021 to the last week of 2023. We concluded that the appropriate model is the hybrid $MA(1) - GARCH(1,1)$ model, which is represented by the Moving Average $MA(1)$ of the average equation and the $GARCH(1,1)$ model of the conditional variance equation.

Key words: time series analysis, exchange rate volatility, maximum likelihood, expectation-maximization(EM algorithm), generalized autoregressive conditional heteroscedasticity.

1. Introduction

The analysis of financial time series is one of the important topics in the field of financial econometrics because of its reflection the various developments of quantitative economics. Therefore, it has attracted the interest of many researchers in this field because it focuses on the evaluation of assets and their volatility, as well as the series of stock returns over longer periods of time that cannot be directly observed. Thus, the importance of statistical methods appears in analyzing real phenomena related to time, including time series analysis as a random uncertain process subject to the laws of probability and uncertainty, which requires us to reveal the mathematical structure of both types (linear and nonlinear) inherent in the data through statistical analysis or modeling and identifying the appropriate model. That phenomenon is defined through diagnosis and estimation, and then it is used in the issue of predicting the future values of the phenomenon under research. In this paper, we

[1]Department of Statistics, University of Sumer, Dhi Qar Governorate, Al-Rifai, Iraq.
E-mail: asalhabeeb@gmail.com. ORCID: https://orcid.org/0000-0003-0037-8292.

considered one of the most difficult topics facing decision-makers in the field of financial policy making in Iraq, which is identifying the best econometric models to modeling the phenomenon of financial fluctuations over a relatively long period of time extending from the first week of January 2021 to the last week of January 2023, with 156 weekly observations (five trading days a week). In addition, what are the appropriate monetary policies in the event of random financial shocks in the Iraqi stock market? One of the first nonlinear models that dealt with the issue of volatility was the autoregressive model with the condition of heterogeneity of variance (ARCH), presented by researcher Engle (1982) in an article that studied the autoregressive model of the first order and the unconditional mean to be zero. Draper and Joha (1982) the presented research based on the nonlinearity test for residual series, while Bollerslev (1986) developed the ARCH model to become GARCH, and Cochrane (1988) investigated a measure of the continuity of volatility in Gross National Product (GNP) based on variation and its differences in the long run and studied a group of models, perhaps the most important of which is the random walk model. Nelson (1991) used the GARCH models to model the relationship between conditional variance and asset risk. These models have at least some major defects when pricing assets, perhaps the most important of which is that random shocks do not continue in GARCH models due to stationarity measurement standards. Hall and Yao (2003) showed that asset aggregation and risk calculations rely heavily on volatility models and provided a comparison of forecast performance using a group of ARCH models and working on daily exchange rate data against the dollar and the importance of specific and correct distribution of asset returns and not focusing only on daily volatility. Awartani and Corradi (2005) studied the relative predictive ability of different types of GARCH models, with a focus on asymmetric models. Pairwise comparisons were made with different models compared with the $GARCH(1,1)$ model; the same finding applies to different longer forecast horizons, although the predictive superiority of asymmetric models is not as striking as in the one-step-ahead comparison. So and Yu (2006) investigated the efficacy of seven distinct GARCH model specifications, incorporating various market indicators and foreign exchange rate to evaluate Value at Risk (VaR) and according to the RiskMetrics framework across multiple error distribution levels, they demonstrated that these models exhibit asymmetric error distributions. Furthermore, their analysis confirmed that daily exchange rate data are characterized by heavy-tailed distributions and long-memory properties. Ghahramani and Thavaneswaran (2008) confirmed that financial returns are often designed as a time series, especially the GARCH model, and the error distribution is often determined from parametric distributions, which was not justified. They confirmed in the research a specific method for diagnosing the identity of GARCH models for a number of data sets. On the other hand, Fan and Xiu (2014)) proposed that the non-Gaussian maximum likelihood estimator be used frequently in GARCH models; with heavy-tailed distributions, the estimator is biased and inconsistent in the case of the parametric likelihood family. They provided an unknown scale parameter to the identification for consistency, and they proposed a three-step quasi-maximum likelihood procedure, particularly when likelihood functions have heavy tails. Kononovicius and Ruseckas (2015) used the normal distribution in modeling volatility even if the return distribution is asymmetric and has excessive kurtosis. Klein and Walther (2016) assumed a log-normal distribution of oil price returns in order to compare different types of ARCH models in terms of accu-

racy of forecasting volatility. Ewing and Malik (2017) found that there is a problem with the issue of asymmetry and excessive kurtosis in oil revenues, which are slightly low when the model calculates the structural breaks. Cerqueti et al. (2020) proposed the application of non-Gaussian GARCH models of volatility in a new and comprehensive study about the cryptocurrency market, evaluating the forecasting performance for three of the most important cryptocurrencies (Bitcoin, Ethereum, and Litecoin) in terms of market capitalization, the results demonstrated the effectiveness of the prediction performance. Massadikov (2021) used the VAR-GARCH model to determine the interaction between oil prices and stock markets in terms of return and volatility in the financial markets of some developing countries. It was demonstrated through the use of this model that it is important in discovering and interpreting the reflection of risks for some economies of oil-producing countries. Habeeb et al. (2021) discussed the method of estimating smoother splines for the Additive Autoregressive Model, as it is one of the most important numerical methods used in curve fitting of the regression function. Tashpulatov (2022) in his paper presents volatility modeling for wholesale electricity market prices in England and Wales over a determined period by the AR-ARCH model with four flexible distributions that account for asymmetry, heavy tails, excess kurtosis, and a peak higher than in normal distribution, observed in the data or model residuals; it applied generalized deviation, and that distribution is suitable to study this case. Chifurita and Chinhamn (2022) dealt with the daily returns of financial market variables, which often have empirical distributions that are heavy or semi heavy tailed, and they also proved that the estimation value-at-risk depends highly on the distributional characteristics of the stock returns. The aim of the study is to investigate the relative performance of the error distributions of the GARCH model, such as the generalized hyperbolic skew Student-t and Pearson $type-IV$ distributions, as this study recommends that the $ARMA(1,1)EGARCH(1,1)$ model be used in the estimation of the VaR for the daily returns from the FTSE/JSE growth index (J280). Ampadu et al. (2024) presented a study to exhibit volatility clusters and effects on the importance of error distribution to improve the performance of the GARCH model through simulation experiments of five different types of error distributions. The study showed that the Skewed Generalized Error Distribution (SGED) is the most efficient and consistent for modeling daily return volatility through the use of different criteria. Antony and Prabakarant (2024) considered the spatial autoregressive model, which is applied to state-wise $COVID19$ confirmed rates. This article aims to build a stochastic model that describes the financial volatilities of the Iraqi dinar exchange rate against the weekly US dollar in Iraq for the period (2021-2023) in the Iraqi stock market using the $GARCH(1,1)$ model, and in the presence of skewed distributions for the error term, particularly those exhibiting heavy tails, deploying a robust mathematical model is essential for phenomenon study. In this paper, we provide the theoretical background by describing the theoretical basis of the weak stationarity of the model and building a theoretical analysis of the ARCH model, paying particular attention to the organization and some important tests in this context. Then we studied the $GARCH(1,1)$ model and discussed the estimation process. Finally, we reach some brief conclusions, including estimating the proposed model.

2. Methodology

Most financial studies deal with a series of asset returns instead of price returns, but in this research, we will deal with price returns (Bala, 2013). If we assume that pt represents the simple price for one period of time (selling price at closing) over a time period (such as a day, a week, or a month), and pt1 represents the simple price for a previous period of time, then the return is $(r_t = \frac{p_t}{p_{t-1}})$. So, the returns can be represented through a logarithmic transformation of the ratio of the return values for the current period relative to the last period. In this paper, we consider the mathematical formula as follows: $y_t = \log(\frac{x_t}{x_{t-1}}) = \log(x_t) - \log(x_{t-1})$,

where y_t: represents the weekly returns of the exchange rate, x_t is the exchange rate of the current period, and x_{t-1} is the exchange rate of the previous period. Volatility represents the standard deviation of returns (symbolized by σ_t)). It measures the price behavior of returns and is sometimes considered to be the accompanying risk.

The mathematical formula adopted to calculate it is: $\sigma_t^2 = \sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(y_t - \mu)^2)}$, where: μ represents the arithmetic mean, and n represents the number of observations in the sample. When representing volatility in returns by the *GARCH* model, there are two distinctive characteristics of these models, the conditional mean and the conditional variance. $E(y_t|y_{t-1},\ldots,y_{t-p}) = \mu_t \quad and, Var(y_t|y_{t-1},\ldots,y_{t-p}) = E(y_t^2|y_{t-1}^2,\ldots) = \sigma_t^2$.

2.1. Definition for Model and Stationary of the Series

In some articles for studying and analyzing the time series, the series $\{y_t\}_{t=1}^{n}$ must be stationary, and it has two types, The first is Strict stationarity, which assumes that the time series is $\{y_t, t \in \mathscr{Z}\}$, where $\mathscr{Z}$ is the set of integers; we say that it is strictly stationary in the distribution function for the time period (t) and has the same distribution for the subsequent time period $(t+k)$, and this type is called strongly stationary and can be expressed when the following condition holds: $F(y_{t_1},\ldots,y_{t_n}) = F(y_{t_1+k},\ldots,y_{t_n+k})$. The joint probability distribution for the variables is the same as for the variables $(y_{t_1+k},\ldots,y_{t_n+k})$. The second type is weak stationarity if we have the following: $i)E(y_t) = \mu, \quad ii)E(y_t^2) < \infty; \quad iii)\gamma_y(s,t) = \gamma_y(s+k,t+k) = Cov(y_t, y_{t+k}).\forall t \in \mathscr{Z}$, where k is integer numbers, this means $\gamma(k), \forall k = 0,1,2,\ldots$. In addition, if $k = 0$, then the covariance of the time series is the variance $(i.e.\gamma(0) = \sigma_t^2)$. The autocorrelation function measures the linear dependence between y_t *and* $y_{(t+k)}$. Likewise, the autocorrelation function is given according to the following mathematical formula: $\rho_k = \frac{Cov(y_t,y_{t+k})}{Var(y_t)} = \frac{\gamma(k)}{\gamma(0)}$, where k is an integer, this means $\rho_k, \forall k = 0,1,2,\ldots$, and it is called the autocorrelation function. In addition, if $k = 0$ then $\rho_0 = 1$. Also, the autocorrelation function is independent of the time series measurements. The most important test that has wide application among researchers is the Augmented Dickey-Fuller Approach (ADF), which is one of the unit root tests in the search for the stationarity of time series, and the expanded form of this test works to estimate the first-order

autoregressive model:

$$\begin{aligned} & \Delta y_t = \phi y_{t-1} + \sum_{j=1}^{p-1} \alpha_j^* \Delta y_{t-j} + \varepsilon_t, \\ with linear Trend \qquad & \Delta y_t = \delta + \phi y_{t-1} + \sum_{j=1}^{p-1} \alpha_j^* \Delta y_{t-j} + \varepsilon_t; \end{aligned} \tag{1}$$

where $\phi = -\alpha(1), \quad \alpha_j^* = -(\alpha_{j+1}, \ldots, \alpha_p); \qquad \Delta y_t = y_t - y_{t-1}$. Therefore, we test the significance of the above model coefficient, which in turn means testing the significance of the unit root (ϕ) through the following hypothesis: $H_0 : \phi = 0 \quad Vs. \quad H_1 : \phi < 1$. This condition is verified for the ARCH(p) model when the following is true: $i) E(y_t^2 < \infty, ii) \sum_{i=1}^{p} \alpha_i < 1$. The roots lie outside the unit circle. When we are interested in the ARCH(1) model, the condition for weak stationary is $(\alpha_1 < 1)$. As for the GARCH(p,q) model, weak stationary is achieved by verifying the following (Andersen et al. 2009) $i) E y_t^2 < \infty, ii) \sum_{i=1}^{p} \alpha_i + \sum_{i=1}^{p} \beta_j < 1$. The roots lie outside the unit circle. When we are interested in the GARCH(1,1) model (Jafari et al. 2007), the condition for weak stationary is $(\alpha_1 + \beta_1 < 1)$. The basic idea is to find a model that can sense the effects of unpredictable volatility that are the main cause of changes (in more detail, e.g. Alberg et al. 2008; Angabini and Wasiuzzaman 2011 and Engle 2001). And uncertainty here indicates the inability of the model to predict the future that will affect the phenomenon's behavior, and that addressing this idea requires a stochastic model that can sense changes over time. And y_t is a random variable that has a conditional mean and conditional variance defined on the set of information $\mathscr{F}_{t-1}$ (generated from the $\sigma-$field) defined through the random process $\{y_t; t \in \mathscr{Z}\}$ and it has weak stationarity. The general form of the model is of the order ARCH(p) for positive integer $(p \geq 1)$, the random process has $E(y_t|\mathscr{F}_{t-1}) = 0 \quad and \quad E(y_t^2|\mathscr{F}_{t-1}) = \sigma_t^2$ (see more details in, e.g. Draper and John (1982); and So and Yu (2006)).

$$y_t = \mu + e_t \implies e_t = \varepsilon_t \sqrt{\sigma_t^2}; \tag{2}$$

where ε_t is the random error series, which is independent and identically distributed with zero mean and specified variance (σ_t^2), and is denoted by the process $\{\varepsilon_t; t \in \mathscr{Z}\}$, and y_t return series. It can be shown that the conditional distribution of $y_t|\mathscr{F}_{t-1} \sim N(0, \sigma_t^2)$, where σ_t^2 is the conditional variance of a time series for different time periods.

$$\begin{aligned} Var(y_t|y_{t-1}, \ldots, y_{t-p}) &= E(y_t^2|y_{t-1}, \ldots) - E^2(y_t|y_{t-1}, \ldots) = \sigma_t^2 \\ \sigma_t^2 &= \alpha_0 + \alpha_1 e_{t-1}^2 + \cdots + \alpha_p e_{t-p}^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i e_{t-i}^2; \end{aligned} \tag{3}$$

where $\alpha_0 > 0, \alpha_i \geq 0; \forall i = 1, 2, \ldots, p.$

2.2. Generalized Autoregressive Conditional Heteroskedicity GARCH(p,q) Model

According to the conditional variance relationship in (3) the current variance is also dependent upon the lag variables of the conditional variance series. $\sigma_{t-1}^2,\ldots,\sigma_{t-q}^2$. In addition, the squares of the time lags for the process $\{y_t; t \in \mathscr{Z}\}_{t=1}^{T}$ become a general model called $GARCH(p,q)$, where $(p \geq 0, q \geq 1)$. (see, Bollerslev (1986); Kononovicius and Ruseckas (2015) and Lundbergh and Teräsvirta (2002)).

$$\begin{aligned} & y_t = \mu + e_t \implies e_t = \varepsilon_t \sqrt{\sigma_t^2}\,. \\ and, \quad & \sigma_t^2 = \alpha_0 + \alpha_1 e_{t-1}^2 + \cdots + \alpha_p e_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_p \sigma_{t-q}^2, \\ & \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i e_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2. \end{aligned} \tag{4}$$

where $\alpha_0 > 0 \quad , \alpha_i \geq 0; \quad \beta_j \geq 0 \quad \forall i, j = 1,2,\ldots,p,q$; respectively. The model parameters to be estimated, and we can write the relationship in (4), (Andersen et al. (2009) and Ding (2021)) as follows:

$$y_t = \mu + e_t \implies e_t = \varepsilon_t \sqrt{\sigma_t^2} \quad and, \, \sigma_t^2 \quad = \alpha_0 + \sum_{i=1}^{p} \alpha_i e_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2. \tag{5}$$

3. Testing the model

In the literature of time series analysis, volatility represents the standard deviation of currency sales returns in dollars (Mapa 2004). It measures the price behavior of the currency and is sometimes expressed as the current exchange rate. In order to verify that the model is free of any problems that negatively affect the accuracy of the model (see, e.g. Bangura et al. (2021) and Draper and Joha (1982)), a set of tests can be used in this regard Lagrange Multiplier test is one of the most important tests used to detect that the model contains the problem of heteroscedasticity, Draper and Joha (1982). To test the existence of the heteroskedasticity problem, the coefficient of determination is calculated, which depends on the regression of the squares of the residuals for the current period with respect to the squares of the residuals for the previous period, and its statistic for this test is

$$ARCH - test = T \times R^2 \sim \chi_{(p)}^2; \tag{6}$$

where is the number of model parameters, T is the size of the sample, and R is the coefficient of determination. We calculate the test statistic; it follows a chi-square distribution with degrees of freedom (p). The process $\varepsilon_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2$. As for the hypothesis that we are testing, $\quad H_0 : \alpha_1 = \cdots = \alpha_k = 0 \quad Vs. \quad H_1 : \alpha_k \neq 0, \quad \forall k = 1,\ldots,K.$

4. Theoretical distributions of the model residuals

Let y_t denote a real-valued discrete-time stochastic process, and let $\mathscr{F}_t$ denote the $\sigma-algebra$ generated by the history time series. The process $\{y_t; t \in \mathscr{Z}\}$ is represented for time series, then y_t follows a GARCH(p,q) model, and this model can also be written in (5) as follows:

$$\begin{aligned} y_t &= \mu + e_t, \qquad (Mean\ Equation) \\ and,\ e_t = \varepsilon_t \sigma_t \implies \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p} \alpha_i e_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2; \quad (Variance\ Equation) \end{aligned} \tag{7}$$

where $\alpha_0 > 0, \quad \alpha_i \geq 0, \quad \beta_j \geq 0, \quad \forall\, i = 1,2,\ldots,p \ \&\ j = 1,2,\ldots,q$. The model in (7) will be stationary if the following condition is met $\left[\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1\right]$.

Here if the process is white noise (i.e. $\varepsilon_t \sim N(0,1)$), we have the following:

$$\begin{aligned} E(y_t|\mathscr{F}_{t-1}) &= E(\sigma_t \varepsilon_t) = E(E(\sigma_t \varepsilon_t|\mathscr{F}_{t-1})) = E(\sigma_t E(\varepsilon_t|\mathscr{F}_{t-1})) = E(\sigma_t E(\varepsilon_t)) = 0, \\ Var(y_t|\mathscr{F}_{t-1}) &= E(y_t^2|\mathscr{F}_{t-1}) - E^2(y_t|\mathscr{F}_{t-1}) = E(\sigma_t^2 \varepsilon_t^2|\mathscr{F}_{t-1}) = E(\sigma_t^2 E(\varepsilon_t^2)) = \sigma_t^2. \end{aligned}$$

Then, $y_t \sim D(0,\sigma_t^2)$, $Cov(y_t y_s) = 0$ for $t \neq s \Rightarrow \sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$. Furthermore,

$$\begin{aligned} E(\sigma_t^2|\mathscr{F}_{t-1}) &= E(\alpha_0|\mathscr{F}_{t-1}) + \sum_{i=1}^{q} \alpha_i E(y_{t-i}^2|\mathscr{F}_{t-i}) + \sum_{j=1}^{p} \beta_j E(\sigma_{t-j}^2|\mathscr{F}_{t-j}), \\ \sigma_t^2 &= \alpha_0 + \sum_{i=1}^{q} \alpha_i \sigma_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 = \frac{\alpha_0}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j}. \end{aligned} \tag{8}$$

This implies $E[y_t|\mathscr{F}_{t-1}] = 0$ and $E[y_t^2|\mathscr{F}_{t-1}] = \sigma_t^2 \implies y_t|\mathscr{F}_{t-1} \sim D(0,\sigma_t^2)$.

For the $GARCH(p,q)$ model, which is defined in (9), assuming that ε_t is the disturbance term, it has three different types of distributions called heavy-tailed distributions (see more details in, e.g. Hall and Yao (2003), Preminger et al., (2017), and Chen and Liu (2013)), which can be given according to the following cases:

Case I The Gaussian distribution: If ε_t is a random variable with probability density function as follows: $f(\varepsilon_t;\sigma_t^2) = \frac{1}{\sqrt{2\pi\sigma_t^2}} e^{-\frac{1}{2}\frac{\varepsilon_t^2}{\sigma_t^2}}$ where $\varepsilon_t \in \mathscr{R}, \mu = 0; \sigma_t \in \mathscr{R}^+$ is the scale parameter, denoted by $\varepsilon_t \sim N(0,\sigma_t^2)$.

Case II The Student's t distribution: If ε_t is a random variable, the probability density function of ε_t, is defined as (see, e.g. Adubisi et al., (2022) and Awalludin et al., (2018)): $f(\varepsilon_t;\sigma_t^2,v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)\sigma_t\sqrt{\pi(v-2)}} \left[1 + \frac{(\varepsilon_t)^2}{(v-2)}\right]^{-\left(\frac{v+1}{2}\right)}$ where $\varepsilon_t \in \mathscr{R}, \mu = 0; \sigma_t \in \mathscr{R}^+$ is the scale parameter, and $2 < v < \infty$ is the positive integer number (degrees of freedom) and $\Gamma(.)$ is the Euler gamma function.

Case III The Generalized Error Distribution (GED): If ε_t is a random variable, and it has probability density function as follows: $f(\varepsilon_t;\sigma_t,v)=\frac{v}{\lambda\sigma_t 2^{(1+1/v)}\Gamma(1/v)}.\exp\left(-\frac{1}{2}\left|\frac{\varepsilon_t}{\lambda}\right|^v\right)$. where $\varepsilon_t\in\mathscr{R},\mu=0,\sigma_t\in\mathscr{R}^+\quad \lambda=\frac{1}{2^{\frac{1}{v}}}.\left[\frac{\Gamma(1/v)}{\Gamma(3/v)}\right]^{\frac{1}{2}}$, and v $\in\mathbb{N}$ is the shape parameter.
Therefore, it is necessary to know the type of distribution of the error term, which varies according to the distributions, as the variance is not constant in this type of model, which has heavy tails, and by plotting the probability density function for financial volatility data, we can identify these important distributions when analyzing time series . We have shown

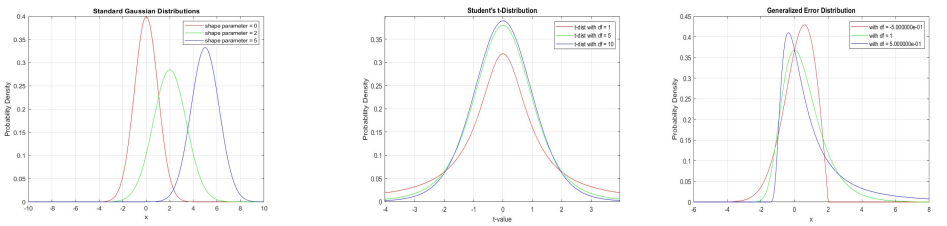

Figure 1. Probability density function for: (a) Plot of the Gaussian distribution with different shape parameters; (b) Plot of the Student's t-distribution with different degrees of freedom; (c) Plot of the GED with different skewed values

in (1) the three cases of the different distributions: the Gaussian distribution for different shape parameters, the Student's t-distribution with different degrees of freedom, and the generalized error distribution with heavy tails.

5. Estimation of the Model

One of the methods commonly used in the literature to estimate model parameters is the maximum likelihood method because it has good and desirable theoretical properties when interested in estimating the parameters of this type of the (ARCH or GARCH) model (in more detail, e.g. Huang et al., (2008) and Rzgar and Kale (2007)), which represents financial volatility, and we can express it by the standard deviation (variance) of the random error term (see more details in, e.g. Augustyniak (2014) and Bera and Higgins (1993)). Therefore, it is necessary to know the type of distribution of the error term, which varies according to the distributions, as the variance is not constant in this type of model, which has heavy tails, and by plotting the probability density function for financial volatility data we can identify these important distributions, in particularly when analyzing time series. We assume that the random error term has a normal distribution, and then the error term is white noise at a mean of zero and a specified variance, and this case is considered optimal (PzCruz et al., (2003)). However, this assumption may be inaccurate and thus affect the accuracy of choosing the appropriate model for the analysis. Therefore, the distribution of the residuals must be tested for the appropriate distribution (Gao et al., (2012)).

5.1. Maximum Likelihood Method

The maximum likelihood method is considered one of the more common traditional approaches used in the statistical literature for estimating model parameters because it has good theoretical properties and gives the best fit for the candidate model. We are interested in estimating the unknown parameters of the ARCH and GARCH models (see, e.g. Huang et al., (2008) and Ruzgar and Kale (2007)). This model represents financial volatility, which is expressed by the standard deviation (variance) of the random error term (in more detail, e.g. Augustyniak (2014); Bera and Higgins (1993) and Horv and Kokoszka (2003)). Furthermore, we used this method as a basic tool to deal with the GARCH(p,q) model. We assume that the residuals of the model have the probability distributions mentioned in section 4.

Let $\{e_t,\, t = 1,2,\cdots,T\}$ be a sequence of random residual variables from a distribution that has a probability density function $f(e_t,\theta)$, where θ is the vector of parameters of the model. The $f(\theta|e_t)$ represents the conditional function of the GARCH(p,q) model, and it has an associated probability distribution function. Therefore, the likelihood function is formally defined as: $\ell(\theta|e_t) = \sum_{t=1}^{T} f(\theta|e_t)$. In other words, the likelihood function is often replaced by its natural logarithm, and we have the log-likelihood function given by $L(\theta) = \ln\sum_{t=1}^{T}(\ell(\theta|e_t))$. Therefore, this approach can be applied to the three cases mentioned in Section 4; we obtain the following:

Case I The log likelihood function of the $GARCH(p,q)$ model if the residual has the Gaussian distribution:

$$L(\theta) = \frac{-T}{2} ln\big(2\pi\big) - \frac{1}{2}\sum_{t=1}^{T}\left(ln\big(\sigma_t^2\big) + \big(\frac{\varepsilon_t^2}{\sigma_t^2}\big)\right). \tag{9}$$

Case II The log likelihood function of the $GARCH(p,q)$ model if the residual has the Student's t distribution:

$$\begin{aligned} L(\theta) = T\left(ln\Gamma\left(\frac{v+1}{2}\right) - ln\Gamma\left(\frac{v}{2}\right) - \frac{1}{2} ln\big(\pi(v-2)\big)\right) \\ - \frac{1}{2}\sum_{t=1}^{T}\left(ln\big(\sigma_t^2\big) + \big(1+v\big) ln\big(1+\frac{\varepsilon_t^2}{v-2}\big)\right). \end{aligned} \tag{10}$$

Case III The log likelihood function of the $GARCH(p,q)$ model if the residual has the Generalized Error Distribution (GED):

$$L(\theta) = \sum_{t=1}^{T}\left(ln\big(\frac{v}{\lambda}\big) - \frac{1}{2} \mid \frac{\varepsilon_t}{\lambda} \mid^{v} - \big(1+v^{-1}\big) ln(2) - ln\Gamma\left(\frac{1}{v}\right) - \frac{1}{2} ln\big(\sigma_t^2\big)\right) \tag{11}$$

where $\lambda = 2^{\frac{-1}{v}}\Gamma\left(\frac{1}{v}\right)^{\frac{1}{2}}\Gamma\left(\frac{3}{v}\right)^{\frac{-1}{2}}$, and T is sample size.

Then the maximum likelihood estimator (MLE) for the parameters is found determining the value that maximize $L(\theta|e_t)$, which is:

$$\hat{\theta}_{MLE} = \underbrace{\mathbf{arg\ max}}_{\theta\in\Theta}\Big(L(\theta)\Big). \tag{12}$$

5.2. Expectation-Maximization Method

The EM algorithm is a general method as an iterative approach used to solve maximum likelihood estimation problems with the presence of incomplete data. Let y denote the incomplete data (observed), and the probability density $f_Y(y;\phi)$ indexed by the parameter vector $\phi \in \Omega \subseteq \mathscr{R}^2$ and. Let ε denote the complete data (see, e.g. Feder and Weinstein (1988) and Zhang et al., (2021)).

$$f_\varepsilon(\varepsilon_t;\phi) = f_{\varepsilon|Y=y}(\varepsilon_t;\phi)\cdot f_Y(y;\phi) \tag{13}$$

where $f_\varepsilon(\varepsilon_t;\phi)$ is the probability density associated with ε, and $f_{\varepsilon|Y=y}(\varepsilon_t;\phi)$ is the conditional probability density of ε given $Y=y$. Taking the logarithm for both sides of (15), we obtain: $\log f_Y(y;\phi) = \log f_\varepsilon(\varepsilon_t;\phi) - \log f_{\varepsilon|Y=y}(\varepsilon_t;\phi)$, The conditional expectation given $Y=y$ at a parameter value ϕ

$$E\left\{\log f_Y(y;\phi)\right\} = E\left\{\log f_\varepsilon(\varepsilon_t;\phi)|Y=y;\phi\right\} - E\left\{\log f_{\varepsilon_t|Y=y}(\varepsilon_t;\phi|Y=y;\phi)\right\} \tag{14}$$

The algorithm starts with an arbitrary initial value $\phi(0)$, and $\phi(j)$ denotes the current estimate of ϕ after j iterations of the algorithm. The iteration cycle of the EM algorithm can be described. Therefore, we can express this algorithm through two steps, namely the E-step and the M-step, which are described below for the GARCH model fitting. Hence, the Φ function of the algorithm (Popovici and Dumitrescu 2007) will be as follows:
E-step: In the first step, we Assume that $\Phi\left(\phi|\phi^{(m)}\right)$ is the function for unknown parameters, and the vector $\Phi = \left(\mu,\theta,\alpha_0,\alpha_1,\beta_1\right)$ denotes the parameters of the model:

$$\begin{aligned}\Phi\left(\phi|\phi^{(m)}\right) &= E_{\varepsilon_t|Y=y,\phi^{(m)}}\left[\log\Pi_{t=1}^n f_\varepsilon(\varepsilon_t;\phi)\right] \quad = E_{\varepsilon|Y,\phi^{(m)}}\left[\sum_{t=1}^n \log f_\varepsilon(\varepsilon_t;\phi\right] \\ &= \sum_{t=1}^n E_{\varepsilon|Y,\phi^{(m)}}\left[\log f_\varepsilon(\varepsilon_t;\phi)\right]\end{aligned} \tag{15}$$

M-step: The second step of the EM algorithm is that for maximizing the expectation. We perform the following: we use $(m+1)$ cycle iteration for from $\phi^{(m+1)}$, which is maximization for

$$\phi^{(m+1)} = \underbrace{\mathbf{arg\ max}}_{\phi\in\Phi}\left(\Phi(\phi|\phi^{(m)})\right). \tag{16}$$

6. Application

To provide an overview of the dataset, and the purpose of getting a general idea for it. We present Table 1 which contains the descriptive statistics summary of the weekly exchange rate volatility of the Central Bank of Iraq's official sales of the US dollar against the Iraqi dinar. The observation period extends from the first week of 2021 to the last week of 2023. Accordingly, the indicators of the phenomenon under study are as follows.

Table 1. Summary of the descriptive statistics for weekly returns

Des. Stat.	Mean	Std.Dev.	Min	Max	Median	Ske.	Kur.	JB
Exchange Rate	1493	46.38	1408	1720	1480	1.915	7.367	219.36

It is noted from Table 1 that the average exchange rate is (1493) dinars per dollar in the currency auction and that the lowest price at which the dollar was sold against the dinar was the 121st week of May 2023, when it reached (1408) dinars per dollar. The highest selling price was in the 108th week of January 2023, when it reached 1720 dinars per dollar. The value of the median is 1480 dinars per dollar, and this indicates the rate of sale during the period studied, without being affected by outliers or extreme values. It is known that the value of the median is a robust indicator of the trend of outliers and extreme values, and the median is considered an appropriate average for measuring volatility in the exchange rate. The standard deviation was 46.38, and this indicates that the data has a high dispersion during the measured period. We also note that the skewness coefficient is far from zero, and it is positive. This indicates that the data has a distribution skewed to the right, in addition to the kurtosis coefficient, whose peak is much higher than 3. This indicates that it has a high peak that is higher than the peak of the normal distribution (2 a), which represents the time series that displays the behavior of the phenomenon of weekly volatilities of the exchange rate of the Iraqi dinar against the dollar during the time period (from the first week of 2021 until the last week of 2023) according to the sales tables published on the Central Bank of Iraq website. From Table 1, it can be noted that the value shown by the Jarque-Bera statistic tests the normal distribution of series data, which proved that the data does not follow a normal distribution. In addition, (2 b) is graphically observed on the $(q-q)$ plot; it shows strong deviation from normal distribution.

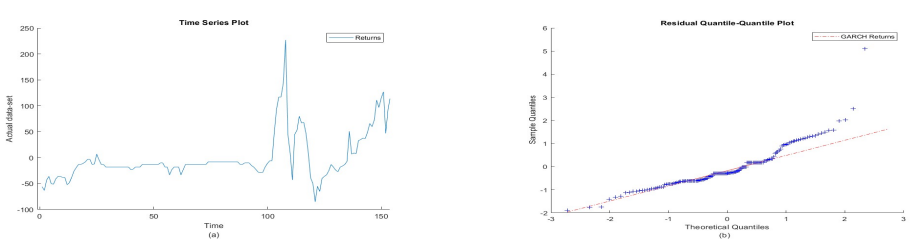

Figure 2. (a) The plot of the weekly returns; (b) The normal Q-Q of the weekly returns

Table 2. The stationary tests

Tests		Exchange Rate series			First Difference for series		
		t-Statistic	**Prob.**	**Test Critical Values**	**t-Statistic**	**Prob.**	**Test Critical Values**
I. Augmented Dickey-Fuller (ADF))	None	0.611	0.847	-2.58	-14.545	<0.01	-2.58
	Constant	-3.216	0.021	-3.47	-14.531	<0.01	-3.47
	Constant + Linear Trend	-4.112	0.008	-4.018	-14.482	<0.01	-4.018
II. Phillips-Perron (PP)	None	0.912	0.903	-2.579	-15.487	<0.01	-2.58
	Constant	-3.216	0.021	-3.47	-15.588	<0.01	-3.47
	Constant + Linear Trend	-3.991	0.011	-4.018	-15.569	<0.01	-4.019

6.1. The Box-Jenkins method for Analyzing and modeling

The time series of returns is nonstationary on average. In addition, there is severe volatility after 2023, as noted in (2.a). These are considered financial shocks (and the reason for this may be due to the entry of dollar sales into the international monitoring platform for currency sales or resulting from severe geopolitical changes that coincided with that period). And, this can be seen through the probability value of (p-value=0.9404); it is greater than the level of significance $(\alpha = 0.01)$, which was calculated by the MATLAB code program. In addition, it is possible to test the stability of the time series by using the Augmented Dickey-Fuller (ADF) test, as in Table 2, which shows the results of this test, which were calculated using the *EViews*10 program. The hypothesis test were evaluated under three cases determined : an intercept, an intercept with a linear trend, and the absence of both. The calculated results of these tests are presented in Table 2 below.

6.2. Augmented Dickey-Fuller Test

In the case of the two tests (Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP)), we see that the calculated value of t-statistics is less than the value of t-theoretical of this test except for the constant with linear trend in ADF, which is not less than the value of t-theoretical, and in this case, we see that this value is very critical, but the PP test supports the decision-maker, and hence the series is not stable in the level and stable in the first difference. Table 2 shows for the $(ADFandPP)$ tests that the probability values using the program $(EViews10)$ were $(P-values = 0.8473, 0.021,$ and $0.008 \simeq 0.01$ for each one from: none, constant, and constant with linear trend, respectively), which is greater than the level of significance $(\alpha = 0.01)$, and therefore the null hypothesis cannot be rejected in (6), and the series has a unit root; it is nonstationary, and the correction is through finding the first difference of the series data. In addition, the graphical representation between the original series data and the logarithmic transformation of the data can be compared in order to find out whether nonlinear transformations as shown in (3) can be used for the purpose of treating instability in the behavior of the phenomenon.

In order to prove this, we conduct the Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) tests for the exchange rate series and the first difference for the series according to what was found in the literature. We note from Table 2 for the Augmented Dickey-Fuller and Phillips-Perron tests that the probability value was (P-value = 0.0000), which is less than the level of significance $(\alpha = 0.01)$ for the first difference, and therefore, the null hypothesis cannot be accepted in (6) and that the series does not have a unit root. The time series is stable on average, and therefore the first difference series can be adopted in the analysis and model building. And to investigate the stationarity of the series and avoid the unit root problem in the exchange rate data, the Phillipa-Perron test was utilized. Additionally, this test was employed to corroborate the findings through the assessment of structural breaks within the series.

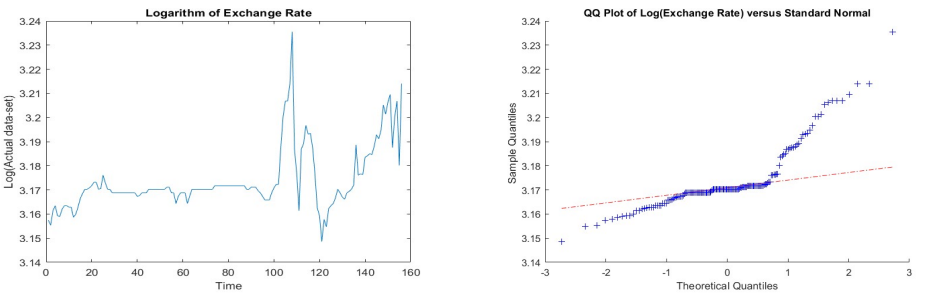

Figure 3. The plot of the logarithm of the weekly returns and the plot of the normal Q-Q of the logarithm of the weekly returns

6.3. Identification of the model

We used the sample ACF and PACF plots to suggest an appropriate parsimonious subclass order for the model, based on the data employed. Hence, we use the sample ACF and PACF plots in (5); it has an exponential decay pattern for PACF, and it cuts off after lag(1) for ACF. Therefore, the proposed appropriate model can be the MA(1) model. So, we believe that the model has been identified (see more detail in, Montgomery (2008), pp. 258, Table 5.2).

6.4. Analyzing and Modeling Using the ARCH/GARCH Method

For the purpose of analysis in this method, we display the logarithmic transformation function of the return series. In order to observe the analysis of the return series data through the modeling of the GARCH model, (5) shows us the conditional variances and residual analysis of the weekly returns for the $GARCH(1,1)$ model after taking the first difference in (4). Furthermore, the volatility of the conditional variances shown in (5) resulted from the use of the maximum likelihood method, which was calculated using $(Eviews10)$ software. This method gives a good fit for the weekly returns series.

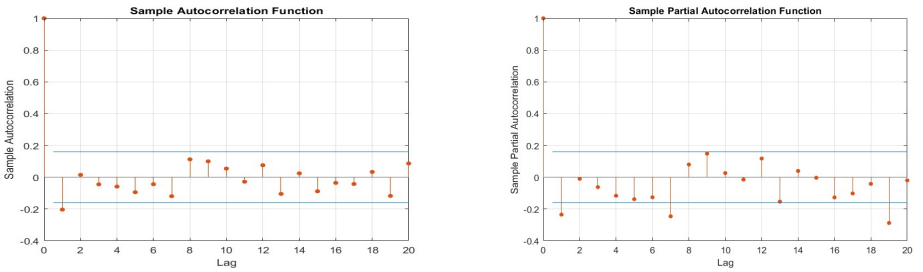

Figure 4. The Autocorrelation Function (ACF) and the Partial Autocorrelation Function (PACF) for the weekly returns

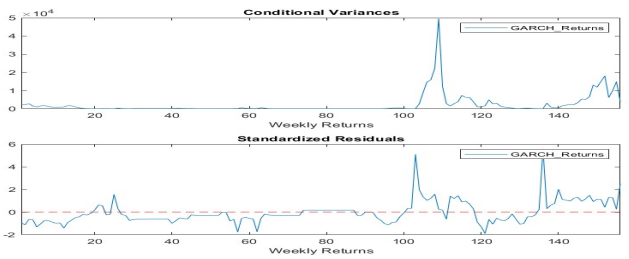

Figure 5. The plot of conditional variances and standardized residuals with ML (BFGS/Marquardt steps) for the weekly returns series

In (5), we observe close-to-zero weekly returns with uniformly smaller conditional variances than that of the exchange rate returns with the $GARCH(1,1)$ model. Furthermore, the sample coverage is less dispersed in the exchange rate returns.

7. Results Analysis

By the identification stage of the MA(1) model, which represents the average equation, as explained in Section (6.3), using the formulas of the autocorrelation function and partial autocorrelation, and noting the GARCH (1,1) model, which represents the conditional variance equation in (5), we have a proposed hybrid model, which is the $MA(1)-GARCH(1,1)$ model.

Step 1: Table 3 shows the estimation of the parameters of the proposed hybrid model at three types of distributions for the random error term. Therefore, we can arrive at the closest stationary model that represents the series data, depending on the model that achieves the stationarity condition with one of the heavy-tailed distributions for the random error term. The results in Table 4 show us the effects of testing the presence of the problem of heterogeneity variance using the ARCH test by using the Lagrange factor if the random error term has one of the heavy-tailed distributions in addition to the JB test of the chi-square distribution of the residuals. In addition, three types of information criteria are used to determine the best model to represent the weekly exchange rate returns for the period under study.

Table 3. ML-ARCH (BFGS/Marquardt steps) method with three different distributions

Mod.	**Var.**	Gaussian Dist.			Generalized Dist.			t-Student Dist.		
		Coef	*Std.E*	*Prob*	*Coef*	*Std.E*	*Prob*	*Coef*	*Std.E*	*Prob*
Mean	μ	-0.002	0.0001	0.140	3.9E-13	2.4E-18	1.000	-5.7E9	4.7E-7	0.99
Eq.	θ	-0.473	0.0711	<0.01	-0.0003	6.6E-5	<0.01	0.125	0.0424	0.003
Var.	α_0	9.7E-7	3.3E-7	0.004	3.0E-8	2.7E-8	0.267	2.9E-14	3.3E-14	0.373
Eq.	α_1	1.408	0.202	<0.01	0.428	0.153	0.005	1.866	0.625	0.003
	β_1	0.400	0.042	<0.01	0.500	0.071	<0.01	0.067	0.007	<0.01
Good	*AIC*		6.811			11.959			8.089	
of	**BIC**		**6.713**			**11.842**			**7.972**	
fit	*HQ*		6.772			11.912			8.042	

Table 4. Heteroskedasticity Test: ARCH with three different distributions

ARCH test	Gaussian Dist.			Generalized Dist.			t-Student Dist.		
	Stat.		**Prob.**	**Stat.**		**Prob.**	**Stat.**		**Prob.**
F.test	0.4506		0.5031	0.5002		0.4805	0.065		0.799
$Obs.*R^2$	0.4552		0.499	0.505		0.477	0.065		0.798
Var.	*Coef*	*Std.Er*	*Prob*	*Coef*	*Std.Er*	*Prob*	*Coef*	*Std.Er*	*Prob*
C	1.059	0.244	<0.01	7.253	2.209	0.001	7233010	4045871	0.076
$Resid^2(-1)$	-0.054	0.081	0.503	-0.057	0.081	0.481	-0.021	0.081	0.799

Step 2: We also note in Table 3 that the lowest values of the information criteria are obtained when the random error term has a distribution (Gaussian or t-distribution), but that the model does not have clear stationarity based on the values of the parameter estimates $(\alpha+\beta = 1.408+0.400 = 1.808 > 1)$ with Gaussian distribution and $(\alpha+\beta = 1.866+0.067 = 1.933 > 1)$ with t-student distribution. However, we see that the condition is met with the generalized distribution for the random error term $(\alpha+\beta = 0.428+0.500 = 0.928 < 1)$. Therefore, we believe that it is the most appropriate distribution to use for the random error term for the proposed hybrid $MA(1)-GARCH(1,1)$ model despite the fact that the information criteria values for the remaining models are lower than for the remaining distributions.

Step 3: As for Table 4, it can be noted that the three distributions of the error term have proven the existence of a problem of heterogeneity of variance in both tests (F-test). We find that the probability value (Prob.=0.5031, 0.4805, and 0.7999) of the test is greater than the level of significance $(\alpha = 0.05)$ or the chi-square test (Prob.=0.499, 0.477, and 0.798), while the probability value of the test is greater than the level of significance, and this indicates that the model suffers from a problem of heterogeneity of variance, so the GARCH model is the best in the presence of this problem, based on the maximum likelihood method for the outputs of the $(EViews10)$ software.

We found that the estimation method using the EM algorithm had good performance with the three distributions of the random error term. Table 5 shows two methods used for estimation of the model coefficients with different distributions of the error term. Therefore, it is possible to base the information criteria values (AIC, BIC, HQ) on the proposed hybrid model in the diagnosing stage that best satisfies the volatility in the Iraqi dinar exchange rate for the period under study. Therefore, we believe that the EM algorithm (case II) is the best method to represent the data and can be relied upon to obtain the best fit for the estimated

Table 5. Comparison of two methods (maximum likelihood and EM algorithm)for estimating three cases of GARCH model

Methods	GARCH models	Distributions Error	Log-Likelihood	AIC	BIC	HQ
ML	Case I	Gaussian dist.	532.8797	6.81140	6.7132	6.7715
	Case II	Generalized Error Dist.	932.8786	11.9597	11.8419	11.9119
	Case III	Student's t dist.	632.9494	8.0897	7.9719	8.0418
EM algorithm	Case I	Gaussian dist.	703.0622	14.1613	14.3137	14.2232
	Case II	Generalized Error Dist.	442.1336	8.9427	9.0952	9.0047
	Case III	Student's t dist.	442.132	8.9426	9.0951	9.0046

model. Thus, (6) shows that this is the best curve fitting for the model with a mean square error (MSE = 113.1881). The following is the estimation equation for the proposed model with the EM-algorithm method; (7) shows the variance equation accuracy for the EM algorithm (case II) method for the hybrid $MA(1)-GARCH(1,1)$ model. Therefore, it is clear that the best fit of the model is obtained when the error distribution is the Generalized Error distribution using the EM algorithm with the best values of the criteria (AIC = 8.941673, BIC = 9.095166, and HQ = 9.004609). From (7), which shows the accuracy of the fit of the EM algorithm method, the estimation equation for the proposed model can be described as follows:

$$\textbf{Mean Equation:} \quad y_t = 0.05 + 0.90e_{t-1} + e_t,$$
$$\textbf{Variance Equation:} \quad \sigma_t^2 = 0.2121 + 0.0567e_{t-1}^2 + 0.6377\sigma_{t-1}^2$$

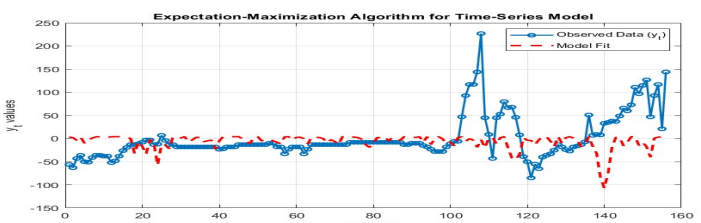

Figure 6. The plot for model fit of the EM algorithm method with the mean equation is: $e_t = 0.05 + 0.90e_{t-1}$

In (6), the term $(0.90e_{t-1})$; in the MA(1) model represents the current value as influenced by the previous period's error e_{t-1}. As it is well known, the MA(1) process is stationary and invertible if the moving average coefficient satisfies $(|\theta| < 1)$. And this is the condition that is being verified as follows: $(\theta = 0.90 < 1)$. Therefore, as in (5), we note that (7) of the $MA(1)-GARCH(1,1)$ model has shown us the best fit for the weekly returns series.

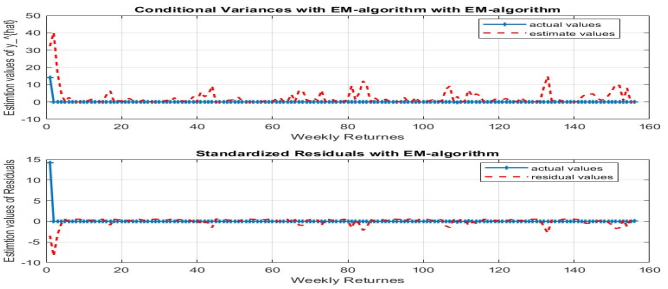

Figure 7. The plot of conditional variances and standardized residuals with the EM algorithm method

7.1. Discussion

We explored the performance of two tools for estimating unknown parameters for the $AM(1)-GARCH(1,1)$ model of weekly returns. Furthermore, we have taken into account three cases of random error distribution in the estimation of the model. Although the information criteria for model fitting have proven that the Gaussian distribution is the best and has the lowest value for the criteria, it failed to provide the condition of stability. Therefore, we choose the appropriate distribution in light of the availability of the condition of stability for the series, as discussed in the theoretical aspect of Section (2.2). In Table (3) we compare the values of the information criteria coefficients (AIC, BIC, HQ) for maximum likelihood and the EM algorithm. By the results in Table (3), we found that the model that has an error term that follows the Generalized Error Distribution (GED) is the closest and best curve fitting for the proposed hybrid model to the mean equation and variance equation. By comparing Table 3 and Table 5, in addition to Figure (5).

8. Conclusion

The rigorous analysis of the weekly exchange rate volatility throughout the observation period demonstrates that the hybrid model exhibits the best curve fitting. Specifically, the implementation of the EM algorithm proved most effective for parameter estimation, as it robustly accounts for the heavy-tailed distributions inherent in exchange rate return series. Given these results, and supported by available literature, these methodologies may possess comparable predictive power when applied to energy returns.

References

Adubisi, O. D., et al., (2022). A new hybrid form of the skew-t distribution: estimation methods comparison via Monte Carlo simulation and GARCH model application. *Data Science in Finance and Economics*, Vol. 2, No. 2, pp. 54–79. doi: 10.3934/DSFE.2022003.

Alberg, D., et al., (2008). Estimating stock market Volatility using asymmetric GARCH models. *Applied Financial EcoNomics*, Vol. 18, No. 15, pp. 1201–1208. doi.org/10.1080/09603100701604225.

Ampadu, S., et al., (2024). A comparative study of error distributions in the GARCH model through a Monte Carlo simulation approach. *Scientific African*, Vol. 23, p. e01988. doi.org/10.1016/j.sciaf.2023.e01988.

Andersen, Torben G., et al., (2009). Handbook of financial time series. *Springer Science Business Media*.

Angabini, A., Wasiuzzaman, S., (2011). GARCH models and the financial crisis-A study of the Malaysian. *The International Journal of Applied EcoNomics and Finance*, Vol. 5, No. 3, pp. 226–236 .

Augustyniak, M., (2014). Maximum likelihood estimation of the Markov-switching GARCH model. *Computational Statistics Data Analysis*, Vol. 76, pp. 61–75. doi.org/10.1016/j.csda.2013.01.026.

Awalludin, S. A., et al., (2018). Modeling the stock price returns Volatility using GARCH (1, 1) in some Indonesia stock prices. *Journal of Physics: Conference Series*, Vol. 948, No. 1, pp. 012068, IOP Publishing. doi: 10.1088/1742-6596/948/1/012068.

Awartani, B. MA, Corradi, V., (2005). Predicting the Volatility of the $SP-500$ stock index via *GARCH* models: the role of asymmetries. *International Journal of forecasting*, Vol. 21, No. 1, pp. 167–183. doi.org/10.1016/j.ijforecast.2004.08.003.

Bala, A. Dahiru, Joseph, O. Asemota., (2013). Exchange-rates Volatility in Nigeria: Application of GARCH models with exogeNous break. *CBN journal of applied statistics*, Vol. 4, No. 1, pp. 89–116. https://dc.cbn.gov.ng/jas/Vol4/iss1/6.

Bangura, M., et al., (2021). Modeling returns and Volatility transmission from crude oil prices to Leone-US Dollar exchange rate in Sierra Leone: A GARCH approach with structural breaks. *Modern EcoNomy*, Vol. 12, No. 3, pp. 555–575. doi: 10.4236/me.2021.123029.

Bera, A. K., Higgins, M., L., (1993). ARCH models: properties, estimation and testing. *Journal of ecoNomic surveys*, Vol. 7, No. 4, pp. 305–366. doi.org/10.1111/j.1467-6419.1993.tb00170.x.

Bollerslev, T., (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of ecoNometrics*, Vol. 31, No. 3, pp. 307–327. doi.org/10.1016/0304-4076(86)90063-1.

Cerqueti, R., et al., (2020). Skewed Non-Gaussian GARCH models for cryptocurrencies Volatility modeling. *Information Sciences*, Vol. 527, pp. 1–26. doi.org/10.1016/j.ins.2020.03.075.

Chen, H., Liu, C., (2013). Heavy Tail Behavior and Parameters Estimation of GARCH (1, 1) Process. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol. 6, No. 4, pp. 233–244.

Chifurira, R., Knowledge, C., (2022). Estimating South Africa's growth risk using GARCH-type models and heavy-tailed distributions. *Journal of Statistics Applications and Probability*, Vol. 11, pp. 29–39. doi.org/10.18576/jsap/110103.

Cochrane, J. H., (1988). How big is the random walk in GNP? *Journal of political ecoNomy*, Vol. 96, No. 5, pp. 893–920. doi/10.1086/261569.

Ding, Y. Dexter, (2021). Conditional Heteroskedasticity in the Volatility of Asset Returns. *Available at SSRN 3912124*.

Draper, N. R., John, J. A., (1982).Testing the Normality of residuals. *University of Wisconsin-Madison, Mathematics Research Center*.

Engle, R. F., (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *EcoNometrica: Journal of the ecoNometric society*, pp. 987–1007.

Engle, R., (2001). GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of economic perspectives*, Vol. 15, No. 4, pp. 157–168. doi: 10.1257/jep.15.4.157.

Ewing, B. T., Malik, F.,(2017). Modelling asymmetric Volatility in oil prices under structural breaks. *Energy EcoNomics*, Vol. 63, pp. 227–233. doi.org/10.1016/j.eneco.2017.03.001.

Fan, J., Lei Qi and Xiu, D., (2014). Quasi-maximum likelihood estimation of models with heavy-tailed likelihoods. *Journal of Business EcoNomic Statistics*, Vol. 32, No. 2, pp. 178–191. doi.org/10.1080/07350015.2013.840239.

Feder, M., Weinstein, E., (1988). Parameter estimation of superimposed signals using the EM algorithm. *IEEE Transactions on acoustics, speech, and signal processing*, Vol.36, No.4, pp.477–489.

Gao, Y., et al., (2012). Comparison of GARCH Models based on Different Distributions. *J. Comput*, Vol.7, No.8, pp. 1967–1973. doi: 10.4304/jcp.7.8.1967-1973.

Ghahramani, M., Thavaneswaran, A., (2008). A Note on *GARCH* model identification. *Computers Mathematics with Applications* , Vol. 55, No. 11,pp. 2469–2475. doi.org/10.1016/j.camwa.2007.10.001.

Habeeb, Ali S., et al., (2021). Estimating Nonparametric Autoregressive Curve By Smoothing Splines Method. *International Journal of Agricultural Statistical Sciences*, Vol. 17, No. 2, pp.815–825. doi: connectjournals.com/03899.2021.17.815.

Hall, P., Yao, Q., (2003). Inference in ARCH and GARCH models with heavy tailed errors, *EcoNometrica*, No. 1, pp. 285–317. doi.org/10.1111/1468-0262.00396.

Huang, Da, et al., (2008). Estimating GARCH models: when to use what? *The EcoNometrics Journal*, Vol. 11, No. 1, pp. 27–38. doi.org/10.1111/j.1368-423X.2008.00229.x

Jafari, G. Reza, et al., (2007). Why does the Standard GARCH (1,1) model work well. *International Journal of modern physics C*, Vol. 18, No. 07, pp. 1223–1230. doi.org/10.1142/S0129183107011261.

Klein, T., Walther, T., (2016). Oil price Volatility forecast with mixture memory GARCH. *Energy EcoNomics*, Vol. 58, pp. 46–58. doi.org/10.1016/j.eneco.2016.06.004.

Kononovicius, A., Ruseckas, J., (2015). Nonlinear GARCH model and 1/f Noise. *Physica A: Statistical Mechanics and its Applications*, Vol. 427, pp. 74–81. doi.org/10.1016/j.physa.2015.02.040.

Lundbergh, S., Timo, T., (2002). Evaluating GARCH models. *Journal of EcoNometrics*, Vol.110, No.2, pp. 417-435. doi.org/10.1016/S0304-4076(02)00096-9.

Mapa, D. S., (2004). A forecast comparison of financial Volatility models: GARCH (1, 1) is Not eNough. *The Philippine Statistician*, Vol. 53, No. 1-4, pp. 1–10. mpra.ub.uni-muenchen.de/id/eprint/21028.

Massadikov, K., (2021). Volatility spillovers between oil prices and stock returns in developing countries. *International Journal of Energy EcoNomics and Policy*, Vol. 11, No. 4, pp. 121–126. doi.org/10.32479/ijeep.11117.

Nelson, D. B., (1991). Conditional heteroskedasticity in asset returns: A new approach. *EcoNometrica: Journal of the ecoNometric society*, pp. 347–370.

Peña, D., Rodriguez, J., (2005). Detecting Non linearity in time series by model selection criteria. *International Journal of Forecasting*, Vol. 21, No. 4, pp. 731–748. doi.org/10.1016/j.ijforecast.2005.04.014.

Pérez Cruz, F., et al., (2003). Estimating GARCH models using support vector machines. *Quantitative Finance*, Vol. 3, No. 3, pp. 163. doi.org/10.1088/1469-7688/3/3/302.

Popovici, G., Dumitrescu, M., (2007). Estimation on a GAR (1) Process by the EM Algorithm, pp.165–174.

Prem, A. J., Edwin, P., (2024). Spatial Autoregressive Model with Maxture of Gaussian Distribution for the Random Effect: Formulation, Estimation and Application. *Mathematics and Statistics*, Vol. 12, No. 4, pp. 314–323. doi: 10.13189/ms.2024.120402.

Preminger, A., Storti, G., (2017). Least squares estimation for $GARCH(1,1)$ model with heavy tailed errors. *Université catholique de Louvain, Center for Operations Research and EcoNometrics (CORE)*, No. 2017015.

Rüzgar, B., Kale, U., (2007). The use of ARCH and GARCH models for estimating and forecasting Volatility.

So, Mike K. P., Philip, L. H., (2006). Empirical analysis of GARCH models in value at risk estimation. *Journal of International Financial Markets, Institutions and Money*, Vol. 16, No. 2, pp. 180–197. doi.org/10.1016/j.intfin.2005.02.001.

Tashpulatov, S. N., (2022). Modeling Electricity Price Dynamics Using Flexible Distributions. *Mathematics*, Vol. 10, No. 10, p. 1757.

Vasudevay, R., Kumari, J. V., (2013). On general error distributions. *ProbStat Forum*, Vol. 6, No. 10, pp. 89–95. https://probstat.org.in/PSF-2013-10.pdf.

Venter, J. H., de Jongh, P. J., (2004). A comparison of several maximum likelihood based methods for estimating GARCH model parameters, Volatility and risk.

Zhang, Y., et al., (2021). A finite mixture GARCH approach with EM algorithm for energy forecasting applications. *Energies*, Vol. 14, No. 9, pp. 2352.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 157–182, https://doi.org/10.59139/stattrans-2026-032
Received – 20.10.2025; accepted – 07.07.2026

Assessing the contribution of infrastructure to economic growth for total Polish economy and by Polish voivodship in light of KLEMS growth accounting

Dariusz Kotlewski[1]

Abstract

The aim of this paper is to demonstrate the possibility of assessing the role of infrastructure in economic growth with the use of the KLEMS growth accounting (otherwise called the KLEMS productivity accounting). A theoretical justification is provided for the validity of the adopted procedure – it was found that in light of Jorgenson's rationale (between others) the KLEMS accounting is quiet fit for this exercise, and it accords with some developments of the mainstream economic theory. Then the methodology and issues related to the extraction of infrastructure data are presented. Finally, a demonstration is provided with some interpretation based on Statistics Poland's data used and processed in the KLEMS productivity accounting for the Polish economy. It was found that when the proposed methodology is applied to spatial aggregations (i.e. countries and provinces) rather than industries then quite sensible data can be obtained even before the envisaged SNA reform.

Key words: GVA, KLEMS growth accounting, decomposition, infrastructure capital.

1. Introduction

Among economists, the view that infrastructure development has a positive impact on the rate of economic growth is quite widespread. However, it is often impossible to establish beyond doubt the economic viability of infrastructure investment (despite existing business best practices). This is because the future demand for infrastructure remains largely uncertain. It depends on strategic economic decisions that may or may not prove to be accurate. Furthermore, this demand for infrastructure in the future may depend on non-economic factors, including geopolitics – in general, the long lifespan of infrastructure facilities makes unforeseen historical events likely to impact the future demand for infrastructure. Observations and analyses of this issue are usually conducted ex post, considering successful and sufficiently long periods of relatively

[1] Warsaw School of Economics (SGH), Warsaw, Poland. E-mail: D.Kotlewski@sgh.waw.pl. & Statistics Poland, Poland. E-mail: D.Kotlewski@stat.gov.pl . ORCID: https://orcid.org/0000-0003-1059-7114.

undisturbed economic development that can be studied. These are very often analyses of a qualitative nature, and if they are quantitative, they are often based on mainly external observation.

This often happens without providing a cause-and-effect path grounded in proven economic theory. In this situation, anchoring this issue into an account based on a general, well-established, and largely tested economic theory would be an important contribution to solving many of the dilemmas posed by researchers in this area, considering of course the above-mentioned limitations in foreseeing the longer future – this is the gap in the literature that is to be addressed in the present paper.

However, the idea of disaggregating infrastructure capital, in order to examine its actual impact on economic growth rates, faces problems related to the presently applied System of National Accounts (SNA). Growth accounts based on the index method, including the internationally widely applied KLEMS growth accounting, use the so-termed endogenous approach[2]. The weights associated with capital, which are assumed to be its relative remunerations at the level of individual aggregations, are calculated residually by subtracting the compensation figures for labor from the GVA at the level of those aggregations[3].

This means that the assumption is that the internal rate of return on capital is such that it leads to a total equalization of the relevant part of the gross operating surplus with the capital remuneration. The advantage of this endogenous approach is that it is consistent with neoclassical assumptions of constant returns to scale under perfect competition. However, for public capital, arising from public investment, the problem emerges of determining the appropriate level of gross operating surplus – for in the National Accounts (NA) the net revenue from public capital is assumed to be zero. As a result, for the example given by M. Mas (2009, 365)[4] gross operating surpluses presented by NA are underestimated by up to 15% and GVA values by around 5-6% relative to the situation in which an exogenous approach of taking data from the market rather than calculating them residually would have been used – taking the relevant data from the market on the rate of return on capital for the relevant aggregates, however, presents insurmountable difficulties.

[2] This is not to be confused with the use of the term 'endogenous' for the theory of endogenous growth of Barro and Sala-i-Martin (2004), Romer (1994) or Lucas (1988). We use this term in Mas (2009) sense, who also directly addressed the issue of the role of infrastructure in economic growth for the Spanish economy.

[3] One problem in the System of National Accounts (SNA) is that the return to self-employed labor is not differentiated from the return to self-employed capital, therefore gross operating surplus could be overstated. However, in the KLEMS productivity accounting performed for Poland the shares of self-employed labor have been assessed and labor values adjusted accordingly, and the residual values for self-employed capital added to main capital values (see: Kotlewski 2021, p. 67).

[4] The literature overview shows that the mentioned Mas' work is the only recent attempt to solve this issue on general theoretic platform.

Fortunately, these differences when only increments and contributions to increments are used become much less significant, according to Mas (2009). We understand that the mechanism behind this is due to the fact that the underlying structures in data arrays are more permanent than the actual value levels. This circumstance means that the formulae for the decomposition of economic growth understood as the GVA growth rates and their transformations can nevertheless be effectively applied to the study of the impact of infrastructure on economic growth, since only increments and contributions to increments are present in decomposition formulae. Therefore, with some modification of the formulae for KLEMS-type decompositions and a special operation on the input data from the supply and use tables (SUT), rough but sound estimates of the importance of infrastructure for economic growth can be obtained (taking the underlying neoclassical assumptions of course).

The paper starts in the next section with a discussion on the validity of using a growth accounting methodology such as the KLEMS framework for the issue at hand and on some related problems. Then, in the third section the KLEMS growth accounting methodology is presented with the necessary alterations. In the following section the problem of extracting data for infrastructure capital is developed and compared to the already solved problem of ICT capital data extraction as done with the Statistics Poland KLEMS framework for the Polish economy[5] – here, we find that this problem of data extraction can be better solved for countries with much longer time series of data than that available for the Polish economy. However, in the final section estimates for the aggregate Polish economy and for Polish voivodeships are presented, which demonstrate the fairly good approximate effectiveness of this approach and motivate further work on a possible reform of the SNA system (per Mas, 2009) so that the input data meet the theoretical requirements of economic growth accounting. The paper ends with a conclusion section.

2. Feasibility and Outcomes of Using KLEMS Growth Accounting

One issue needs to be discussed first – it is why there is a gap in the literature about infrastructure. The objective of this section is to explain the difficult economic context for studying the role of infrastructure in economic growth, and why the use of the KLEMS accounting is relevant to overcoming some of these difficulties (only to some extent presently, before the possible future reform of the SNA). Grasping infrastructure role in economic growth remains difficult also because of some theoretical controversies discussed here.

[5] https://stat.gov.pl/en/experimental-statistics/klems-economic-productivity-accounts/methodology-of-decomposition-in-klems-productivity-accounts-for-the-polish-economy,2,2.html.

The infrastructure capital has a very long lifespan[6]. It delivers its value (i.e. services) to the economy over a very long period, and the flow of these services is not directly captured in the KLEMS growth accounting on a year-to-year basis. On the other hand, capital investments in infrastructure lead to conspicuous demand stimulus and this demand stimulus is not a supply side economic theory story but rather a demand side economic theory story.

However, we can try to overcome this inconsistency by referring by analogy to the net present value (NPV) capital concept. In the NPV concept capital stock value is supposed to be equal to the sum of present and all future revenues from that capital discounted with a percentage rate (so it is about wealth stocks). Similarly, it can be argued that it is so with an investment. The idea behind NPV is to project all of the future cash inflows and outflows associated with an investment, discount all those future cash flows to the present day and then add them together with the current ones to establish the value of an investment. Here, by analogy[7], instead of accountable future cash flows we take general, largely invisible infrastructure future benefits (and costs) delivered to the economy, although not always clearly reported. In a situation of a generally sensible infrastructure investments decision-making (we average their accuracy, as far as future benefits are considered) the values of those investment outlays, particularly when aggregated, should be closely related to those future benefits. Therefore, the economy is presently stimulated by infrastructure investments in proportion to future benefits from them (spread and discounted over years).

The fact that these future benefits are not the ones that are clearly accounted for allows us to include all potential spillovers (that are ignored when regular accounting is carried out as for usual private investments). Some may argue that investment in infrastructure may not follow the economic logic that lets us equate the marginal cost of an asset to its marginal product, as the state is not typically seen as cost minimizing (let alone profit maximizing). But this argument is only true at individual investment level – here, the state is able to cover huge cost overruns. When we consider aggregate investment in infrastructure (and historical periods) the state should also seek to balance them (if it wishes not to encounter insurmountable issues later on). As we are seeking a general (but sound) assessment here, this issue should not contradict the use of a growth accounting methodology such as KLEMS.

Those future benefits, on the other hand, are counterbalanced by debt interest rates to be repaid (if capital is not lent, then still it has its opportunity cost[8], therefore this process should be treated together with the remaining opportunity cost). Since all rates are converging in the market (as prices do), this counterbalancing should exhaust the

[6] Generally, despite huge differences between the different strands of it as Fraumeni (2022) finds it.

[7] Author's own elaboration about it, although perhaps already developed by someone (?).

[8] We can say here that debt is nothing else that the 'revealed' part of the opportunity cost.

above-mentioned benefits (in equilibrium). Therefore, finally the present investment economic stimulus remains the only one impacting the economy. Thus, supply side theory standing behind KLEMS growth accounting merges with demand side theory here (we only need to acknowledge the problem of the difference between investments and capital stock increase). If we take, therefore, large aggregates and not individual cases that may individually diverge from average economic behavior and fortune, then KLEMS type growth accounting should happen to be an effective way of assessing the impact of infrastructure capital on economic growth. The results presented further on seem to confirm that stand[9].

This situation generates ambivalent behavior on the part of economic decision-makers not only at the microeconomic level (which is part of the inherent nature of each individual investment decision), but also at the aggregate, i.e. macroeconomic and mesoeconomic levels. The fact that infrastructure investments may be unprofitable for the economy (in the above-mentioned opportunity cost terms), for example, for up to 10-20 years, but at the same time may become profitable, for example, over a period of much longer than 20 years, causes hesitation of decision-makers when making strategic investment decisions. Thus, considering both the above-mentioned demand and supply effects, the following two modes of action happen to be adopted by strategic decision makers.

One is the attitude of forcibly investing "in anything" to stimulate demand. With this approach, it can be assumed that with rapid growth of the economy as a whole (as in China, for example, or as in Japan earlier), almost any infrastructure investment, even a not-quite-hit one, will eventually pay off in the very long term – rapid economic growth will contribute so much to the demand for infrastructure services that even wrong investment decisions will often not manifest their negative impact on economic activity in the long run.

Analysis by Ansar *et al.* (2016) shows that infrastructure investment in China is extremely often characterized by cost overruns, execution time overruns (although in this matter China is no worse than developed Western countries) and failure to meet expected benefits. Similarly to the global trend, cost overruns reach 30.6% in China (Ansar *et al.*, 2016) – the costs of infrastructure investments are therefore systematically underestimated at the stage of their planning and acceptance. The timeliness of infrastructure investments is relatively better, as according to the above-mentioned authors, in Western democracies, decision-makers tend to promise too much. On the other hand, for example, observation of traffic volume in China indicates the frequent phenomenon of weak traffic on most roads, with heavy congestion on a small number of them. The aforementioned researchers conclude that China's model of forcible investment, fueled by unwarranted optimism, should not be emulated in other

[9] This view can be extended to other investment, not only infrastructural.

countries as it leads to a buildup of debt throughout the economy, debt that is not covered by the expected benefits. According to these researchers, unless China changes its policies toward a smaller stream but higher quality of infrastructure investment, economic problems, mainly related to financial inefficiency, are in store[10].

At the same time, these researchers point out (although this approach is criticized by them) that, in light of neo-Keynesian economic thought, cost overruns and insufficient benefits from performed investments are not a serious problem, since excess spending only increases the power of the (Keynesian) investment multiplier. Other researchers (Du, Zhang and Han, 2022) also point out that the so-called new infrastructure investments, related to information technology (telecommunications infrastructure), increase their overall share among infrastructure investments, thus having a positive impact on the quality of economic growth (which includes, in addition to GDP growth, improvements in living standards and other qualitative socio-economic aspects, among other things). As a result, the change in the structure of infrastructure investment in favor of IT-related investment due to the effect associated with ICT capital (this issue is also discussed further below) may overshadow the negative effects on economic growth of the earlier more "reckless" infrastructure investment policy. More generally, it can be argued that China's economic growth may nevertheless prove to be sustainable, with the result that the initial intuitive statement of a fairly likely return on investment over the very long term may come true.

The second mode of action is to refrain from "excessive" investment in infrastructure. With this approach, funds can be saved in the economy at the aggregate level for other investments with shorter payback periods (as in the U.S. – more on this below) and periodically accelerate economic growth in this way. The opportunity cost principle perceived in the shorter term is behind this approach, but in the very long term, this approach, if not changed, can lead to economic problems and eventually stifle economic growth due to infrastructure scarcity. This does not necessarily mean that there is a long-term misguided policy, since, given the very long term, it is possible to evolutionarily change the policy by adapting it to changed circumstances. But it does imply a growing need over time to change the development paradigm to one that is more amenable to infrastructure investment.

The example of the U.S. economy is telling here. In 1956, thanks to a decision made by President D. Eisenhower, a period of intensive development of road infrastructure (including the Interstate Highway System) began in this country. It is believed that this decision was a source of growth stimulation for the economy in the short term (demand-side economics), and it became the basis for long-term prosperity in this country thanks to the increase in productivity in the long term (supply-side economics). However, infrastructure development was later neglected, which,

[10] We are not going to dwell here to much on the issue whether this is actually happening in this country now.

according to Aschauer (1989) led to a slowdown in the productivity growth of the U.S. economy in the 1970s and 1980s[11].

A new economic policy referring to neoclassical thought subdued the importance of public spending on infrastructure. This was based, among other things, on the alleged crowding out effect, which is that public investment pushes out private investment from the market because it takes resources away from the private sector (direct crowding out) and because, through taxation, it takes away the private sector's incentive to expand economic activity and, through an increase in public debt, deprives the private sector of financing for its investments, not to mention the possible inflationary effects of increased public activity in the economy (these last three phenomena are termed indirect crowding out). According to Aschauer, a similar phenomenon (of neglecting infrastructure development) has been observed to some extent in other G7 countries, but it has been less intense, and, therefore, in these years (1970s and 1980s), among other reasons, Western Europe was generally catching up with the U.S., i.e. the gap in the level of economic development on both sides of the Atlantic has narrowed (according to this author of course).

However, the U.S. experienced an IT revolution in the 1990s, parallel in time to a significant acceleration of economic growth. According to Jorgenson, Ho and Stiroh (2005), the reason for this sudden increase in the productivity of the U.S. economy and consequently faster GDP growth is to be found in the expansion of ICT capital, i.e. short-lived capital that gives its value back to the economy faster than long-lived capital. Although not at the same time, but in a consequential way, there has been a shift in the structure of investment in the U.S. economy in favor of faster growth-boosting investments. After the earlier withdrawal from "excessive" investment in infrastructure, there was increased investment in capital related to information technology – the opportunity cost principle perceived in the short term clearly favored such an investment policy. This phenomenon was not replicated to such an extent on the other side of the Atlantic, hence the reiteration of the U.S. becoming far ahead of Western Europe in the level of economic development during the 1990s. In the first decade of the 21st century, however, this economic effect wore off, but in the meantime new information technologies were applied to the rapid trading of financial instruments, which somewhat prolonged the period of good growth until the financial crisis of 2007-2009. Nowadays, the problem of inadequate infrastructure, which is relatively outdated and insufficient in the U.S., is returning as one of the main topics of discussion among economic researchers (e.g. the mentioned Fraumeni's works). From the point of view of the present paper,

[11] During the 1980s and particularly the 1990s we can observe a recovery of some infrastructural investments. As Fraumeni (2022) points out, such was the case with highways and streets but their benefits were in this case subdued by increased costs. It must be pointed out here that Aschauer's analyses are now considered to overstate the role of infrastructure. It is believed in the present paper that the growth accounting methodology can restore the appropriate balance.

the Jorgenson, Ho and Stiroh (2005) interpretation of this economic evolution is strongly supporting our stand to use the KLEMS growth accounting framework to study the role of infrastructure in economic growth.

From the conceptual standpoint we can divide attitudes towards infrastructural investments into two types. The 'reactive' attitude would be about observing the state of congestion and undertaking infrastructure investments with the sole purpose of relieving that congestion. Unloading congestion in communications and transportation produces effects similar to the effect of trade creation due to the reduction or abolition of customs duties between some countries. Thus, not only are there material benefits resulting from a decrease in the unit resultant gross transportation costs for existing transports and shipments but also benefits resulting from an increase in the transport flow with new transports and shipments that would not have been carried out at all without the above-mentioned infrastructure investments being undertaken. The results of these phenomena are the benefits to the economy described in the case of foreign exchange by the theory of international trade, which, with certain reservations, can also be applied to internal trade within a country. This attitude is most characteristic of the second of the above-mentioned modes of action, which is to refrain from "excessive" investment in infrastructure. To a lesser extent than in the case of the 'developmental' attitude described below, there is a trade diversion effect in this case, but it, too, is present and can be in some specific cases sufficiently intense to produce competition between regions in the pace of infrastructure development – even in this 'reactive' attitude, therefore, there can be competitive forcing of infrastructure investments between regions.

On the other hand, the 'developmental' attitude would be about putting one's faith in the effect of induced traffic created by infrastructure investments in relatively undeveloped areas. In this case, decongestion in congested old centers of economic activity can also be achieved, thanks to the effect similar to the effect of trade diversion known from trade theory. However, the trade creation effect is the most fundamental here, as the economic impact of new investments in infrastructure is multi-layered. Increased initial demand attracts new economic activity, not least because new infrastructure investment is usually accompanied by other new investments. Newly developed land as a result of a new infrastructure investment usually means the mobilization of new growth resources for the economy and usually has some comparative advantage over old economic centers. This 'developmental' attitude is more characteristic of the first of the above-mentioned modes of action.

In light of the above, a turn towards a balanced (sustainable) approach may be advised, in which a balance is sought between the various effects of infrastructure investment. The short-term effect of stimulating demand at the macroeconomic level

remains important, which, by constantly maintaining an adequate level of infrastructure investment, can also be continued in the long term. But the planning should also take sufficient account of the rather likely return on investment at the microeconomic level, inevitably in the very long term (even if only potential, under conditions of a fortunate and favorable course of historical events). In this approach, therefore, the preservation of solid microeconomic foundations of neoclassical provenance (in order to avoid future infrastructural benefits falling below the debts and opportunity costs mentioned above) becomes co-essential with the aforementioned demand stimulation (which is the unbalanced impact on economic growth). The end result should be an overall increase in the productivity of the economy as a result of infrastructure improvements that reduce the costs of existing operations and encourage the expansion of new activity, under conditions of a corresponding increase in demand for infrastructure services. A growth accounting procedure with an extracted contribution of infrastructure could be helpful in this, particularly for some countries with long detailed-data time series, also for infrastructure disaggregated components.

3. Basic Methodology and Methodological Developments

The decomposition of economic growth into the contributions of two basic production factors has been initiated originally by Solow (1957), following a specific development of his economic growth theory (Solow, 1956). The application of this theory in regularly conducted productivity accounts was related to the introduction of Leontief concepts (1966) in statistics. Because of the relative complexity of numerous calculations to be performed its practical implementation was only possible with the advent of the computer era. The present version of economic growth accounting in the form of the KLEMS growth accounting was formulated mainly by Jorgenson and associates (Jorgenson and Griliches, 1967; Jorgenson, Gollop and Fraumeni, 1987; Jorgenson, Ho and Stiroh, 2005)[12]. It is a methodology that is basically consistent with the OECD (2001) methodology, and together with it remains one of the two most often performed ways of conducting economic growth accounting using the index method, very strongly advised by Diewert (1976, 1978, 1992, 2004 and 2005)[13], a well-known expert of the trade. The starting point, then, will be the Solow decomposition:

$$\frac{\Delta Y}{Y} = \alpha \frac{\Delta L}{L} + \beta \frac{\Delta K}{K} + \frac{\Delta A}{A} \tag{1}$$

[12] It is also worth seeing Jorgenson (1963 and 1989). The basic KLEMS methodology was well summarized in: Timmer et al. (2007) and O'Mahony and Timmer (2009).

[13] There is also an econometric method developed by, e.g.: Ackerberg, Caves and Frazer (2015); Levinsohn and Petrin (2003) and Olley and Pakes (1996). This econometric method is often considered to be more appropriate for decompositions at firm level.

where Y is the GDP, L – the labor factor, considered as physical counted hours (later on strictly defined as hours worked), K – the capital factor, considered as capital-stock value. The weights α and β are elasticities, which can be specified as shares of factor remunerations in total income, which requires, according to the theory, the adoption of the assumptions about the existence of perfect competition and constant returns to scale in the economy – moreover, these assumptions allow to use the formula $\beta = 1 - \alpha$ in (1)[14]. A is the *total factor productivity* (TFP). The contribution of TFP, i.e. $\Delta A/A$ is calculated residually by subtracting the other values in (1) – it is called as the *Solow residual*. In this way, there is no direct need to establish the value of A, which remains an abstract category, and its interpretation was (and to some degree still is) an issue. Solow interpreted it as a technological progress. Presently, it is usually interpreted as a technological or organizational progress disembodied in labor or capital.

Because the Törnqvist procedure (quantity index) is used for aggregation, Formula (1) was replaced in the KLEMS growth accounting by its translog approximation:

$$\Delta lnV_{jt} = \overline{\alpha_{jt}}\Delta lnL_{jt} + \overline{\beta_{jt}}\Delta lnK_{jt} + \Delta lnA^{V}_{jt}, \qquad (2)$$

which is consistent with this procedure. It has been established that in this procedure average shares between two time periods *t* and *t-1* should be used, according to formula $\bar{\alpha}_{jt} = (\alpha_{jt} + \alpha_{j(t-1)})/2$ and similarly for $\overline{\beta_{jt}}$ (however, β_{jt} is usually calculated residually (i.e. by subtraction of $\bar{\alpha}_{jt}$ from unity) in a similar way as for the Solow formulae). By definition, these shares are shares in the gross value added (GVA) here, and it is the growth rate of GVA (V_{jt}) that is present on the left-hand side of formula (2).

In addition to the above, the important change in the KLEMS growth accounting decomposition compared to the Solow decomposition is the introduction of different definitions for the factors labor and capital. In the KLEMS growth accounting, instead of using the resources (stocks) of these factors in formula (2), the quantities of services of these factors are used. The magnitudes of the factors' services are calculated by means of the Törnqvist quantity index aggregation procedure, in which factor growths are weighted by their relative shares in total income before being added up to the total contribution of the factor (at the given aggregation, including industry aggregations *j*). Therefore, in the KLEMS growth accounts, residual productivity is referred to as *multifactor productivity* (MFP), which is a variant of TFP. Thanks to its translog shape formula (2) is strictly conformable with the original Cobb-Douglas production function[15].

[14] In the original work of Solow (1957), the symbol α refers to the capital factor and the symbol β refers to the labor factor.

[15] However, in the instance when growths are high (much over 10%) the logarithm values become more discrepant with the classic relative growths from formula (1).

Formula (2) can be developed by introducing an additional variable, related to the intermediate inputs, to the original production function. In the theory developed after Solow, it was established that only the decomposition of gross output growth rate (with an additional factor-alike contribution in the form of intermediate inputs' contribution to gross output growth) allows to establish technological or organizational progress disembodied in labor or capital. This gross-output-based MFP contribution is different than the value-added-based MFP contribution, but in an ideal situation they should be related with each other by the ratio between the gross output and GVA. Otherwise, formula (2) allows to establish the contribution to growth of technological or organizational progress disembodied in labor or capital only approximately – it can be inconsistent (i.e. not related to a known ratio as mentioned above) because of the phenomenon of substitution between the production factors (labor and capital) and the intermediate inputs[16]. That is why the contribution of the *A* variable in (2) is presently rather considered as the industry capacity to capture the value. *Rather than technical change itself, the value-added based measure reflects an industry's capacity to translate technical change into income and into a contribution to final demand* (OECD, 2001, p. 27–28).

But the use of gross output growth rate decomposition is associated with data issues. Data insufficiency means that for most countries, for which the KLEMS growth accounting is performed, only the GVA growth rate decomposition according to formula (2) is being done[17]. Fortunately, it remains the central backbone of KLEMS growth accounting, providing the most essential information about the economy. Therefore, despite its limitations, it remains the basis for most analyses based on the method of decomposition in the framework of this accounting. Performing the GVA growth rate decomposition as in (2) instead of gross output growth rate decomposition facilitates also international comparisons since the issue of huge differences in the vertical integration of firms between the countries related with intermediate inputs is lifted. This issue is also important in the case discussed in the present paper, as those differences in vertical integration are also present between the different provinces of a given country. Therefore, for the present study oriented towards regional comparisons the choice of the GVA decomposition in the framework of the KLEMS growth accounting seems to be well justified.

The difference between the contribution of labor factor services and the contribution of its resource (hours worked) is, according to the KLEMS methodology,

[16] This discrepancy is observable but not that extremely large as to totally refute the simpler GVA growth rate decomposition.

[17] In the last EU KLEMS 2021, 2023 and 2025 releases (https://www.rug.nl/ggdc/productivity/eu-klems/) despite the use of extended decomposition framework with intangibles there is no mention about gross output growth decomposition. For the theoretical basics of these releases see: Bontadini et al (2023) and Corrado (2005 and 2009).

the quality of labor, otherwise-called labor composition. Although it is theoretically possible to divide the contribution of capital services similarly into the contribution of capital quality and the contribution of its resource (stock), in the adopted practice of the KLEMS growth accounts the capital services are divided differently, i.e. by capital type aggregates – into the contributions of ICT capital services and of non-ICT capital services.[18] Formula (2). Therefore, develops into the formula:

$$\Delta lnY_{jt} = \overline{\alpha_{jt}}\left(\Delta lnLC_{jt} + \Delta lnH_{jt}\right) + \overline{\beta_{jt}}\Delta ln\left(K_{ITjt} + K_{NITjt}\right) + \Delta lnA_{jt}, \quad (3)$$

where LC stands for labor quality (otherwise called *labor composition*), H – hours worked, K_{IT} – ICT capital services, K_{NIT} – non-ICT capital services. As can be seen, the contribution of capital services K_{jt} from formula (2) is divided differently than the contribution of labor services L_{jt}, as there is a logarithmic expression $\overline{\beta_{jt}}\Delta ln$ before the parentheses in addition to the weight of the factor, while for the labor factor there is only the weight α_{jt} before the parentheses. Hence, equation (3) can be further transformed into:

$$\Delta lnY_{jt} = \overline{\alpha_{jt}}\Delta lnLC_{jt} + \overline{\alpha_{jt}}\Delta lnH_{jt} + \overline{\beta_{ITjt}}\Delta lnK_{ITjt} + \overline{\beta_{NITjt}}\Delta lnK_{NITjt} + \Delta lnA_{jt}, \quad (4)$$

in which the expressions in brackets are separated. The difference in the way the capital factor has been split up is evident from the fact that the weights $\overline{\beta_{ITjt}}$ and $\overline{\beta_{NITjt}}$ differ, while the weights for the components of the labor factor are identical. Thus, it can be said that the capital factor has been separated into separate sub-factors, while the labor factor has only been separated into its distinct aspects (labor quality and labor stock).

The quality of labor is determined by the level of its hourly remuneration (it is assumed that markets function perfectly or at least correctly and that the hourly remuneration reflects the value of labor to the economy) – it is usually related to the level of education and experience (age), so according to this differentiation this quality of labor is calculated. The calculation of the quality of capital would also be related to its level of remuneration according to its types but in this case, we have the special situation that the high level of remuneration of capital is directly due to its short lifespan. If capital is short-life then the return on capital must be rapid for such capital to be of use in the economy, so the remuneration of short-life capital is high in relation to its stock value. The opposite is true for long-life capital – here, a low remuneration of capital relative to its stock value is possible, as there is a lengthy period of return on capital. Hence, the 'temptation' arises, or just the possibility, to stimulate an increase in the application of short-life capital at the expense of long-life capital in order to accelerate economic growth on an ad hoc basis. The American growth resurgence

[18] For the standard KLEMS growth accounting, i.e. for all releases before the last 2021, 2023 and 2025 releases, where (as above mentioned) intangible capital contributions have been extracted. The additional extraction of intangible capital lastly performed on EU KLEMS site is omitted here as not relevant for the issue at hand.

(Jorgenson, Ho and Stiroh, 2005)) is based on a similar phenomenon, and happened not necessarily on purpose but because of technological evolution (or revolution according to some).

In the KLEMS growth accounts, it was considered that there is no need to divide the service input of capital into the input of its quality and the input of its stock, since, in general, the capital with particularly high 'quality' in this sense is precisely ICT capital, while the remaining capital is capital with a rather more standard (albeit differentiated) 'quality'. While an increase in the quality of labor can and should be stimulated through an increase in the level of education, which is not controversial, in mathematical terms the same 'quality' of capital is controversial, as it may induce the above-mentioned behavior leading to negligence in the development of infrastructure – i.e. capital of low 'quality' in light of the definition of this term in this instance, for infrastructure capital is capital with a generally very long life and a very long period of return on capital[19] because of low yearly remuneration against its stock value.

Notwithstanding these considerations, the fact that the analysis of the contribution of capital services is carried out in the KLEMS productivity account somewhat differently from the analysis of the contribution of labor services is a fortunate circumstance for the issue at hand since in formula (4) it is sufficient to divide the contribution of non-ICT capital services K_{NITjt} to the GVA growth into the contribution of infrastructure capital services K_{INFjt} and the contribution of other capital services K_{Ojt}[20]:

$$\Delta lnY_{jt} = \overline{\alpha_{jt}}\Delta lnL_{jt} + \overline{\beta_{ITjt}}\Delta lnK_{ITjt} + \overline{\beta_{Ojt}}\Delta lnK_{Ojt} + \overline{\beta_{INFjt}}\Delta lnK_{INFjt} + \Delta lnA_{jt}. \quad (5)$$

In formula (5), the contributions of labor services have been combined back as in formula (2) because the focus of this article is not on separating out the quality of labor. The contribution of capital services K_{jt}, on the other hand, was finally separated into the contribution of ICT capital K_{ITjt} with a generally shorter lifespan, the contribution of other capital K_{Ojt} with a generally medium lifespan and the contribution of infrastructure capital K_{INFjt} with a generally longer lifespan. Now, we need to solve the problem of extracting the relevant data.

4. Infrastructure Capital Extraction compared to ICT Capital Extraction

The above positioning of infrastructure capital in the KLEMS growth accounts effectively solves the problem posed on the side of economic growth accounting

[19] Such views are sometimes advanced by researchers, see: Hirschman (1964).

[20] According to the author's knowledge the only attempt to extract infrastructure capital within growth accounting has been performed by Mas (2009). However, a large work on infrastructural data (which could end by a possible growth accounting exercise) has been performed also by Fraumeni (e.g. 1999, 2007, 2009 and 2022).

methodology, as only general estimates are sought here. However, the issue of obtaining appropriate data for equation (5) remains a problem. By analogy, we will first refer to the separation of ICT capital in the KLEMS growth accounts for the Polish economy, which has already been implemented by Statistics Poland.

The statistical data in the Polish National Accounts distinguish 6 capital types: *dwellings, other structures and buildings, transport equipment, other machinery and equipment, cultivated assets, intangible assets.* However, three ICT types of capital are not distinguished. The *computer hardware* and *telecommunications equipment* types are not extracted from the *other machinery and equipment* type, and the *computer software* type is not extracted from the *intangible assets* type. These three types of capital not extracted under Polish conditions are aggregated in the internationally practiced (e.g. on the EU KLEMS website) standard KLEMS growth accounts into one super-category of ICT capital, and the remaining types of capital into one super-category of non-ICT capital using the Törnqvist quantity index.[21] From this also follows the methodologically justified necessity to separate ICT capital in order to be consistent with the way accounts are performed for countries present on the EU KLEMS platform[22], to which the Polish KLEMS growth accounting results are most often compared (regardless of the fact that the importance of ICT capital for the Polish and many non-Polish economies is small).

Referring to the infrastructure capital type, it can be noted that it is not separated from the *other structures and buildings* wider capital type already used in the KLEMS accounting. Theoretically, an attempt could therefore be made to extract infrastructure capital in a manner analogous to the already performed extraction of ICT capital. At the same time, as hard infrastructure constitutes the 'lion's share' of the *other structures and buildings* type, it can be used as a first approximation in the analyses without separating 'pure' hard infrastructure capital from it (i.e. without separating other non-residential structures contained in it, such as offices, warehouses and factory buildings), which in turn can enable the accounts to be performed outright. This simplification turn is not possible for the ICT capital, which is not the 'lion's share' of the categories from which it has been extracted, but, as we will show further, at this stage it is not only possible but necessary for the infrastructure capital.

The adopted operation in the Polish KLEMS accounts was therefore to separate these three types of ICT capital before aggregating them into a common ICT capital category using the Törnqvist quantity index. This was done on the basis of the Supply and Use Tables (SUT), i.e. on the basis of the figures in the investment outlays column

[21] With the exception of 2021, 2023 and 2025 EU KLEMS releases mentioned above, where additionally intangible capital has been extracted.

[22] The lastly performed extraction of intangible capital contribution on EU KLEMS site is planned to be implemented in Statistics Poland's KLEMS accounts in the next round of data release.

for each of the above-mentioned three types of ICT capital (in the SUT Tables up to 2007 in the NACE 1.1 system these are groups 296, 316 and 430, while in the NACE Rev. 2 system these are groups 250, 252 and 489 respectively) – see further details of this already applied technique described in Kotlewski (2021).

A similar approach could be followed for the extraction of the infrastructure capital type, selecting the relevant items in the investment outlays column, and using the appropriate structures to distribute their values. In the NACE 1.1 system, these would be groups: 384, 385, 387 and 389 for investment outlays and the structures of NACE 1.1 divisions numbered 38, 39, 40 and 42. In the NACE 2 system, on the other hand, these would be groups 444-450, i.e. the entire division numbered 42 for investment outlays and the structures of NACE 2 divisions numbered 49-53. In doing so, it should be noted that the representation of infrastructure capital is more accurate in the NACE Rev. 2 system.

However, the effectiveness of this method may not be much better than the simplifying assumption mentioned above, based on the observation that the infrastructure capital type makes up the 'lion's share' of the *other structures and buildings* capital type, not least because this would be (especially for pre-2008 data in the NACE 1.1 system) a 'narrowly understood' infrastructure capital (i.e. without some components of it). Thus, an undertaking that was necessary in the case of ICT capital (thus 'narrowly understood' including only the most essential components such as computers, software, etc. without the rest of the equipment, including peripheral equipment), because otherwise it could not be extracted at all, may be not so necessary in the case of infrastructure capital. Furthermore, the following problem, also explained here by comparison with the performed ICT capital extraction by Statistics Poland, is presently unsolvable without going outside the SNA system.

When separating out ICT capital, the assumption was made that, since this capital has a short life in general, there is no need, given their very low value, to separate out the older parts of this capital from the existing broader aggregates. In order to determine the current state of fixed assets in this respect, it was considered unnecessary to consider the initial ICT capital stock from just few years before 2005 (the starting year of Polish KLEMS accounts). The total value of this initial ICT capital stock, due to its high rates of depreciation, becomes much less than 10% of its initial value after just a few years – the addition of ongoing investments and depreciation of ICT capital means that after just a few years the resulting total includes almost all ICT capital stock (albeit narrowly defined ICT capital, as only this can be effectively separated based on the SUT tables).

Unfortunately, this operation cannot be effectively conducted for infrastructural capital, as it has a very long life. When determining the initial fixed capital stock for this capital as negligible (in a similar way as for the ICT capital), the time horizon for this

operation would have to be extended many times – meanwhile, SUT tables for the period before 2000 are not drawn up for the Polish economy (and no initiatives to draw them up are advanced, because of unsurmountable difficulties). Thus, based on data from the SUT tables, the perpetual inventory method (PIM) cannot be effectively performed under Polish conditions for infrastructure capital. The other two methods of estimating the value of capital, i.e. observing the value of market transactions or the capital insurance market, also cannot be used effectively in this case, as infrastructure capital is predominantly owned by the state and is not normally traded on the market.

In this situation, it remains to check approximately how large the above-mentioned 'lion's share' can be attributed to infrastructure sensu stricto within the above-mentioned category *other structures and buildings*. This can be roughly estimated by observing the SUT tables, but only in the NACE 2 system, i.e. from 2008 onwards. It turns out that this infrastructure capital sensu stricto is rather by far the predominant part of this category. Thus, if we devise for the needs of the present analysis the category of a 'broadly understood infrastructure capital' then the above-mentioned *other structures and buildings* capital type can be a good approximation of it. When only increments are considered in decomposition formulae, then this approach becomes even more sensible. Therefore, in the following analysis, this capital type data are used as data source on infrastructure capital.

Before continuing, however, we need to make one important remark. It is true that for the Polish economy there are no SUT tables from before the year 2000 and no reliable data on capital stocks from before 2000 (not to mention the requirement for these data to be conformable with SNA 2008 or ESA 2010 systems). But there are countries where this kind of data collection and processing have started earlier[23]. In such a case the method suggested above (and applied in the Polish KLEMS growth accounts only for the ICT capital extraction) could become an effective tool for infrastructure capital extraction. KLEMS sites' data for some countries begin from the 1970s, awaiting also to be used for this purpose.

5. Empirical Findings

In order to limit the issue at hand to its essence we will transform equation (5) into:

$$\Delta lnY_{jt} = \overline{\alpha_{INFJt}} \Delta lnK_{INFjt} + \sum W_{jt} \tag{6}$$

[23] Direct data on capital stock, using the perpetual inventory method (PIM) are more available for some countries. This is the case for instance for the U.S. and they have been largely processed by Fraumeni (1999, 2007, 2009 and 2022). With quite rich, although disparate, array of data for this country, it is possible to choose narrow infrastructure components to be studied more closely as in the case of Fraumeni's works cited here, where strong evidence is provided for that *more disaggregated approach*. It is plausible to advance that a growth decomposition exercise could possibly be performed on those data.

where all contributions from equation (5) other than infrastructure capital (broadly understood as above explained) contribution have been added up to $\sum W_{jt}$. Thus, we can perform the calculations and present the data as in Table 1 – firstly, for the aggregate Polish economy (code "00") and later for the 16 Polish voivodships (codes "02" – "32").

Table 1. Decomposition of GVA growth into infrastructure contribution and other contributions for total Polish economy and by Polish voivodship

Specification		Years															
		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Code	Decomposition	in % for GVA and in pp for contributions															
00	GVA growth	3.32	5.93	6.87	3.88	3.07	3.71	4.62	1.37	1.19	3.31	4.00	3.06	4.57	5.17	4.51	-2.61
	Other contributions	3.00	5.95	5.32	3.46	2.16	1.93	2.64	0.41	-0.03	1.82	3.33	2.05	3.90	4.37	3.55	-3.47
	Infrastructure contribution	0.32	-0.02	1.55	0.42	0.91	1.77	1.98	0.95	1.22	1.49	0.66	1.02	0.68	0.79	0.96	0.86
02	GVA growth	5.38	9.59	8.90	3.08	3.97	7.31	5.33	1.26	0.39	3.18	3.94	2.64	4.15	4.38	4.41	-3.26
	Other contributions	4.90	9.68	7.13	2.54	3.19	5.51	3.53	-0.02	-1.08	1.76	3.44	1.41	3.74	4.11	3.37	-4.51
	Infrastructure contribution	0.49	-0.09	1.77	0.54	0.78	1.80	1.79	1.28	1.47	1.41	0.50	1.23	0.41	0.27	1.04	1.25
04	GVA growth	1.78	6.30	6.25	4.25	1.61	3.25	2.99	0.61	1.71	2.66	3.67	2.72	3.69	5.35	2.24	-2.39
	Other contributions	2.27	6.19	5.33	3.97	0.67	1.75	0.70	0.30	0.36	1.53	2.30	2.36	3.02	4.61	1.91	-3.06
	Infrastructure contribution	-0.50	0.11	0.92	0.27	0.95	1.50	2.30	0.32	1.35	1.13	1.37	0.36	0.68	0.75	0.33	0.67
06	GVA growth	1.79	4.06	7.12	5.23	0.23	3.28	4.26	2.02	1.84	1.98	1.90	2.85	3.49	2.40	4.54	-1.72
	Other contributions	1.73	4.11	6.75	5.17	0.13	2.52	3.44	1.04	1.25	0.92	1.37	2.44	3.16	2.29	3.93	-2.12
	Infrastructure contribution	0.06	-0.05	0.37	0.06	0.09	0.77	0.82	0.98	0.59	1.06	0.53	0.41	0.33	0.11	0.61	0.40
08	GVA growth	5.56	5.54	5.91	1.14	1.53	3.37	2.17	1.56	1.78	4.40	2.72	2.97	2.30	4.37	2.98	-3.07
	Other contributions	4.90	5.30	3.70	1.57	0.23	1.70	1.03	-1.26	-1.76	3.63	1.51	1.75	1.30	3.53	2.56	-3.17
	Infrastructure contribution	0.65	0.23	2.21	-0.42	1.30	1.67	1.14	2.82	3.54	0.77	1.21	1.22	0.99	0.84	0.42	0.10
10	GVA growth	3.39	5.47	6.40	4.64	1.42	4.34	4.47	1.34	0.85	3.35	2.90	2.66	3.83	4.99	5.21	-2.58
	Other contributions	2.87	5.42	5.74	4.26	0.79	4.28	1.59	0.30	-0.28	0.85	2.52	1.80	3.20	4.66	4.45	-2.99
	Infrastructure contribution	0.53	0.05	0.65	0.38	0.63	0.07	2.87	1.04	1.13	2.50	0.38	0.86	0.63	0.33	0.76	0.41
12	GVA growth	3.45	8.17	5.75	4.61	2.58	2.95	6.37	1.26	1.49	4.13	5.49	3.97	6.07	5.96	4.19	-2.47
	Other contributions	3.12	8.30	4.72	4.52	1.55	1.56	4.88	0.32	0.37	3.38	4.67	2.92	5.78	5.60	3.75	-2.84
	Infrastructure contribution	0.33	-0.14	1.04	0.09	1.03	1.38	1.49	0.94	1.12	0.75	0.83	1.05	0.29	0.36	0.44	0.37
14	GVA growth	5.34	6.55	7.58	2.71	4.82	5.05	5.01	2.30	2.19	3.72	4.51	3.79	5.58	6.22	6.09	-2.15
	Other contributions	5.03	6.84	5.90	1.93	3.73	2.95	2.82	1.75	0.47	2.07	4.08	2.05	4.47	5.38	4.97	-3.22
	Infrastructure contribution	0.32	-0.29	1.67	0.78	1.09	2.11	2.19	0.55	1.72	1.65	0.44	1.74	1.11	0.84	1.12	1.08
16	GVA growth	1.18	3.30	9.28	5.81	-1.72	1.58	4.40	-0.42	-0.06	3.72	2.02	1.02	4.31	4.88	3.61	-2.39
	Other contributions	0.60	3.49	7.50	7.01	-2.29	-0.69	4.40	-0.31	-0.95	3.03	1.65	0.68	3.55	4.39	0.28	-2.75
	Infrastructure contribution	0.58	-0.19	1.78	-1.20	0.57	2.27	0.00	-0.11	0.89	0.68	0.37	0.34	0.76	0.50	3.33	0.36

Table 1. Decomposition of GVA growth into infrastructure contribution and other contributions for total Polish economy and by Polish voivodship (cont.)

Specification		Years															
		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Code	Decomposition	in % for GVA and in pp for contributions															
18	GVA growth	2.68	5.48	5.83	6.12	2.17	3.04	5.52	0.90	2.56	3.14	3.49	2.33	3.99	6.13	4.50	-2.82
	Other contributions	2.57	5.33	4.66	5.88	1.55	1.82	4.58	0.83	1.36	2.20	3.12	1.73	3.67	5.62	4.28	-3.44
	Infrastructure contribution	0.11	0.14	1.17	0.24	0.62	1.23	0.94	0.07	1.20	0.94	0.37	0.60	0.31	0.52	0.22	0.62
20	GVA growth	3.48	3.73	7.73	2.34	4.75	2.87	3.67	-0.98	2.33	3.05	1.72	1.88	4.66	4.43	4.55	-1.53
	Other contributions	3.37	3.65	7.19	2.14	4.42	1.95	2.17	-1.27	1.06	1.60	0.88	1.53	4.31	3.60	3.88	-2.23
	Infrastructure contribution	0.11	0.07	0.54	0.19	0.32	0.92	1.50	0.29	1.27	1.45	0.84	0.36	0.35	0.83	0.67	0.70
22	GVA growth	4.26	6.37	7.25	1.27	6.45	2.74	5.36	3.47	0.06	2.41	4.96	3.97	5.15	6.19	5.29	-2.78
	Other contributions	4.11	6.18	5.65	0.94	5.64	0.95	3.62	3.34	-2.13	0.72	4.33	2.95	4.59	5.25	4.87	-3.89
	Infrastructure contribution	0.15	0.19	1.60	0.33	0.80	1.79	1.74	0.13	2.19	1.69	0.63	1.02	0.56	0.95	0.42	1.12
24	GVA growth	-0.36	3.57	6.31	4.10	1.13	3.03	4.14	0.03	-0.36	2.51	4.07	2.65	3.79	4.79	2.86	-3.76
	Other contributions	-0.49	3.40	4.88	3.80	0.36	1.21	2.23	-0.11	-2.60	0.81	3.41	1.51	3.12	3.52	2.28	-5.23
	Infrastructure contribution	0.13	0.17	1.43	0.30	0.77	1.82	1.91	0.14	2.24	1.70	0.66	1.14	0.68	1.27	0.58	1.46
26	GVA growth	-0.87	7.52	7.91	7.31	-0.72	1.78	3.06	-0.67	-2.06	3.26	1.83	1.42	3.78	5.51	2.33	-2.03
	Other contributions	-0.96	7.41	6.95	7.11	-1.26	0.45	1.70	-0.76	-3.74	1.91	1.31	0.55	3.27	4.54	1.87	-3.20
	Infrastructure contribution	0.09	0.11	0.96	0.21	0.54	1.33	1.36	0.10	1.69	1.35	0.52	0.86	0.50	0.97	0.46	1.17
28	GVA growth	3.12	5.03	5.27	4.19	3.71	2.96	3.37	0.47	0.71	3.44	2.59	2.62	2.18	3.18	3.18	-2.03
	Other contributions	3.00	4.88	4.04	3.93	3.03	1.39	1.82	0.35	-1.36	1.80	1.97	1.56	1.56	2.05	2.66	-3.37
	Infrastructure contribution	0.12	0.15	1.24	0.26	0.67	1.57	1.55	0.12	2.07	1.64	0.62	1.06	0.63	1.13	0.52	1.34
30	GVA growth	4.48	5.31	6.58	4.92	5.76	1.75	4.85	2.03	2.58	3.78	5.10	3.59	5.08	4.43	5.07	-2.64
	Other contributions	4.33	5.12	5.05	4.61	4.98	-0.05	3.08	1.91	0.47	2.12	4.48	2.58	4.51	3.42	4.63	-3.79
	Infrastructure contribution	0.15	0.19	1.53	0.31	0.78	1.80	1.77	0.13	2.11	1.66	0.62	1.01	0.57	1.01	0.44	1.14
32	GVA growth	3.54	4.92	4.48	4.74	0.72	2.68	3.05	1.24	0.07	3.89	3.81	1.62	4.16	4.43	3.57	-2.45
	Other contributions	3.40	4.74	2.99	4.43	-0.07	0.87	1.22	1.10	-2.21	2.05	3.09	0.40	3.47	3.22	3.00	-3.92
	Infrastructure contribution	0.14	0.18	1.48	0.31	0.79	1.81	1.83	0.14	2.28	1.84	0.72	1.22	0.69	1.21	0.57	1.47

Note: Data in NACE Rev. 2 classification. Code numbers in the left-hand column are code numbers used in the Polish statistics: 00 – for the total Polish economy aggregate, 02-32 – for the 16 Polish voivodships. The names of the voivodships are given in figure 2.
Source: own work based on Statistics Poland data.

As can be seen in Figure 1, concerning the total Polish economy, in the year 2006 there was almost no contribution of infrastructure to economic growth. This can be explained by the fact that economic actors were waiting for the EU funds to be approved, as Poland accessed the European Union (EU) in May 2004. Following that, the

role of infrastructure in economic growth became greater in 2007, followed by a slump in connection with the global financial crisis of 2007-2009. The role of infrastructure was even more important in the years preceding the Euro 2012 soccer event. For the purposes of this sporting event, a number of highways and other roads were built and some other elements of infrastructure (e.g. railways) were modernized – 2010-2011 was in fact the period of the most intensive development of infrastructure in the entire adopted period of analysis (2005-2020). This investment acceleration was performed in a country very needy in this respect, and the Euro 2012 sporting event provided a good socio-political momentum for it. However, infrastructure captured in this way also includes facilities associated with infrastructure, such as hotels and others.

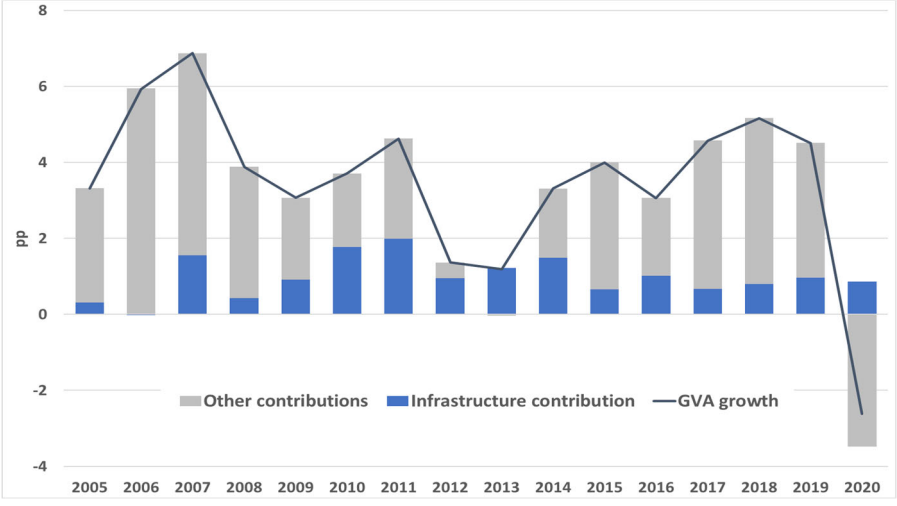

Figure 1. Decomposition of GVA growth into infrastructure contribution and other contributions for the total Polish economy.

Source: as for Table 1.

After 2012, the contribution of infrastructure to economic growth becomes smaller and fluctuates in a two or three-year cycle, which may be related to the two or three-year construction cycles (planning and preparation and then construction). In general, in the period after the above-mentioned soccer event the role of infrastructure in economic growth shows a decreasing trend. A parallel, unpublished yet, research conducted by the author of this paper indicates that this is probably due to the slower price increase in investment goods in Poland during that period, compared to consumer goods. It can be also seen that during the Covid-19 related crisis infrastructural investments were rather continued, as only a slight drop against the previous year can be observed, which is conformable to observation. This picture of the data seems confirming the validity of the method for the scope of the analysis.

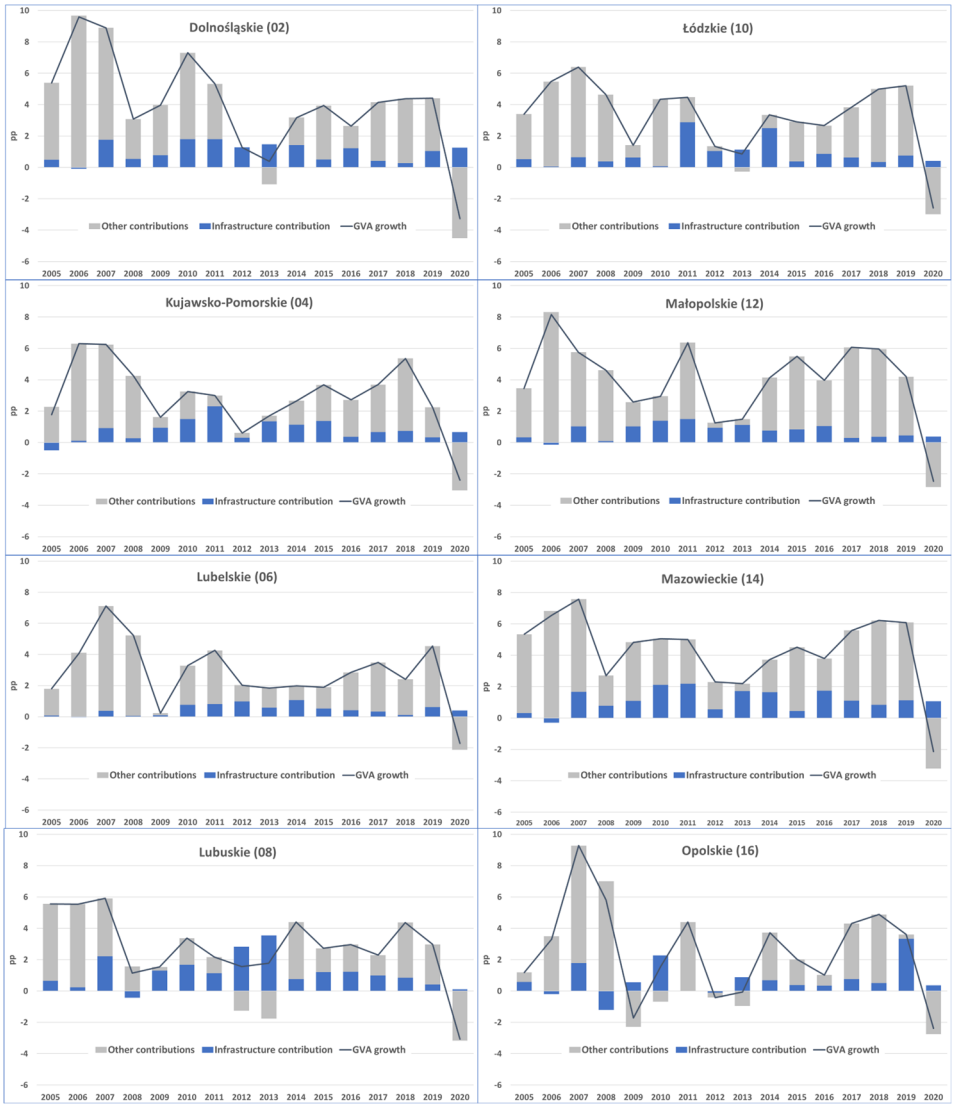

Figure 2. Decomposition of GVA growth into infrastructure contribution and other contributions for the 16 Polish voivodships.

Note: Data in NACE Rev. 2 classification. Code numbers by the names of voivodships are code numbers used in the Polish statistics: 00 – for the total Polish economy aggregate, 02-32 – for the 16 Polish voivodships.

Source: as for Table 1.

Figure 2. (cont.) Decomposition of GVA growth into infrastructure contribution and other contributions for the 16 Polish voivodships.

Note: Data in NACE Rev. 2 classification. Code numbers by the names of voivodships are code numbers used in the Polish statistics: 00 – for the total Polish economy aggregate, 02-32 – for the 16 Polish voivodships.

Source: as for Table 1.

However, one problem remains, namely that the presented data are not at the industry level but only at the aggregate level. This does not mean that the computations have not been done at the industry level, but that their validity is debatable, and therefore data at the industry level have not been included in the analysis. The question arises as to how to attribute shares in infrastructure components such as highways, railroads, etc., to individual industries distinguished in the KLEMS framework. Do industries participate equally, i.e. proportionally, in infrastructure services? Should individual consumers be treated differently from businesses? Which part of the tax revenue should be considered as infrastructure capital remuneration in growth accounting, as infrastructure fees are rather not directly charged on users (with some exceptions), etc.

However, despite difficulties in conducting the KLEMS type decomposition exercise by industry, as far as infrastructure contribution is considered, it remains nevertheless possible to develop it spatially. Therefore, the analysis based on Figure 1 can be significantly deepened if we refer to Figure 2, where the same kind of data from Table 1 were presented for the 16 Polish voivodships. It can be observed that the voivodships more involved in the above-mentioned 2012 soccer event and also those more centrally located, contrary to the more peripheral ones, have greater infrastructure contributions, which is conformable to expectations and is once again confirming the validity of the adopted method. It is well known that government's investments (and government supported or promoted ones) translate into SNA data reported in Statistics Poland. These data enter the KLEMS growth accounting and impact the contributions to economic growth with different effectiveness (some of the investment may not translate effectively into economic growth). They can be nevertheless traced qualitatively.

Three new stadiums were built in Gdańsk (Pomorskie voivodship), Wrocław (Dolnośląskie voivodship) and Warsaw (Mazowieckie voivodship). One stadium was utterly modernized in Poznań (Wielkopolskie voivodship). Around these soccer 2012 locations airports and train stations and lines were modernized, and more road connections established (more on this in Sobiera, 2012). Hotels were also built. Zachodniopomorskie (with Szczecin) and Lubuskie voivodships are important entry hubs to Poland from Western Europe. Kujawsko-Pomorskie and Łodzkie voivodships are centrally located, Śląskie (the most industrial voivodship) and Małopolskie (with Kraków, a renown touristic destination) are the remaining hubs that also participated in the infrastructure development momentum, as this event was also considered to be an occasion to promote Poland. We can observe that in many cases infrastructure investments were almost stopped dead in 2012 itself – this is the case for 11 out of 16 voivodships. The intention was not to annoy soccer participants (road and other works were temporarily restrained) and to deviate resources to withstand the additional soccer

event expenditures. Other differences from Figure 1 at the voivodship level are related to local circumstances and developmental needs.

A profound study of government and provincial policy documents could produce an ample book-size report but, nevertheless, it can be asserted that a sensible initial assessment of the role of infrastructure in economic growth was achieved – it is important but should not be overstated at the same time.

6. Conclusion

The main issue of the paper was to demonstrate the viability of a growth accounting methodology, the KLEMS growth accounting in particular, to study the role of infrastructure in economic growth, and an empirical case study for the Polish economy has been conducted for this purpose.

The paper demonstrates that there are strong theoretical premises for the use of this accounting methodology in this context, including the work of Jorgenson, Ho and Stiroh (2005). The specific interrelation between demand-side and supply-side impacts of infrastructure and its very long life-span lead to different attitudes of policy makers in regard to infrastructure investments – one is to forcibly invest in it in order to stimulate the economy on the demand side, the other one is to refrain from "excessive" investment in infrastructure in order to avoid the negative crowding out effect of those investment on business activity. A turn towards ICT capital investment in the U.S. during the 90s, as explained by Jorgenson, Ho and Stiroh (2005), was the reason for the U.S. economic growth resurgence in those years, and this is another justification to see this issue through the lenses of the KLEMS growth accounts.

As shown in the paper, it is possible to slightly alter the KLEMS formulae in order to distinguish the contribution of infrastructure to economic growth. There is, however, the issue of extracting appropriate data. This can be done thanks to the use of structures from the supply and use tables (SUT) but only for countries with very long time series of data conformable with the system of national accounts (SNA). However, despite the fact that this kind of data is not available for the Polish economy, it is still possible to make an acceptable assessment of the role of infrastructure in economic growth for Poland thanks to adopting the notion of a 'broadly understood infrastructure capital'. The obtained results do sensibly explain the impact of infrastructure on economic growth in the years 2005-2020, which is the period for which the KLEMS growth accounting exercise for the Polish economy has been conducted.

The most important limitation of the study is at the sectoral level of aggregation. Infrastructure, contrary to private capital owned by firms, is owned in Poland (as in many other countries) by the state, and therefore it is difficult to attribute it to industries or sectors. The solution to this problem could be to conduct a growth accounting

exercise at the firm level, using econometric methods that would allow to attribute infrastructure shares to firms and therefore also to industries. But this still remains a matter of future research. Fortunately, the lack of industry results does not affect the aggregate results. The second limitation is the measurement of infrastructure types. It would be interesting to analyze the contribution of different types of infrastructure – not treated *en bloc* – to economic growth. Innovative research is also required in this area.

Further research should extend the analysis in two directions. Firstly, it should compare the Polish infrastructure contributions with other countries using the EU KLEMS data, although the availability of data limits such comparisons, as only limited result data are available for other countries. Secondly, it should combine the KLEMS results with firm-level data in order to identify differences between groups of firms and, consequently, industries.

References

Ackerberg, D., Caves, K. and Frazer, G., (2015). Identification Properties of Recent Production Function Estimators. *Econometrica,* 83(6), pp. 2411–2451.

Ansar, A. *et al.*, (2016). Does infrastructure investment lead to economic growth or economic fragility? Evidence from China. *Oxford Review of Economic Policy*, 32(3), pp. 360–390.

Aschauer, D. A., (1989a). Does Public Capital Crowd Out Private Capital? *Journal of Monetary Economics*, 24(2), pp. 171–188.

Aschauer, D. A., (1989b). Is Public Investment Productive? *Journal of Monetary Economics*, 23(2), pp. 177–200.

Aschauer, D. A., (1989c). Public Investment and Productivity Growth in the Group of Seven. *Economic Perspective*, 13(5), pp. 17–25.

Bontadini, F. *et al.*, (2023). *EUKLEMS & INTANProd: industry productivity accounts with intangibles – Sources of growth and productivity trends: methods and main measurement challenges,* ECFIN/2020/O.P./0001 – Provision of Industry level growth and productivity data with special focus on intangible assets – 2020/S 114-275561.

Barro R. J., Sala-i-Martin X., (2004). One-Sector Models of Endogenous Growth, in: *Economic Growth* (Second ed.), New York: McGraw-Hill, pp. 205–237.

Corrado, C., Hulten, C. and Sichel, D., (2005). Measuring Capital and Technology: An Expanded Framework in: Corrado C., Haltiwanger J. and Sichel D. (Eds.), *Measuring Capital in the New Economy*. University of Chicago Press, pp. 11–46.

Corrado, C., Hulten, C. and Sichel, D., (2009). Intangible Capital and U.S. Economic Growth. *Review of Income and Wealth*, 55(3), pp. 661–685.

Diewert, W. E., (2004). Basic Index Number Theory in: *Consumer price index manual: theory and practice. International Monetary Fund*, Chapter 15, pp. 263–288.

Diewert, W. E., (1976). Exact and Superlative Index Numbers. *Journal of Econometrics*, 4(2), pp. 115–145.

Diewert, W. E., (2005). Issues in the Measurement of Capital Services, Depreciation, Asset Price Changes and Interest Rates, in: Corrado C., Haltiwanger J. and Sichel D. (Eds.), *Measuring Capital in the New Economy*. University of Chicago Press, pp. 479–542.

Diewert, W. E., (1978). Superlative Index Numbers and Consistency in Aggregation. *Econometrica*, 46(4), pp. 883–900.

Diewert, W. E., (1992). The Measurement of Productivity. *Bulletin of Economic Research*, 44(3), pp. 163–198.

Du, X., Zhang, H. and Han, Y., (2022). How Does New Infrastructure Affect Economic Growth Quality? Empirical Evidence from China. *Sustainability*, 14, 3511.

Fraumeni, B. M., (2022). *How should we measure infrastructure? The case of highways and streets*, NBER Working Paper, 30045.

Fraumeni, B. M., (1999). *Productive Highway Capital Stock Measures,* under subcontract to Battelle Memorial Institute, Contract DTFH61-97-C-00010, BAT-98-006, Federal Highway Administration, U.S. Department of Transportation.

Fraumeni, B. M., (2007). *Productive Highway Capital and the Contribution of Highways to Growth in GDP,* Volume I and II, under subcontract to Battelle Memorial Institute, subcontract no. 208937 to work order BAT-03-21, Federal Highway Administration, U. S. Department of Transportation.

Fraumeni, B. M., (2009). *The Contribution of Highways to GDP Growth*. NBER Working Paper, 14736.

Hirschman, A. O., (1964). Stratégie du developpement économique. *Les éditions ouvrières*, pp. 105–112.

Jorgenson, D. W., Griliches, Z., (1967). The explanation of Productivity Change. *The Review of Economic Studies*, 34(3), pp. 249–283.

Jorgenson, D. W., (1963). Capital Theory and Investment Behavior. *The American Economic Review*, 53(2), pp. 247–259.

Jorgenson, D. W., (1989). Productivity and Economic Growth, in: Berndt R. E., Triplett E. J. (ed.), *Fifty years of economic measurement: the jubilee of the conference on research in income and wealth*. University of Chicago Press, pp. 19–118.

Jorgenson, D. W., Gollop F. M. and Fraumeni, B. M., (1987). *Productivity and US economic growth.* Harvard University Press.

Jorgenson, D. W., Ho M. S. and Stiroh, K. J., (2005). *Information technology and the American growth resurgence,* MIT Press.

Kotlewski, D., (2021). *KLEMS Productivity Accounting for the Polish Economy*, Statistics Poland.

Leontief, W., (1966). *Input-Output Economics*, Oxford University Press 1986.

Levinsohn, J., Petrin, A., (2003). Estimating Production Functions Using Inputs to Control for Unobservables.*Review of Economic Studies*, 70, pp. 317–341.

Lucas, R. E., (1988) On the mechanics of Economic Development. *Journal of Monetary Economics,* 22, pp. 3–42.

Mas, M., (2009). Infrastructures and New Technologies: as Sources of Spanish Economic Growth, in: *OECD. Productivity Measurement and Analysis*, OECD Publishing, pp. 357-378.

OMahony, M., Timmer, M., (2009). Output, Input and Productivity Measures at the Industry Level: The EU KLEMS Database. *The Economic Journal,* 119 (June), pp. F374-F403.

OECD, (2001). *Measuring productivity, OECD manual, Measurement of aggregate and industry-level productivity growth.*

Olley, G. S., Pakes, A., (1996). The Dynamics of Productivity in the Telecommunications Equipment Industry. *Econometrica,* 64(6), pp. 1263–1297.

Romer, P. M. (1994). The Origins of Endogenous Growth, *The Journal of Economic Perspectives,* 8(1), pp. 3–22.

Sobiera, A., (2012). *Infrastruktura sportowa i komunikacyjna na Euro 2012* (Sport and communication infrastructure for Euro 2012), Nowoczesne Budownictwo Inżynieryjne, pp. 82–89.

Solow, R. M., (1956). A Contribution to the Theory of Economic Growth. *The Quarterly Journal of Economics*, 70(1), pp. 65–94.

Solow, R. M., (1957). Technical Change and the Aggregate Production Function. *Review of Economics and Statistics*, 39(3), pp. 312–320.

Timmer, M. *et al.,* (2007). *EU KLEMS growth and productivity accounts. Version 1.0. part I. Methodology,* EU KLEMS Consortium.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 183–206, https://doi.org/10.59139/stattrans-2026-033
Received – 11.08.2024; accepted – 07.07.2026

Seasonal ARIMA modeling of inflation in Nigeria

Saheed A. Ajobo[1], **Oluwaseun A. Adesina**[2], **David I. Ohida**[3], **Kafayat A. Alobaloke**[4], **OlaOluwa S. Yaya**[5], **Mary M. Adepoju**[6]

Abstract

The success of monetary and fiscal policies in controlling inflation rates relies largely on the availability of accurate inflation forecasting models. Thus, updated models that require seasonality are required. This study applies the Seasonal Autoregressive Integrated Moving Average (SARIMA) process to model the consumer price indices in Nigeria, specifically headline, food and core inflation. The study analyses monthly inflation rate data for the period from January 2003 to December 2023, obtained from the Central Bank of Nigeria's website. The findings of the study identify the best-fitting SARIMA models for each inflation measure. The SARIMA (1,1,0)(2,0,1)12 model is selected as the most appropriate for headline inflation, while the SARIMA(2,1,2)(2,0,2)12 and SARIMA(1,0,1)(2,0,1)12 models are chosen for food inflation and core inflation, respectively. Caution is then required when applying SARIMA models in forecasting inflation, with emphasis on the importance of considering other relevant economic and socio-political factors that may influence inflation trajectories.

Key words: CPI inflation, Nigeria, SARIMA, food inflation.
JEL Classification Codes: C22.

1. Introduction

Inflation rate is the key metric used to measure price stability in an economy, and due to the severe microeconomic and macroeconomic consequences of rising inflation, countries have always sought to keep it low. In 2007, the Central Bank of Nigeria (CBN)

[1] Department of Statistics, University of Ibadan, Ibadan, Nigeria. E-mail address: sa.ajobo@ui.edu.ng..
[2] Department of Statistics, Ladoke Akintola University of Technology, Ogbomosho, Nigeria. E-mail address: oaadesina26@lautech.edu.ng.
[3] Department of Statistics, University of Ibadan, Ibadan, Nigeria.
[4] Department of Statistics, University of Ibadan, Ibadan, Nigeria. E-mail address: kalobaloke96@gmail.com.
[5] Department of Statistics, University of Ibadan, Ibadan, Nigeria. E-mail address: os.yaya@ui.edu.ng.
[6] Department of Statistics, Ladoke Akintola University of Technology, Ogbomosho, Nigeria. https://orcid.org/0009-0002-1445-9317.

Act, which made monetary and price stability the number one priority of the CBN, was introduced to fight rising inflation in the country. Over the past decades, in line with its core mandate, the CBN has sought to control and stabilize inflation in Nigeria through various monetary policies at its disposal. This goal is very crucial as the rate of inflation in the country has far-reaching and spreading effects on various sectors of the economy.

Nigeria has continued to experience double-digit headline, food, and core inflation in recent years, indicating that the fight against inflation has not been effective. There has been increased pressure on the country's inflation rates as a result of high levels of insecurity, high demands for foreign currency from its citizens due to the nation's higher imports than exports, the exodus of people leaving the country, and a decline in foreign investments. The country is experiencing a complicated web of economic misery as a result of high inflation rates, despite the government and monetary authorities' best efforts to combat it.

Despite the initial drop in prices, Nigeria experienced unpredictable inflation rates in the years that followed the implementation of SAP. The reasons for this inflationary volatility were identified as monetary policy fluctuations and extra-budgetary spending practices. Monetary policy measures such as exchange rate policies, money supply management, and interest rate adjustments were inconsistent, which led to economic uncertainty and volatility in inflation. Trade restrictions and fluctuating foreign exchange rates impeded growth and reduced living standards. The rate of inflation varied greatly, rising to more than 50% in 1988–1989 before momentarily declining.

Inflation forecasting is highly useful to central banks, as it plays a crucial role in guiding their monetary policy decisions. To preserve price stability, encourage economic growth, and successfully carry out their mandates, central banks can make better-informed and timely decisions by relying on accurate and up-to-date inflation forecasts (see Tumala et al., 2017; 2019).

Inflation forecasting is essential for anchoring inflation expectations in addition to its influence on monetary policy. These expectations, which are essentially how businesses and consumers view future inflation, have a significant impact on actual inflation. Businesses and consumers are more likely to boost prices and wages in anticipation of high inflation, which can lead to a self-fulfilling prophecy. On the other hand, expectations that are well-anchored and in line with the central bank's aim can promote an atmosphere of inflation that is more stable. Central banks can influence these expectations by communicating inflation forecasts in a transparent and consistent manner. This promotes a more predictable and controllable inflation trajectory by creating an environment in which consumers and companies base decisions about wages and pricing on the expected level of prices in the future.

Moreover, accurate inflation forecasts enable central banks to proactively address possible inflationary threats. Often, inflationary pressures intensify gradually before

showing up as significant price increases. Through accurate forecasting, central banks can detect these patterns and take preemptive measures to mitigate them. This could entail taking proactive steps like managing the money supply through open market operations or changing reserve requirements for banks, which impacts the amount of money they can lend. By taking proactive measures based on reliable inflation forecasts, central banks can prevent inflation from spiraling out of control and causing significant economic disruption.

Beyond central bank policy, accurate inflation forecasts also boost market confidence. By giving businesses and investors a better idea of what to expect from inflation in the future, they can make more confident decisions, which lowers the risk of unexpected spikes in inflation disrupting the financial system. Also, the release of inflation forecasts shows a central bank's dedication to price stability and transparency, which builds public trust and fosters accountability, allowing the public to hold the bank responsible for meeting its inflation targets.

In summary, inflation forecasting gives central banks an effective instrument for navigating the intricacies of the economy. It gives them the ability to anticipate future price changes, control inflation expectations, and take proactive measures to avert inflationary threats by giving them insights into future price movements. Ultimately, this promotes an economic climate that is predictable and steady, which is favorable for long-term prosperity and growth.

Although conventional forecasting techniques have been used by central banks for many years, they frequently fail to adequately represent the intricacies of the dynamics of inflation in developing nations such as Nigeria. These methodologies might not adequately account for seasonal fluctuations, a characteristic feature of inflation in many countries due to factors like agricultural cycles. Agricultural cycles often drive food production and prices, and hence, the monthly food prices are expected to show seasonality. This is where the Seasonal Autoregressive Integrated Moving Average (SARIMA) models emerge as a potentially powerful tool. These models incorporate the strengths of Autoregressive Integrated Moving Average (ARIMA) models by taking into account seasonality in the data. By incorporating these crucial components, SARIMA models have the potential to generate more accurate and nuanced forecasts of inflation, particularly in economies with seasonal variations. The ARIMAX model is entirely different from the ARIMA model since X component is the exogenous variable. This can be extended to the SARIMAX model, which allows for the seasonal component, but we are limited in terms of data gathering. The SARIMAX model requires that the additional variable X be an exogenous variable, which serves as a predictor in addition to the seasonal/non-seasonal autoregressive and moving average parameters in the SARIMA model. So, this is not the case in this study.

This study, therefore, will examine in more detail the dynamics of headline, food and core inflation in Nigeria using the SARIMA processes for inflation forecasting in Nigeria. It will investigate how this methodology can address the drawbacks of conventional approaches and potentially provide valuable insights for central banks, businesses, and households alike. The SARIMA model has been used among researchers around the globe for modeling and forecasting due to its usefulness in factoring the seasonality of data.

2. Literature Review

Inflation, defined as the persistent increase in the general price level of goods and services in an economy, has been a critical economic issue facing Nigeria and other countries around the globe. As highlighted by the World Bank, developing countries like Nigeria grapple with the adverse impacts of sustained high inflation, evident in the continuous rise of overall price levels. This poses major barriers to achieving macroeconomic stability and growth. Given these multifaceted challenges, economists and researchers have dedicated substantial efforts towards studying drivers of inflationary pressures and modeling inflation dynamics.

The relative ability of structural and monetary factors to explain the underlying causes of rising prices is a significant topic of discussion in the economics of inflation. The Monetarist school, which is most closely linked to Milton Friedman, places a strong emphasis on how interest rates and the money supply affect changes in prices. (Friedman, 1970), (Friedman, 1971) and (Laidler & Parkin, 1975). Structuralist theories, on the other hand, identify economic forces like fiscal balances, output gaps, cost-push shocks, taxes, import prices, and currency depreciations as the main drivers of inflation (Streeten, 1962), (Baumol, 1967), and (Maynard & Ryckeghem, 1976). This theoretical divergence has significant policy implications regarding the most suitable strategies for controlling inflation. To inform this discussion and guide policymakers, extensive empirical literature has evolved that attempts to model and forecast inflation patterns in different countries.

The fundamental causes of inflation in developing nations like Nigeria are still up for debate. In the Nigerian example, significant and sustained inflation notwithstanding tighter monetary policy has rendered this an open question with crucial stakes. To model Nigerian inflation during the past few decades, several studies have examined this topic, evaluating the importance of interest rates, money supply, government spending, production gaps, exchange rates, and other proposed drivers arising from the aforementioned theories. As the diversity of approaches continues expanding, further explaining Nigerian inflation dynamics through this theoretical lens can define effective policy responses. Sani, Abdullahi, & Adamu (2016) examined Nigeria's inflation

dynamics' drivers from 1981 to 2015. Their results are consistent with the classical quantity theory of money, indicating that monetary issues are the main cause of inflation in Nigeria. As a result, they propose that the implementation of strict monetary policies for a prolonged duration may be able to successfully control inflation in Nigeria.

Okwori and Abu (2017) examined how well monetary policy works in Nigeria to manage inflation. They did this by using time series data spanning from 1986 to 2015 and a vector error correction (VEC) model for analysis. While their findings confirm the statistically significant influence of monetary policy on inflation, they reveal weak and maybe insufficient impacts in reducing inflationary pressures. The authors contend that the existence of a sizable informal sector in the Nigerian economy is the cause of this low efficacy. This sector, which operates outside of the official financial system, is still mainly unresponsive to traditional monetary policy tools. Recognizing the substantial economic and social costs associated with inflation, Olubusoye and Oyaromade (2018) explored the intricate dynamics of inflationary processes in Nigeria. Using an error correction model framework, their analysis reveals that lagged CPI, anticipated inflation, petroleum prices, and real exchange rate exert the most significant influence on inflation fluctuations. Notably, the research highlights the limited impact of the level of output on inflation dynamics, as well as the unexpected negative coefficient of the lagged value of money supply, which was found to be statistically significant only at 10%. The study indicates that monetary authorities' attempts to stabilize domestic prices are significantly hampered by the volatility of crude oil prices around the world. Consequently, the authors posit that to effectively reduce inflation in Nigeria, solutions that go beyond conventional monetary policy tools and a broader examination of global and economic factors may be required. Olubusoye, Akanbi, and Akintande (2017) explored the application of Bayesian model averaging (BMA) to identify key determinants of inflation in Nigeria and generate inflation forecasts. The analysis reveals that the real interest rate is the strongest single predictor of inflation in Nigeria, with a posterior inclusion probability of 95.6%, and has an inverse relationship with inflation. Other papers that applied BMA in modeling Nigeria inflations are Tumala et al. (2017) and Tumala et al. (2019). The literature suggests that monetary and fiscal policy interact in a complicated way to influence inflation in Nigeria.

There is a rich body of literature examining inflation dynamics and predicting future trends across advanced, emerging, and developing economies worldwide. Nigeria is no different, with its high and volatile inflation making reliable forecasts critical but challenging. A range of statistical and econometric techniques have been employed over the past few decades in modeling Nigerian inflation. While accuracy remains constrained by model uncertainties, this steadily growing literature has provided vital insights for policymakers in Africa's largest economy as they aim to

achieve low and stable inflation rates. Oyewale, et al. (2019) compared the performance of artificial neural network (ANN) models, a non-parametric technique, and traditional ARIMA models for predicting inflation rates in the United States of America and Nigeria. The results show that the ANN models, including Standard Backpropagation (SBP), Scaled Conjugate Gradient (SCG), and Backpropagation (BPR), outperform the ARIMA models in forecasting inflation rates, as measured by five performance indices. This suggests that ANN-based models have the potential to provide forecasts that are more accurate and insightful than those produced by traditional ARIMA techniques.

Using the monthly inflation rate series that spanned January 2003 to October 2020, Adelakun, Abiola and Constance (2020), modelled and forecasted inflation in Nigeria and provided three years monthly forecast for the inflation rate in Nigeria. They examined 169 ARMA, 169 ARIMA, 1521 SARMA, and 1521 SARIMA models to determine which model would be most appropriate for modeling the inflation rate in Nigeria. Their findings indicate that out of the 3380 models examined, SARMA (3, 3) x (1, 2)12 is the best model for forecasting the monthly inflation rate in Nigeria. In a bid to deliver precise short-term predictions of headline, food, and core inflation in Nigeria, Doguwa and Alade (2013) proposed four short-term headline inflation forecasting models using seasonal auto-regressive integrated moving average (SARIMA) and seasonal auto-regressive integrated moving average with exogenous factors (SARIMAX) methodologies. These models underwent assessment through a pseudo-out-of-sample forecasting technique spanning from July 2011 to September 2013. The results demonstrated the greater performance of the streamlined SARIMAX model in predicting headline inflation over a period of one to twelve months. The model incorporates external elements including fuel prices, core and food indices, reserve money, and rainfall in certain agricultural regions. In addition, compared to SARIMAX models, SARIMA models showed better accuracy in predicting inflation for food and core.

A comparative examination of generalized auto-regressive conditional heteroskedasticity (GARCH)-type models was carried out by Fasanya and Adekoya (2017) to evaluate how well they captured the volatility of inflation in Nigeria between January 1995 and October 2016. The study assessed how well these dynamics are represented by symmetric (GARCH and GARCH in Mean (GARCH-M)) and asymmetric (exponential GARCH (EGARCH) and threshold GARCH (TGARCH)) models. The findings indicate that symmetric models are less capable of accurately representing inflation volatility than their asymmetric equivalents. Interestingly, the EGARCH model fits the data best. In a study that also aimed to model inflation in Nigeria, Shaibu and Osamwonyi (2020) developed and compared the performance of two time series models: a univariate ARIMA model and a multivariate autoregressive distributed-lag (ARDL) model. Based on predetermined criteria, the study of quarterly

data from 1988 to 2017 revealed that the ARDL(4, 2, 2, 1) and ARIMA(2, 1, 3) models were the most appropriate methods to represent inflation in Nigeria.

The effectiveness of different machine learning algorithms in predicting inflation in Nigeria was examined by Nakorji and Aminu (2022). Their comparative analysis assessed the performance of ridge regression and ANNs against least absolute shrinkage and selection operator (LASSO), elastic net, and partial least squares (PLS) regression models. The study finds that ANNs and ridge regression perform better than the other techniques at accurately predicting inflation in Nigeria. Uko and Nkoro (2012) examined the relative accuracy of ARIMA, vector auto-regression (VAR), and error correction model (ECM) models in forecasting inflation in Nigeria. The study compares the forecasting performances of the models and the simulated series' ability to track real data using annual data from 1970 to 2010. The results of the analysis show that the choice of the best model depends on the forecasting horizon. The VAR model performs better for short-term forecasts, the ECM model performs better for long-term projections, and the ARIMA model is good as a benchmark model.

The efficacy of univariate and multivariate time series models for predicting inflation in Nigeria is examined by Kelilume and Salami (2013). They make use of the VAR and ARIMA models. The study compares the forecasting accuracy of the two models and uses monthly CPI data from the National Bureau of Statistics for the years 2003–2012 to forecast future inflation rates. The result of the analysis reveals that while both models track actual inflation values to a certain extent, the VAR model exhibits a lower mean squared error, indicating its superior ability to closely approximate current inflation dynamics in the Nigerian context. Nyoni and Nathaniel (2018) modelled and predicted inflation in Nigeria using time series analysis. They assess the effectiveness of several models, such as the ARMA, ARIMA, and GARCH models, using data on inflation rates from 1960 to 2016. Based on predetermined selection criteria, the study determines that the ARMA(1, 0, 2) model is the best option, outperforming the ARIMA(1, 1, 1) and AR(3)-GARCH(1, 1) models.

Seasonal ARIMA (SARIMA) models have proven useful across diverse disciplines and fields for time series forecasting tasks involving seasonal fluctuations properties. When it comes to managing seasonal non-stationarities, SARIMA models offer more flexibility than normal ARIMA specifications since they integrate seasonal differencing components. This makes the SARIMA model a better forecasting method than the normal ARIMA model when the variable to be modelled has seasonal fluctuations properties. Kabbilawsh et al. (2020) used the SARIMA model in developing temperature forecasting models for the state of Kerala. The study utilized mean maximum and mean minimum monthly temperature data spanning 47 years from seven stations and identified SARIMA(2, 1, 1)(1, 1, 1)12 as the ideal forecasting model for eight out of the fourteen time-series datasets.

Another property of time series found in inflation is the persistence due to the unit root level in the series. Yaya and Shittu (2020) observed long memory in the Nigerian and US inflation rates. By including structural breaks, Gil-Alana et al. (2012) found long memory and breaks in Nigeria inflations. In the case of G7 and OECD groups, it was shown by Yaya et al. (2019), Yaya (2018) and Gil-Alana et al. (2016) that inflation rates could contain series of breaks which can influence the unit root decision. Thus, properties of the inflation series need to be checked thoroughly before further estimation since inflation data could contain nonlinearity and seasonality.

3. Data and Statistical Methods

3.1. The Data

This study makes use of monthly inflation data from January 2003 to December 2023 obtained from the Central Bank of Nigeria's website. The data consists of 249 observations. R software is the main statistical software used for analysis and estimation in this study.

3.2. Seasonal Autoregressive Integrated Moving Average Model

This study applied the Seasonal Autoregressive Integrated Moving Average (SARIMA) model in modeling Nigeria's Consumer Price Index. SARIMA is an extension of the ordinary ARIMA model to analyze time series data which contain seasonal and non-seasonal behavior. The SARIMA model accounts for the seasonal property in the time series and therefore has a forecasting advantage over the ordinary ARIMA model. A time series is said to be seasonal of order d if there exists a tendency for the series to exhibit periodic behavior at regular or almost regular time interval d. The time series $\{X_t\}$ is said to have a multiplicative (p, d, q) × (P, D, Q)s seasonal ARIMA model.

The Autoregressive process AR expresses a dependent variable as a function of past values of the dependent variable. An autoregressive process of order p [AR(p)] is given as:

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t,$$

where $\phi_1, \phi_{2,\ldots,}\phi_p$ are autoregressive parameters measuring the effect of $X_{t-1}, X_{t-2,} \ldots, X_{t-p}$ on X_t and ε_t is the error term at time t with mean zero and a constant variance. Using the backward shift operator, $B^p X_t = X_{t-p}$, the AR(p) process can be written as:

$$X_t = \phi_1 B X_t + \phi_2 B^2 X_t + \cdots + \phi_p B^p X_t + \varepsilon_{t,},$$
$$X_t - \left(\phi_1 B X_t + \phi_2 B^2 X_t + \cdots + \phi_p B^p X_t\right) = \varepsilon_t,$$
$$\left(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p\right) X_t = \varepsilon_t.$$

This implies

$$\Phi(B)X_t = \varepsilon_t.$$

The Moving Average (MA) process expresses a dependent variable as a function of the present and past residuals. The Moving Average process of order q [MA(q)] is defined as:

$$X_t = \varepsilon_t - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q},$$

where ε_t is a white noise process with mean zero and constant variance. Using the backward shift operator, the MA(q) process can be written as:

$$X_t = \varepsilon_t - \theta_1 B\varepsilon_t - \theta_2 B^2\varepsilon_t - \cdots - \theta_q B^q\varepsilon_t,$$
$$X_t = \left(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q\right)\varepsilon_t.$$

This implies

$$X_t = \Theta(B)\varepsilon_t.$$

If both AR(p) and MA(q) components are present in a time series process, there is an autoregressive moving average [ARMA(p, q)] process, satisfying,

$$X_t - \phi_1 BX_t - \phi_2 B^2X_t - \cdots - \phi_p B^pX_t = \varepsilon_t - \theta_1 B\varepsilon_t - \theta_2 B^2\varepsilon_t - \cdots - \theta_q B^q\varepsilon_t .$$

In a compact form, this

$$\Phi(B)X_t = \Theta(B)\varepsilon_t,$$

where $\Phi(B)$ and $\Theta(B)$ give the set of autoregressive and moving average parameters, and ε_t is the white noise process.

The integrated ARMA or ARIMA model is a broadening class of ARMA models which includes a differencing term d. The ARIMA model is a combination of autoregressive and moving average models. A process is said to be ARIMA (p, d, q) if

$$\nabla^d X_t = (1 - B)^d X_t.$$

This is generally written as:

$$\Phi(B)(1 - B)^d X_t = \Theta(B)\varepsilon_t,$$

where ε_t follows a white noise process, B is the lag operator, d is the order of integration that makes the data stationary, $\Phi(B)$ and $\Theta(B)$ are the autoregressive operator and the moving average operator, respectively.

The SARIMA model is denoted as ARIMA (p, d, q)x(P, D, Q)s. This can be written in the lag form as:

$$\Phi_{\mathrm{p}}(B^s)\phi_p(B)(1 - B)^d(1 - B^s)^D X_t = \theta_q(B)\Theta_{\mathrm{q}}(\mathrm{B}^{\mathrm{s}})\varepsilon_t,$$

where P, d, and q are the order of non-seasonal AR, differencing and MA respectively,

P, D, and Q are the order of the seasonal AR, differencing and MA respectively,

X_t represents the observable time series data at time t,

ε_t represents white noise or error (random shock) at period t,

B is the backward shift operator,

s represents the seasonal order,

$\theta_p(B)$ and (B) are the autoregressive and moving average polynomials, respectively, defined as:

$$\theta_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p ,$$
$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q,$$

and $\Phi_{\mathrm{p}}(B^s)$ and $\Theta_{\mathrm{q}}(\mathrm{B}^{\mathrm{s}})$ are the seasonal autoregressive and moving average polynomials, respectively, defined as:

$$\Phi_{\mathrm{p}}(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \phi_p B^{ps} ,$$
$$\Theta_{\mathrm{q}}(\mathrm{B}^{\mathrm{s}}) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots - \Theta_{\mathrm{q}} B^{qs} .$$

4. Findings

Table 1 gives a summary of the three consumer price indices. The summary statistics shed light on the behavior of inflation in Nigeria over the period from January 2003 to December 2023. It is evident from the statistics that there have been significant fluctuations in inflation rates, as seen by the large differences between the minimum and maximum values recorded for headline, food, and core inflation measures.

Table 1. Summary statistics of the Consumer Price Indices

Specification	Headline Inflation	Food Inflation	Core Inflation
Minimum	3.00	-3.70	-0.4
1st Quartile	9.40	9.90	8.785
Median	12.32	13.54	11.30
Mean	13.18	14.08	12.33
3rd Quartile	15.98	17.86	14.50
Maximum	28.92	38.50	41.20

Source: Computed by the author in R.

The headline inflation rate, which captures the overall increase in prices for goods and services, ranged from a low of 3% to a high of 28.92%, with a mean of 13.18%. This suggests that the general price level in the economy has experienced significant fluctuations, with periods of comparatively low and high inflation. Food inflation, which measures the change in prices for food items, exhibited even greater variability. The minimum food inflation rate was -3.7%, indicating a period of deflation or falling food prices. However, at its peak, food inflation reached a staggering 38.5%. The mean food inflation rate of 14.08% was higher than both headline and core inflation, emphasizing the significance of food prices in driving overall inflation dynamics in Nigeria. Core inflation, which excludes volatile food and energy prices, ranged from -0.4% to 41.2%, with a mean of 12.33%. This measure is often used to assess underlying inflationary pressures in the economy, and the observed variability suggests that core inflationary pressures have been significant and persistent. Furthermore, the distributions of all three inflation measures were positively skewed, with mean values

exceeding the respective medians. This implies that periods of high inflation were more frequent than periods of low inflation, which could have significant economic and social implications.

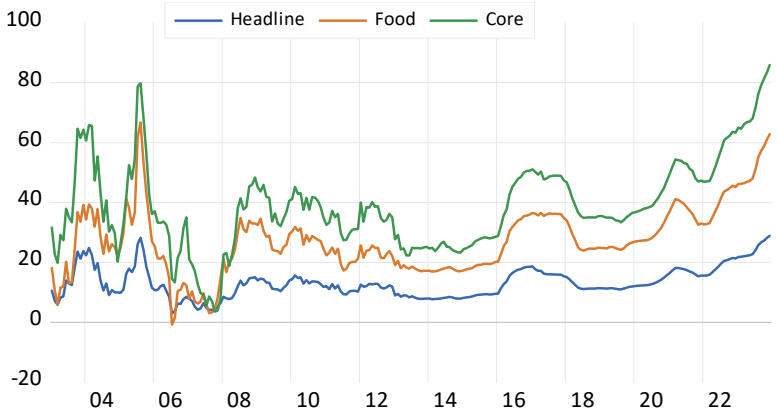

Figure 1. Time plot of headline inflation

Figure 1 displays the trend of headline, core and food inflations over an extended period from around 2003 to 2023. The inflation rate has shown considerable volatility over the given period, with notable peaks and troughs, as seen in the chart. This implies that the economy has experienced both deflationary or low inflation situations and times of high inflationary pressures. The inflation rate has experienced multiple separate surges, each of which can be attributed to a different economic issue such as disruptions in the supply chain, increased energy prices, or external shocks. The chart shows a considerable increase in inflation towards the end of the time period, reaching levels not seen since the previous spikes. This recent surge in inflation could be a significant concern for policymakers and could have implications for economic growth, consumer spending, and monetary policy decisions.

The volatility experienced by food inflation suggests that the food sector in the economy has experienced periods of high inflationary pressures as well as periods of low or even deflationary conditions. The spikes noticed in the chart could be attributed to various factors, such as supply chain disruptions, weather-related events affecting agricultural production, rising input costs (e.g. fertilizers, fuel), or other external shocks.

After stationarity has been accomplished in the non-seasonal part of the data. There is a need to check for stationarity also in the seasonal component of the data before further analysis is carried out on the data. The result of the HEGY unit root test carried out on the differenced data for headline inflation and food inflation and the already stationary data of core inflation is presented in Table 2.

Table 2. HEGY unit root test result

Specification	Headline Inflation		Food Inflation		Core Inflation	
	Statistic	P-value	Statistic	P-value	Statistic	P-value
t_1	-8.2394	0.0000	-6.9107	0.0000	-5.709	0.0000
t_2	-6.4023	0.0000	-5.7459	0.0000	-3.8541	0.0001
F_3:4	42.009	0.0000	50.8489	0.0000	26.1446	0.0000
F_5:6	35.42	0.0000	30.703	0.0000	25.0181	0.0000
F_7:8	25.9675	0.0000	37.5116	0.0000	42.0432	0.0000
F_9:10	57.2815	0.0000	38.7425	0.0000	75.4356	0.0000
F_11:12	35.0101	0.0000	41.7703	0.0000	46.9615	0.0000
F_2:12	54.8	0.0000	48.0916	0.0000	366.3669	0.0000
F_1:12	54.0987	0.0000	47.9752	0.0000	351.8738	0.0000

Source: Computed by the author in R.

From Table 2, for headline inflation, the test statistic t_1 is -8.2394 with a p-value of 0.0000, which indicates strong evidence against the presence of a non-seasonal unit root (regular non-stationarity) in the headline inflation series. The test statistic t_2 is -6.4023 with a p-value of 0.0000, which suggests strong evidence against the presence of a semi-annual unit root (non-stationarity at the 6-month frequency). The F-statistics (F_3:4, F_5:6, F_7:8, F_9:10, F_11:12) all have p-values of 0.0000, indicating significant evidence against the presence of complex unit roots at various seasonal frequencies. The F_2:12 and F_1:12 statistics, which test for the joint presence of all unit roots, have p-values of 0.0000, providing strong evidence against any form of non-stationarity (regular or seasonal) in the headline inflation series.

For food inflation, the test statistic is -6.9107 with a p-value of 0.0000, which suggests strong evidence against the presence of a non-seasonal unit root in the food inflation series. The test statistic t_2 is -5.7459 with a p-value of 0.0000, which offers strong evidence against the presence of a semi-annual unit root. The F-statistics (F_3:4, F_5:6, F_7:8, F_9:10, F_11:12) all have p-values of 0.0000, providing significant evidence against the presence of complex unit roots at various seasonal frequencies. The F_2:12 and F_1:12 statistics have p-values of 0.0000, suggesting strong evidence against any form of non-stationarity (regular or seasonal) in the food inflation series. For core inflation, the test statistic t_1 is -5.709 with a p-value of 0.0000, which presents strong evidence against the presence of a non-seasonal unit root in the core inflation series. The test statistic t_2 is -3.8541 with a p-value of 0.0001, which provides significant evidence against the presence of a semi-annual unit root. The F-statistics (F_3:4, F_5:6, F_7:8, F_9:10, F_11:12) all have p-values of 0, providing significant

evidence against the presence of complex unit roots at various seasonal frequencies. The F_2:12 and F_1:12 statistics have p-values of 0.0000, indicating strong evidence against any form of non-stationarity (regular or seasonal) in the core inflation series.

Overall, the HEGY seasonal unit root test results consistently reject the presence of unit roots (both non-seasonal and seasonal) across all three-inflation series (headline, food, and core). This suggests that these series do not exhibit non-stationarity, either in the regular form or at various seasonal frequencies. (Note that the non-stationarity in the non-seasonal component of the headline inflation series and food inflation series is due to the first-order differencing that had been carried out on them before carrying out the HEGY test).

To determine the most appropriate SARIMA model specifications for forecasting headline inflation, food inflation, and core inflation in Nigeria, a rigorous model selection approach was employed. Initially, a range of potential SARIMA models was fitted to each inflation series, capturing different combinations of autoregressive (AR), integrated (I), and moving average (MA) components, as well as potential seasonal effects. For each inflation series, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were computed for the candidate SARIMA models. These information criteria provide a systematic way to balance the trade-off between model fit and parsimony, penalizing for excessive complexity. The models with the lowest AIC and BIC values for each inflation series were identified as the tentative best-fitting models, as highlighted in bold in the respective tables. However, to further validate the model selection and ensure accurate forecasting performance, the tentative best models were subjected to an out-of-sample evaluation. The models were used to generate forecasts for the test set, and their forecasting accuracies were assessed using the Root Mean Squared Error (RMSE) metric. The RMSE quantifies the average magnitude of the forecasting errors, with lower values indicating better forecasting precision.

For each inflation series, the model with the lowest RMSE on the test set was ultimately selected as the final SARIMA model specification. This approach ensured that the chosen models not only exhibited good in-sample fit, as indicated by the AIC and BIC, but also demonstrated superior out-of-sample forecasting performance, as measured by the RMSE.

By combining the information criteria (AIC and BIC) for in-sample model selection and the RMSE for out-of-sample evaluation, this methodological framework aimed to strike a balance between model parsimony, goodness-of-fit, and forecasting accuracy. The resulting SARIMA model specifications, optimized for each inflation series, provide a robust foundation for generating reliable inflation forecasts, which are crucial for informing economic policymaking and maintaining price stability in Nigeria.

4.1. Model Identification

In Table 3, Table 4, and Table 5, the models with the least AIC and BIC have been bolded for each series and the models are presented in Table 6. The models with the least AIC and BIC from each inflation series are presented in the table. From the possible models for the three-inflation series, we select the models with the least AIC and BIC from each of them for further comparisons. The models with the lowest AIC and BIC have been bolded in the tables, and each of them is used to forecast the values of the test set and we check for their RMSE. The model with the least RMSE for each CPI is then selected as the model for the CPI.

Table 3. The considered SARIMA models for headline inflation and their AICs and BICs

Model	AIC	BIC
SARIMA (2,1,2) $(1,0,1)_{12}$	813.73	838.07
SARIMA (1,1,0) $(1,0,0)_{12}$	862.32	872.74
SARIMA (0,1,1) $(0,0,1)_{12}$	817.04	827.47
SARIMA (1,1,0) $(2,0,0)_{12}$	829.03	842.94
SARIMA (1,1,0) $(2,0,1)_{12}$	807.66	**825.05**
SARIMA (1,1,0) $(1,0,1)_{12}$	813.51	827.41
SARIMA (0,1,0) $(2,0,0)_{12}$	835.32	845.75
SARIMA (2,1,0) $(2,0,0)_{12}$	830.81	848.19
SARIMA (2,1,0) $(1,0,0)_{12}$	864.31	878.22
SARIMA (2,1,0) $(2,0,1)_{12}$	808.64	829.50
SARIMA (2,1,0) $(1,0,1)_{12}$	814.87	832.25
SARIMA (2,1,0) $(1,0,2)_{12}$	**805.5**	826.36
SARIMA (1,1,0) $(2,0,2)_{12}$	806.07	826.92
SARIMA (1,1,1) $(2,0,2)_{12}$	807.68	832.02

Source: Computed by the author in R.

Table 4. The considered SARIMA models for food inflation and their AICs and BICs

Model	AIC	BIC
SARIMA (2,1,2) $(1,0,1)_{12}$	990.71	1015.05
SARIMA (1,1,0) $(1,0,0)_{12}$	1060.64	1071.07
SARIMA (0,1,1) $(0,0,1)_{12}$	999.75	1010.18
SARIMA (0,1,1) $(1,0,1)_{12}$	997.97	1011.88
SARIMA (0,1,1) $(1,0,0)_{12}$	1058.55	1068.98
SARIMA (0,1,1) $(2,0,1)_{12}$	989.34	**1006.72**
SARIMA (0,1,1) $(2,0,0)_{12}$	1009.77	1023.67
SARIMA (0,1,1) $(2,0,2)_{12}$	991.33	1012.19
SARIMA (0,1,1) $(1,0,2)_{12}$	992.15	1009.53
SARIMA (0,1,0) $(2,0,2)_{12}$	995.22	1012.60
SARIMA (1,1,1) $(2,0,2)_{12}$	992.72	1017.06
SARIMA (0,1,2) $(2,0,2)_{12}$	991.19	1015.53

Table 4. The considered SARIMA models for food inflation and their AICs and BICs (cont.)

Model	AIC	BIC
SARIMA (0,1,2) (1,0,2) $_{12}$	991.97	1012.83
SARIMA (0,1,2) (2,0,1) $_{12}$	989.26	1010.12
SARIMA (0,1,2) (1,0,1) $_{12}$	998.72	1016.10
SARIMA (1,1,2) (2,0,2) $_{12}$	990.65	1018.46
SARIMA (1,1,2) (1,0,2) $_{12}$	992.04	1016.37
SARIMA (1,1,2) (2,0,1) $_{12}$	988.67	1013.00
SARIMA (1,1,2) (1,0,1) $_{12}$	998.58	1019.44
SARIMA (2,1,2) (2,0,2) $_{12}$	**983.17**	1014.46
SARIMA (1,1,3) (2,0,2) $_{12}$	990.03	1021.31
SARIMA (0,1,3) (2,0,2) $_{12}$	988.53	1016.35
SARIMA (0,1,3) (1,0,2) $_{12}$	990.11	1014.45
SARIMA (0,1,3) (2,0,1) $_{12}$	986.54	1010.88
SARIMA (0,1,3) (1,0,1) $_{12}$	995.21	1016.07
SARIMA (0,1,4) (2,0,2) $_{12}$	988.68	1019.98

Source: Computed by the author in R.

Table 5. The considered SARIMA models for food inflation and their AICs and BICs

Model	AIC	BIC
SARIMA (2,0,2) (1,0,1) 12	1032.34	1060.18
SARIMA (1,0,0) (1,0,0) 12	1100.96	1114.88
SARIMA (0,0,1) (0,0,1) 12	1335.35	1349.27
SARIMA (2,0,2) (0,0,1) 12	1030.34	1054.70
SARIMA (2,0,2) (1,0,0) 12	1098.94	1123.31
SARIMA (2,0,2) (1,0,2) 12	1034.13	1065.46
SARIMA (1,0,2) (2,0,1) 12	1029.51	1057.36
SARIMA (1,0,2) (1,0,1) 12	1031.89	1056.26
SARIMA (1,0,2) (2,0,0) 12	1080.88	1105.24
SARIMA (1,0,2) (1,0,0) 12	1103.86	1124.75
SARIMA (0,0,2) (2,0,0) 12	1248.4	1269.29
SARIMA (1,0,1) (2,0,0) 12	1078.91	1099.80
SARIMA (1,0,1) (1,0,0) 12	1101.88	1119.28
SARIMA (1,0,1) (2,0,1) 12	1027.56	1051.92
SARIMA (1,0,1) (1,0,1) 12	1029.92	1050.80
SARIMA (0,0,1) (2,0,0) 12	1343.98	1361.38
SARIMA (1,0,0) (2,0,0) $_{12}$	1077.5	1094.91
SARIMA (2,0,1) (2,0,0) $_{12}$	1076.12	1100.48
SARIMA (0,0,0) (2,0,0) $_{12}$	1567.23	1581.16
SARIMA (1,0,1) (2,0,0) $_{12}$	1078.91	1099.80

Source: Computed by the author in R.

Table 6. The potential SARIMA models with the lowest AIC and BIC for each of the CPIs

Specification	Model	AIC	BIC
Headline Inflation	SARIMA (2,1,0) (1,0,2) $_{12}$	805.5	826.36
	SARIMA (1,1,0) (2,0,1) $_{12}$	807.66	825.05
Food Inflation	SARIMA (2,1,2) (2,0,2) $_{12}$	983.17	1014.46
	SARIMA (0,1,1) (2,0,1) $_{12}$	989.34	1006.72
Core Inflation	SARIMA (1,0,1) (2,0,1) $_{12}$	1027.56	1051.92

Source: Computed by the author in R.

From Table 6, the SARIMA (2,1,0) (1,0,2) $_{12}$ has the least AIC and the SARIMA (1,1,0) (2,0,1) $_{12}$ has the least BIC among the models considered for the headline inflation rates series, whereas the SARIMA (2,1,2) (2,0,2) $_{12}$ has the least AIC and the SARIMA (0,1,1) (2,0,1) $_{12}$ has the least BIC among the models considered for the food inflation rates series. Among the models considered for the core inflation series, the SARIMA (1,0,1) (2,0,1) $_{12}$ model is the only selected model because it is the one with both the lowest AIC and the lowest BIC among them.

Figure 2, Figure 3 and Figure 4 show how well the models fit the training set data when applied to them.

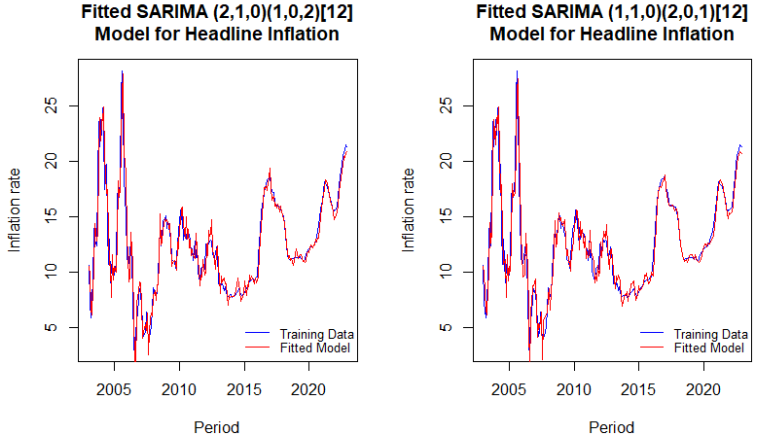

Figure 2. Fitted potential models for headline inflation

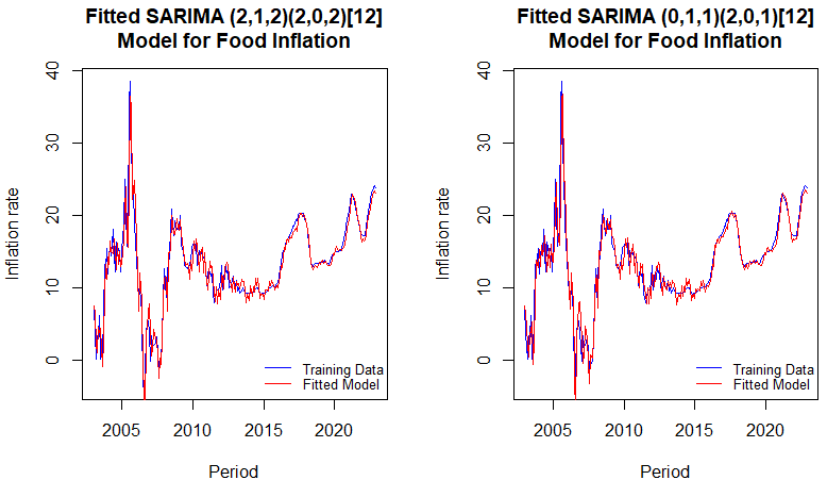

Figure 3. Fitted potential models for food inflation

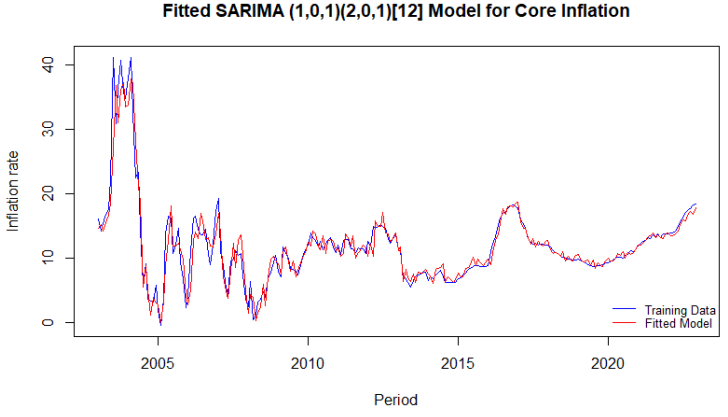

Figure 4. Fitted potential model for core inflation

The selected models are applied to the test set to determine their RMSE. Table 7 gives the result of the RMSE.

Table 7. The potential SARIMA models for each of the CPIs with their RMSE when applied to the test set

Specification	Model	RMSE
Headline Inflation	SARIMA $(2,1,0)\ (1,0,2)_{12}$	2.3519
	SARIMA $(1,1,0)\ (2,0,1)_{12}$	2.1256
Food Inflation	SARIMA $(2,1,2)\ (2,0,2)_{12}$	2.7727
	SARIMA $(0,1,1)\ (2,0,1)_{12}$	3.0498
Core Inflation	SARIMA $(1,0,1)\ (2,0,1)_{12}$	2.5503

Source: Computed by the author in R.

For headline inflation, the SARIMA (1,1,0) $(2,0,1)_{12}$ model emerged as the superior choice between the two models, exhibiting an RMSE of 2.1256 on the test set. In the case of food inflation, the SARIMA (2,1,2) $(2,0,2)_{12}$ model demonstrated the best forecasting performance between the two models, with an RMSE of 2.7727. For core inflation, the SARIMA (1,0,1) $(2,0,1)_{12}$ model was selected as the optimal specification, exhibiting an RMSE of 2.5503.

Model Estimation

The parameters of the 3 models for the CPI are estimated and the results of the estimation are provided.

Table 8. Parameter estimation of the selected SARIMA model for each of the CPIs

Specification	Model	Parameter	Estimate	Std. Error	z value	Pr(>|z|)
Headline Inflation	SARIMA(1,1,0)(2,0,1) $_{12}$	ar1	0.16182	0.064846	2.4955	0.012578 *
		sar1	0.11934	0.092958	1.2838	0.199224
		sar2	-0.2603	0.089145	-2.9205	0.003495 **
		sma1	-0.7631	0.086139	-8.8588	< 2.2e-16 ***
Food Inflation	SARIMA(2,1,2)(2,0,2) $_{12}$	ar1	0.56993	0.164129	3.4725	0.0005157 ***
		ar2	-0.7439	0.115236	-6.4558	1.076e-10 ***
		ma1	-0.4268	0.191911	-2.2237	0.0261655 *
		ma2	0.5687	0.153424	3.7067	0.0002100 ***
		sar1	0.04189	0.351682	0.1191	0.905188
		sar2	-0.278	0.120592	-2.3053	0.0211486 *
		sma1	-0.7244	0.364953	-1.9849	0.0471544 *
		sma2	-0.0273	0.335836	-0.0813	0.935195
Core Inflation	SARIMA(1,0,1)(2,0,1) $_{12}$	ar1	0.965653	0.023734	40.6868	< 2e-16 ***
		ma1	0.02958	0.069008	0.4286	0.66818
		sar1	-0.00563	0.068949	-0.0817	0.93492
		sar2	-0.18066	0.084864	-2.1288	0.03327 *
		sma1	-0.99999	0.089439	-11.1807	< 2e-16 ***
		Intercept	11.26829	0.430739	26.1604	< 2e-16 ***

Source: Computed by the author in R.

Table 8 gives information about the estimates of the parameters of the selected models. For the selected headline inflation the SARIMA (1,1,0)(2,0,1) $_{12}$ model, ar1 (0.16182) is the coefficient of the non-seasonal autoregressive component of order 1, suggesting a positive influence of the previous period's headline inflation on the current value. This coefficient is statistically significant (p-value = 0.012578). sar1 (0.11934) and sar2 (-0.2603) are the coefficients of the seasonal autoregressive components of orders 1 and 2, respectively, capturing the influence of headline inflation values from the same period in the previous year and two years earlier. Only sar2 is statistically significant (p-value = 0.003495). sma1 (-0.7631) is the coefficient of the seasonal moving average component of order 1, accounting for the impact of past seasonal errors on the current

headline inflation. This coefficient is highly significant (p-value < 2.2e-16). For the selected food inflation, the SARIMA (2,1,2)(2,0,2) $_{12}$ model, ar1 (0.56993) and ar2 (-0.7439) are the coefficients of the non-seasonal autoregressive components of orders 1 and 2, respectively, capturing the influence of previous food inflation values. Both coefficients are statistically significant (p-values < 0.001). ma1 (-0.4268) and ma2 (0.5687) are the coefficients of the non-seasonal moving average components of orders 1 and 2, accounting for the impact of past errors on current food inflation. Both coefficients are statistically significant (p-values < 0.05). sar1 (0.04189) and sar2 (-0.278) are the coefficients of the seasonal autoregressive components of orders 1 and 2, respectively. Only sar2 is statistically significant (p-value = 0.0211486). sma1 (-0.7244) and sma2 (-0.0273) are the coefficients of the seasonal moving average components of orders 1 and 2. Only sma1 is statistically significant (p-value = 0.0471544).

For the selected core inflation, the SARIMA (1,0,1)(2,0,1) $_{12}$ model, ar1 (0.965653) is the coefficient of the non-seasonal autoregressive component of order 1, indicating a strong positive influence of the previous period's core inflation on the current value. This coefficient is highly significant (p-value < 2e-16). ma1 (0.02958) is the coefficient of the non-seasonal moving average component of order 1, but it is not statistically significant (p-value = 0.66818). sar1 (-0.00563) and sar2 (-0.18066) are the coefficients of the seasonal autoregressive components of orders 1 and 2, respectively. Only sar2 is statistically significant (p-value = 0.03327). sma1 (-0.99999) is the coefficient of the seasonal moving average component of order 1, accounting for the impact of past seasonal errors on current core inflation. This coefficient is highly significant (p-value < 2e-16). Intercept (11.26829) represents the constant term in the core inflation model, indicating a baseline level of core inflation. This coefficient is highly significant (p-value < 2e-16).

Diagnostic Analysis

Diagnostic analysis is carried out on the models to verify the models' adequacy. The residuals of the models were tested for correlation the Ljung-Box test. The result of the test is presented in Table 9.

Table 9. Ljung-Box test result

CPI	Model	Test Statistic	Pvalue
Headline Inflation	SARIMA $(1,1,0)(2,0,1)_{12}$	0.013797	0.9065
Food Inflation	SARIMA $(2,1,2)(2,0,2)_{12}$	0.055919	0.8131
Core Inflation	SARIMA $(1,0,1)(2,0,1)_{12}$	0.00069325	0.9790

Source: Computed by the author in R.

Based on the Ljung-Box test results provided for the selected SARIMA models for headline inflation, food inflation, and core inflation in Table 9, the following interpretations are made.

For the headline inflation model, the Ljung-Box test statistic is 0.013797, with an associated p-value of 0.9065. Since the p-value is greater than 0.05, we fail to reject the null hypothesis of independently distributed residuals. This result suggests that the residuals from the SARIMA (1,1,0)(2,0,1)12 model for headline inflation do not exhibit significant autocorrelation, indicating that the model adequately captures the dynamics of the headline inflation series. Similarly, for the food inflation model, the Ljung-Box test statistic is 0.055919, with a p-value of 0.8131. Since the p-value is greater than 0.05, we fail to reject the null hypothesis of independent residuals. This result provides evidence that the SARIMA (2,1,2)(2,0,2)12 model for food inflation appropriately accounts for the underlying patterns and dependencies in the data, as the residuals do not exhibit significant autocorrelation.

For the core inflation model, the Ljung-Box test statistic is 0.00069325, with a p-value of 0.9790. Once again, the p-value is substantially larger than the 0.05 significance level, leading to the null hypothesis of independent residuals not being rejected. This outcome indicates that the SARIMA (1,0,1)(2,0,1)12 model for core inflation adequately captures the essential dynamics and patterns in the data, as evidenced by the lack of significant autocorrelation in the residuals. The Ljung-Box test results across all three inflation series unanimously fail to reject the null hypothesis of independent residuals, providing strong evidence that the selected SARIMA models are appropriate for modeling and forecasting the respective inflation series.

The result reported in Table 9 suggests that the models are well specified. The diagnostics indicate that the residuals of the models are normally distributed.

5. Conclusion

Inflation rate is a key indicator for assessing price stability within an economy, particularly given the significant microeconomic and macroeconomic impacts associated with rising inflation. Also, precise inflation forecasts empower central banks to take proactive measures against potential inflationary threats. Consequently, this study delves deeper into the dynamics of headline, food, and core inflation in Nigeria by employing SARIMA processes for inflation forecasting. Monthly inflation data from January 2003 to December 2023 obtained from the CBN website were analyzed and the data consisted of 249 observations.

Findings showed that the selected SARIMA specifications for the data achieved the best forecasting performance among the evaluated candidate models. The results revealed that these models do not fully capture the intricate dynamics underlying the

inflation series. Despite the rigorous model identification and diagnostic procedures, the forecasts generated by the SARIMA models exhibited deviations from the actual observed inflation rates when evaluated on the test set. Specifically, the models demonstrated a propensity to overestimate or underestimate the true inflation values at certain time points, highlighting the challenges in accurately modeling complex economic phenomena such as inflation.

The discrepancies between the forecasts and the observed data underscore the limitations of the SARIMA framework in accounting for all the potential factors influencing inflation rates. Inflation is a multifaceted phenomenon, driven by a myriad of interrelated variables, including monetary policy decisions, global market dynamics, supply chain disruptions, and consumer behavior, among others. The SARIMA models, while capturing the general trends and seasonality, may fail to fully incorporate the intricate interplay of these factors, leading to forecast inaccuracies. It is crucial to acknowledge that forecasting inflation rates is a complex endeavor, and no single model can perfectly capture the complex dynamics involved. The SARIMA models employed in this study represent a valuable tool for understanding and forecasting inflation patterns, but their limitations should be recognized, and their outputs should be interpreted with caution.

Ultimately, the findings of this study highlight the importance of continuous model evaluation, refinement, and the integration of multiple forecasting techniques to better understand and anticipate the complex dynamics of inflation in Nigeria. Accurate inflation forecasting is crucial for informed policymaking and effective economic management, underscoring the need for ongoing research and methodological advancements in this domain. Future research can explore SARIMAX modeling of inflation by assuming that inflation is driven by certain exogenous variables.

References

Adelekun, O. G., Abiola, O. H. and Constance, A. U., (2020). Modeling and Forecasting Inflation Rate in Nigeria using ARIMA Models. *KASU Journal of Mathematical Sciences,* 1(2). Retrieved from http:/www.journal.kasu.edu.ng/index.php/kjms

Akaike, H., (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on automatic control*, pp. 716–723.

Baumol, W. J., (1967). Macroeconomics of Unbalanced Growth: The Anatomy of Urban Crisis. *The American Economic Review, 57*(3), pp. 415–426. https://doi.org/10.2307/1812111.

Box, G. E., Jenkins, G. M., Reinsel, G. C. and Ljung, G. M., (2015). *Time Series Analysis: Forecasting and Control, 5th Edition.* Wiley.

Doguwa, S. I., & Alade, S. O., (2013, December). Short-Term Inflation Forecasting Models for Nigeria. *CBN Journal of Applied Statistics,* 4(3), pp. 1–29.

Fasanya, I. O., & Adekoya, O. B., (2017). Modeling inflation rate volatility in Nigeria with Structural Breaks. *CBN Journal of Applied Statistics,* 8(1), pp. 175–193.

Friedman, M., (1970). A Theoretical Framework for Monetary Analysis. *Journal of Political Economy,* 78(2), pp. 193–238.

Friedman, M., (1971). A Monetary Theory of Nominal Income. *Journal of Political Economy,* 79(2), pp. 323–337.

Gil-Alana, L. A., Shittu, O. I. and Yaya, O. S., (2012). Long memory, Structural breaks and Mean shifts in the Inflation rates in Nigeria. *African Journal of Business Management*, 6(3), pp. 888–897.

Gil-Alana, L. A., Yaya, O. S. and Solademi, E. A., (2016). Testing unit roots, structural breaks and linearity in the inflation rates of the G7 countries with fractional dependence techniques. *Applied Stochastic Models in Business and Industry*, 32, pp. 711–724.

Hylleberg, S., Engle, R., Granger, C. and Yoo, B., (1990). Seasonal Integration and Cointegration. *Journal of Econometrics,* 44, pp. 215–238.

Kabbilawsh, P., Sathish Kumar, D. and Chithra, N., (2020). *Trend analysis and SARIMA forecasting of mean maximum and mean minimum monthly temperature for the state of Kerala, India* (Vol. 68). Acta Geophys. Retrieved from https://doi.org/10.1007/s11600-020-00462-9.

Kelilume, I., Salami, A., (2013). Modeling and Forecasting Inflation with ARIMA and VAR: The Case of Nigeria. *Global Conference on Business and Finance Proceedings,* 8(1), pp. 62–72.

Laidler, D. D., Parkin, M. (1975). Inflation: A Survey. *The Economic Journal,* 84(340), pp. 741–809. https://doi.org/10.2307/2230624.

Ljung, G. M., Box, G. E., (1978). On a measure of lack of fit in time series models. *Biometrika,* 65(2), pp. 297–303.

Maynard, G., Ryckeghem, V., (1976). *A World of Inflation.* London: B.T. Bastford.

Nakorji, M., Aminu, M., (2022). Forecasting Inflation Using Machine Learning Techniques. *The Review of Finance and Banking,* 14(1), pp. 45–55. Retrieved from http//dx.doi.org/10.24818/rfb.22.14.01.04.

Nyoni, T., Nathaniel, S., (2018). Modeling rates of inflation in Nigeria: an application of ARMA, ARIMA and GARCH models. *Munich Personal RePEc Archive*.

Okwori, J., Abu, J., (2017). Monetary Policy and Inflation Targeting in Nigeria. *International Journal of Economics and Financial Management,* 2(3).

Olubusoye, O. E., Oyaromade, R., (2018). Modeling the Inflation Process in Nigeria. *African Economic Research Consortium, Nairobi*.

Olubusoye, O. E., Akanbi, O. B. and Akintande, O., (2017). Forecasting and Determining the Predictors of Inflation in Nigeria: A Bayesian Model Averaging Approach.

Oyewale, A. M., Kasali, A. O., M., K. P., Abiodun, M. V., Obioma, E. N. and Adeyinka, A. I., (2019). Forecasting Inflation Rates Using Artificial Neural. *International Journal of Statistics and Applications,* 9(6), pp. 201–207. https://doi.org/10.5923/j.statistics. 20190906.05.

Sani, B., Abdullahi, I. S. and Adamu, I., (2016, June). Analysis of Inflation Dynamics in Nigeria (1981–2015). *CBN Journal of Applied Statistics,* 7(1(b)).

Shaibu, I., & Osamwonyi, I. O., (2020, February). A Comparative Analysis of Inflation Dynamics Models in Nigeria. *International Journal of Academic Research in Business and Social Sciences,* 10(2), pp. 558–557. doi:10.6007/IJARBSS/v10-i2/6949.

Streeten, P. (1962). Wages, Prices, and Productivity. *Kyklos,* 15(4), pp. 723–733. https://doi.org/10.1111/j.1467-6435.1962.tb00089.x.

Tumala, M. M., Olubusoye, O. E., Yaaba, B. N., Yaya, O. S. and Akanbi, O. B., (2017). Forecasting and Determining the Predictors of Inflation in Nigeria: A Bayesian Model Averaging approach. A Technical Report Submitted to the Department of Statistics, Central Bank of Nigeria, Abuja. 78 pp.

Tumala, M. M., Olubusoye, O. E., Yaaba, B. N., Yaya, O. S. and Akanbi, O. B., (2019). Forecasting Nigerian Inflation using Model Averaging methods: Modeling Frameworks to Central Banks. *The Empirical Economics Review*, 9(1), pp. 47–72.

Uko, A. K., Nkoro, E., (2012). Inflation Forecasts with ARIMA, Vector Autoregressive and Error Correction Models in Nigeria. *European Journal of Economics, Finance and Administrative Sciences*, pp. 71–87. Retrieved from http://www.eurojournals.com/EJEFAS.htm.

Yaya, O. S., Shittu, O. I., (2010). Long Memory and Estimation of Memory Parameters: Nigerian and US Inflation Rates. *International Journal of Physical Sciences*, 2(3), pp. 120–131.

Yaya, O. S., (2018). Another Look at the Stationarity of Inflation rates in OECD countries: Application of Structural break-GARCH-based unit root tests. *Statistics in Transition new series*, 19(3), pp. 477–492.

Yaya, O. S., Ogbonna, A. E. and Atoi, N. V., (2019). Are inflation rates in OECD countries actually stationary during 2011-2018? Evidence based on Fourier Nonlinear Unit root tests with Break. *The Empirical Economics Review*, 9(4), pp. 309–325.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 207–228, https://doi.org/10.59139/stattrans-2026-034
Received – 06.11.2024; accepted – 20.10.2025

Moments and estimation of Burr Hatke Exponential model based on generalized order statistics with real data applications

Farhan Ansari[1], **M. J. S. Khan**[2]

Abstract

This paper presents a recurrence relation for single and product moments of the Burr-Hatke exponential distribution based on generalized order statistics and record values. Moreover, we have derived the explicit and exact expressions for the single moment of generalized order statistics from the Burr-Hatke exponential distribution. Furthermore, the classical and Bayesian estimation procedures for estimating the Burr-Hatke exponential distribution parameter are discussed when the data are in the form of records and *gos*. In addition, the asymptotic confidence interval, boot-p confidence interval, and credible interval for the parameter of the Burr-Hatke exponential distribution are discussed. Monte Carlo simulations were conducted to evaluate the performance of the estimators derived from both classical and Bayesian approaches, with a particular emphasis on their mean squared error. Moreover, two real datasets are included for illustrative purposes in this study.

Keywords: Generalized order statistics, single moments, recurrence relations, Burr Hatke exponential distribution, MLE, Bayesian estimation, credible interval.

1. Introduction

Kamps (1995) introduced the concept of generalized order statistics (*gos*). This concept incorporates various significant models related to ordered random variables (*rv's*). These models include order statistics (OS), non-integral sample size order statistics, sequential order statistics, progressive Type II censored order statistics, record values, and Pfeifer's record values, all embedded within the framework of *gos*.

Let $n \in \mathbb{N}, k \geq 1, \widetilde{m} = (m_1, m_2, \ldots, m_{n-1}) \in \mathbb{R}^{n-1}$, be the parameters such that, for all $r \in \{1, \ldots, n-1\}$, $\gamma_r = k + n - r + \sum_{j=r}^{n-1} m_j \geq 1$, then $X(1,n,\widetilde{m},k), X(2,n,\widetilde{m},k), \ldots, X(n,n,\widetilde{m},k)$ are termed as *gos*, if their joint probability density function (*pdf*) is given by

$$\begin{aligned} & f_{X(1,n,\widetilde{m},k),X(2,n,\widetilde{m},k),\ldots,X(n,n,\widetilde{m},k)}(x_1, \ldots, x_n) \\ & = k \left(\prod_{j=1}^{n-1} \gamma_j \right) \left(\prod_{i=1}^{n-1} (\bar{F}(x_i))^{m_i} f(x_i) \right) (\bar{F}(x_n))^{k-1} f(x_n), \end{aligned} \tag{1}$$

[1]Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, India. E-mail: farhanamu.stats@gmail.com. ORCID: https://orcid.org/0000-0002-8199-0047.
[2]Department of Statistics and Operations Research, Aligarh Muslim University, Aligarh, India. E-mail: jahangirskhan@gmail.com. ORCID: https://orcid.org/0000-0003-2264-9815.

on the cone $F^{-1}(1) > x_n \geq \ldots \geq x_1 > F^{-1}(0)$ of $\mathbb{R}^n$, where $\bar{F}(x) = 1 - F(x)$ denotes the survival function.

By modifying the parameters $\widetilde{m}$ and k, it is possible to construct a wide variety of significant models that fall under the category of ordered *rv's*. For instance, OS can be obtained by setting $m_j = 0,\ k = 1$ *and* $\gamma_j = n - j + 1$, where $1 \leq j \leq n-1$. Similarly, order statistics with non-integral sample size are generated when $m_j = 0,\ k = \beta - n + 1$ *and* $\gamma_j = \beta - j + 1$. In the case of progressive Type II censored order statistics, $m_j = R_j \in \mathbb{N}_0,\ k = R_n + 1$ *and* $\gamma_j = n - j + 1 + \sum_{i=r}^{n} R_i$. Sequential order statistics arise when $k = \beta_n,\ m_j = (n-j+1)\alpha_j - (n-j)\alpha_{j+1} - 1$ and $\gamma_j = (n-j+1)\beta_j; \beta_1, \ldots, \beta_n > 0$, while record values are produced by letting $m_j \to -1,\ k = 1$. Extending this further, k^{th} record values are defined by $m_j \to -1,\ k \in \mathbb{N}$, whereas Pfeifer's record values are obtained by setting $m_j = \beta_j - \beta_{j+1} - 1,\ k = \beta_n$ and $\gamma_j = \beta_j$.
When γ_j's are pairwise different, i.e. $\gamma_j \neq \gamma_i\ ; \forall j \neq i = 1, 2, \ldots, n-1$. Kamps and Cramer (2001) derived the *pdf* of $X(r, n, \widetilde{m}, k)$ as:

$$f_{X(r,n,\widetilde{m},k)}(x) = c_{r-1}\left[\sum_{i=1}^{r} a_i(r)[\bar{F}(x)]^{\gamma_i}\right]\frac{f(x)}{\bar{F}(x)}, \quad -\infty < x < \infty \tag{2}$$

and for $1 \leq r < s \leq n$, the joint *pdf* of $X(r, n, \widetilde{m}, k)$ and $X(s, n, \widetilde{m}, k)$ is given by (Kamps and Cramer, 2001)

$$f_{X(r,n,\widetilde{m},k),X(s,n,\widetilde{m},k)}(x,y) = c_{s-1}\left[\sum_{t=r+1}^{s} a_t^{(r)}(s)\left(\frac{\bar{F}(y)}{\bar{F}(x)}\right)^{\gamma_t}\right]\left[\sum_{i=1}^{r} a_i(r)(\bar{F}(x))^{\gamma_i}\right] \times \frac{f(x)}{\bar{F}(x)}\frac{f(y)}{\bar{F}(y)}, \quad -\infty < x < y < \infty. \tag{3}$$

In case of records, i.e. $m \to -1,\ k = 1$, the *pdf* of $X(r, n, -1, 1)$ is given by

$$f_{X(r,n,-1,1)}(x) = \frac{1}{(r-1)!}[-\ln \bar{F}(x)]^{r-1} f(x), \quad -\infty < x < \infty. \tag{4}$$

The joint *pdf* of $X(r, n, -1, 1)$ and $X(s, n, -1, 1)$, $1 \leq r < s$, is given by

$$f_{X(r,n,-1,1),X(s,n,-1,1)}(x,y) = [(r-1)!(s-r-1)!]^{-1}[-\ln \bar{F}(x)]^{r-1}[\ln \bar{F}(x) - \ln \bar{F}(y)]^{s-r-1} \times \frac{f(x)}{\bar{F}(x)} f(y), \quad -\infty < x < y < \infty. \tag{5}$$

Further, j^{th} moment of r^{th} *gos* is defined as:

$$\mu^{j}_{r,n,\widetilde{m},k} = E[X^j(r, n, \widetilde{m}, k)] = c_{r-1}\sum_{i=1}^{r} a_i(r)\int_{-\infty}^{\infty} x^j [\bar{F}(x)]^{\gamma_i}\frac{f(x)}{\bar{F}(x)}\,dx,$$

and j^{th} and l^{th} product moment of r^{th} and s^{th} *gos* is defined as:

$$\mu^{j,l}_{r,s,n,\widetilde{m},k} = E[X^j(r,n,\widetilde{m},k)Y^l(s,n,\widetilde{m},k)] = c_{s-1}\sum_{i=1}^{r}\sum_{t=r+1}^{s} a_i(r)a_t^r(s)\int_{-\infty}^{\infty}\int_{x}^{\infty} x^j y^l \times [\bar{F}(y)]^{\gamma_t}[\bar{F}(x)]^{\gamma_i-\gamma_t}\frac{f(x)}{\bar{F}(x)}\frac{f(y)}{\bar{F}(y)}\,dy\,dx.$$

Where, $c_{r-1} = \prod_{j=1}^{r}\gamma_j$, $a_i(r) = \prod_{j=1}^{r}(\gamma_j-\gamma_i)^{-1}$, $\forall\ \gamma_i \neq \gamma_j \quad 1 \leq i \leq r \leq n$, and $a_t^{(r)}(s) = \prod_{j=r+1,t\neq j}^{s}(\gamma_j-\gamma_t)^{-1}$, $\gamma_t \neq \gamma_j$, $\forall\ \gamma_j \neq \gamma_t$, $r < t \leq s$.

Another important case of *gos* arises when the parameters $m_1 = m_2 = \ldots = m_{n-1} = m \neq -1$, termed as *m-gos*. In this case, $a_i(r)$ and $a_t^{(r)}(s)$ may further be simplified as (Khan and Khan, 2012)

$$a_i(r) = (-1)^{r-i}((r-1)!)^{-1}(m+1)^{-(r-1)}\binom{r-1}{r-i}, \tag{6}$$

$$a_t^{(r)}(s) = (-1)^{s-t}((s-r-1)!)^{-1}(m+1)^{-(s-r-1)}\binom{s-r-1}{s-t}. \tag{7}$$

With applications ranging from reliability to medicine, the one-parameter exponential distribution is frequently used to characterize the amount of time that passes between events. Furthermore, exponential distribution is a fundamental tool for the study of Poisson and Markov processes. A non-negative *rv* T is said to follow an exponential distribution if its cumulative distribution function (*cdf*) and the *pdf* are given by respectively

$$G(t;\theta) = 1-\exp(-\theta t), \quad 0 < t < \infty, \quad \theta > 0, \tag{8}$$

$$g(t;\theta) = \theta\exp(-\theta t), \quad 0 < t < \infty, \quad \theta > 0. \tag{9}$$

The Burr-Hatke exponential (BHE) distribution was introduced by Yadav et al. (2021) using the Burr-Hatke-G (BH-G) family of distributions, which is given by Yousof et al. (2018). The BHE distribution's *pdf* is right-skewed, and its hazard rate function (*hrf*) exhibits a declining pattern, as shown in Figures 1-2.

A *rv* X follows the BHE distribution if its *cdf* is of the form:

$$F(x) = 1-\frac{e^{-\theta x}}{1+\theta x}, \quad 0 < x < \infty, \quad \theta > 0. \tag{10}$$

Furthermore, its *pdf* is given as follows:

$$f(x) = \theta e^{-\theta x}\frac{2+\theta x}{(1+\theta x)^2}, \quad 0 < x < \infty, \quad \theta > 0. \tag{11}$$

The BHE distribution's *hrf* can be found as:

$$h(x) = \frac{\theta(2+\theta x)}{(1+\theta x)}. \tag{12}$$

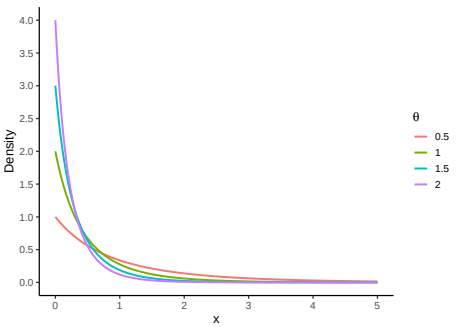

Figure 1. *pdf* plot of the BHE distribution

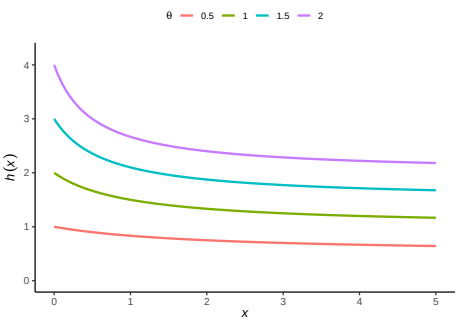

Figure 2. *hrf* plot of the BHE distribution

Quantile Function of BHE Distribution: The quantile, when it can be derived analytically, forms the basis for data generation. This facilitates the simulation of random samples, thereby enabling the determination of a model's behavior. The quantile function (QF) can be obtained by adopting the technique of Jodrá (2010) using $X = F^{-1}(U)$, where *rv* U is generated from $U(0,1)$. Notably, for certain probability distributions, the QFs can be formulated using the the Lambert W function. A review of the relevant literature, including citations such as Jodrá and Arshad (2022), Barco et al. (2017), and others, provides compelling evidence to support this claim.

For the sake of completeness, it is essential to recall the definition of the Lambert W function, which is a multivalued, complex function. This function can be expressed by the equation $w(z)e^{w(z)} = z$, where z represents the complex number. For real $z \geq -1/e$, the Lambert W function possesses two distinct real branch points, designated as W_0 (principal branch) and W_{-1} (negative branch), respectively. The function $W_0(z)$ assumes values within the interval $[-1,\infty)$ for $z \geq -1/e$, whereas $W_{-1}(z)$ assumes values within the interval $(-\infty,-1]$ for $z \in [-1/e,0)$ (see Jodrá and Arshad (2022) and Jodrá (2010)).

Theorem 1. For any $\theta \in \mathbb{R}^+$, the BHE distribution's quantile function may be represented by

$$Q(u;\theta) = \frac{1}{\theta}\left(W_0\left(\frac{e}{1-u}\right) - 1\right), \quad 0 < u < 1. \tag{13}$$

Proof: For any $u \in (0,1)$, the function $F(x) = u$ has to be solved *w.r.t.* x. After simplification, we get

$$(1+\theta x)e^{(1+\theta x)} = \frac{e}{1-u}.$$

By combining the aforementioned equation with the definition of the Lambert W function, we can derive the following result:

$$W\left(\frac{e}{1-u}\right) = 1 + \theta x, \tag{14}$$

where $\frac{e}{1-u} \geq -\frac{1}{e}$, the result is true. In light of the aforementioned properties of W_0, it can be concluded that the branch of W in Equation (14) is the principal branch. Ultimately, the proof is completed upon solving Equation (14) for x, which yields Equation (13).

In the last few years, various studies have been conducted on the extensions and properties of the Burr-Hatke exponential distribution and its variants. Abouelmagd (2018) studied logarithmic BHE distribution and obtained its moments and maximum likelihood estimates (MLEs). Yousof et al. (2018) proposed the BH-G family of distributions by using the BH differential equation, whereas discrete forms of the BHE distribution were discussed in Morshedy et al. (2020) and Pandey et al. (2021). Afify et al. (2021) and El-Morshedy et al. (2021) discussed the classical and Bayesian estimation for the two-parameter and exponentiated BH distribution, respectively. Bhatti et al. (2021) derived the extended Chen distribution using a generalized BH differential equation and the relation between the exponential and gamma variables. The authors also obtained the MLE of an unknown parameter of the extended Chen distribution. Yadav et al. (2021) discussed Bayesian and classical estimation under Type II hybrid censoring. Khan et al. (2021) derived the moments and estimation of the inverted Topp-Leone distribution based on record values. Some other related works are Almuqrin (2023), Shirozhan et al. (2023), and Ragab et al. (2024), who discussed new BH-type models and their inferential properties.

Several authors have discussed classical and Bayesian inference under *gos*. For instance, Kumar (2017), Khan et al. (2018), Khan and Iqrar (2019), Khan and Khatoon (2020), Khan et al. (2020), established several moment relations and estimation procedures. Other studies, such as Alawady et al. (2022), Aly (2023), and Arshad et al. (2023), dealt with Bayesian and classical estimations, recurrence relations, predictive inference, and reliability estimation using *gos*.

The manuscript is structured into eight sections. In Section 2, we established the recurrence relations for the BHE distribution based on *gos* and records. In Section 3, we derive the exact moment of BHE distribution based on *gos*. In Section 4, the MLE of BHE distribution parameters is presented. In Section 5, we derive a Bayesian estimation of the scale parameter of the BHE distribution based on *gos*. In Section 6, two algorithms for generating a random sample from the BHE distribution based on *gos* are developed. Furthermore, the performance of the BHE distribution parameter is evaluated using the Markov Chain Monte Carlo (MCMC) simulation technique. To demonstrate the flexibility and applicability of the results, two real datasets are considered in Section 7. Moreover, the manuscript is concluded in Section 8.

2. Recursive relations for the moments of the BHE distribution

The recursive relations for the single and product moments of the BHE distribution derived from *gos* are presented in this section. Utilizing these recursive relations, one can compute the higher order moments.

Theorem 2. For any $j_1 \in \mathbb{N}_0$ and $r = 1,2,\ldots,n$, the recurrence for single moment of the BHE distribution based on gos is given by

$$\begin{aligned}&\mu^{j_1}_{r,n,\widetilde{m},k} + \theta\left(1 - \frac{2\gamma_r}{j_1+1}\right)\mu^{j_1+1}_{r,n,\widetilde{m},k} + \left(\frac{2\theta\gamma_r}{j_1+1}\right)\mu^{j_1}_{r-1,n,\widetilde{m},k}\\ &= \left(\frac{\theta^2\gamma_r}{j_1+2}\right)\left(\mu^{j_1+2}_{r,n,\widetilde{m},k} - \mu^{j_1+2}_{r-1,n,\widetilde{m},k}\right),\end{aligned}$$

and in the case of records ($m \to -1,\ k = 1$), the recursive relation for a single moment of the BHE distribution is given by

$$\begin{aligned}&\mu^{j_1}_{r,n,-1,1} + \theta\left(1 - \frac{2}{j_1+1}\right)\mu^{j_1+1}_{r,n,-1,1} + \left(\frac{2\theta}{j_1+1}\right)\mu^{j_1}_{r-1,n,-1,1}\\ &= \left(\frac{\theta^2}{j_1+2}\right)\left(\mu^{j_1+2}_{r,n,-1,1} - \mu^{j_1+2}_{r-1,n,-1,1}\right).\end{aligned}$$

Proof : In light of (2) and (12), we get

$$\begin{aligned}\mu^{j_1}_{r,n,\widetilde{m},k} + \theta\mu^{j_1+1}_{r,n,\widetilde{m},k} &= 2\theta c_{r-1}\sum_{i=1}^{r} a_i(r)\int_0^\infty x^{j_1}[\bar{F}(x)]^{\gamma_i}\,dx\\ &\quad + \theta^2 c_{r-1}\sum_{i=1}^{r} a_i(r)\int_0^\infty x^{j_1+1}[\bar{F}(x)]^{\gamma_i}\,dx.\end{aligned} \tag{15}$$

Simplifying the aforementioned integration, treating $[\bar{F}(x)]^{\gamma_i}$ as the part of integrand and simplifying, we get the theorem.

Similarly, in case the of records ($m \to -1,\ k = 1$), and in view of (4) and (12), we have

$$\begin{aligned}\mu^{j_1}_{r,n,-1,1} + \theta\mu^{j_1+1}_{r,n,-1,1} &= \frac{2\theta}{r!}\int_0^\infty x^{j_1}[-\ln\bar{F}(x)]^{r-1}\bar{F}(x)\,dx\\ &\quad + \frac{\theta^2}{r!}\int_0^\infty x^{j_1+1}[-\ln\bar{F}(x)]^{r-1}\bar{F}(x)\,dx.\end{aligned} \tag{16}$$

Simplifying the integration (16) by parts, $[-\ln\bar{F}(x)]^{r-1}\bar{F}(x)$ is considered part of the integrand. Thus, the theorem is proven.

Theorem 3. For any $j_1, j_2 \in \mathbb{N}_0$ and $1 \le r \le s+1 \le n$, the recursive relation for the product moment of the BHE distribution is given by

$$\begin{aligned}&\mu^{j_1,j_2}_{r,s,n,\widetilde{m},k} + \theta\left(1 - \frac{2\gamma_r}{j_2+1}\right)\mu^{j_1,j_2+1}_{r,s,n,\widetilde{m},k} + \left(\frac{2\theta}{j_2+1}\right)\mu^{j_1,j_2}_{r,s-1,n,\widetilde{m},k}\\ &= \left(\frac{\theta^2}{j_2+2}\right)\left[\mu^{j_1,j_2+2}_{r,s,n,\widetilde{m},k} - \mu^{j_1,j_2+2}_{r,s-1,n,\widetilde{m},k}\right],\end{aligned}$$

and in the case of records ($m \to -1,\ k = 1$), the recursive relation for the product moment of the BHE distribution is as follows:

$$\begin{aligned}&\mu^{j_1,j_2}_{r,s,n,-1,1} + \theta\left(1 - \frac{2}{j_2+1}\right)\mu^{j_1,j_2+1}_{r,s,n,-1,1} + \left(\frac{2\theta}{j_2+1}\right)\mu^{j_1,j_2+1}_{r,s-1,n,-1,1}\\ &= \left(\frac{\theta^2}{j_2+2}\right)\left[\mu^{j_1,j_2+2}_{r,s,n,-1,1} - \mu^{j_1,j_2+2}_{r,s-1,n,-1,1}\right].\end{aligned}$$

Proof : In view of (3) and (12), we have

$$\begin{aligned}&\mu^{j_1,j_2}_{r,s,n,\widetilde{m},k} + \theta\mu^{j_1,j_2+1}_{r,s,n,\widetilde{m},k}\\ &= 2\theta c_{s-1}\sum_{t=r+1}^{s}\sum_{i=1}^{r} a_i(r)a_t^{(r)}(s)\int_0^\infty\int_x^\infty x^{j_1}y^{j_2}\left(\frac{\bar F(y)}{\bar F(x)}\right)^{\gamma_t}(\bar F(x))^{\gamma_t}\,dydx\\ &+ \theta^2 c_{s-1}\sum_{t=r+1}^{s}\sum_{i=1}^{r} a_i(r)a_t^{(r)}(s)\int_0^\infty\int_x^\infty x^{j_1}y^{j_2+1}\left(\frac{\bar F(y)}{\bar F(x)}\right)^{\gamma_t}(\bar F(x))^{\gamma_t}\,dydx\end{aligned} \tag{17}$$

Simplifying the above integration (17) by parts, $\left(\frac{\bar F(y)}{\bar F(x)}\right)^{\gamma_t}$ is considered as the parts of the integrand. The theorem is thus proven.

Similarly, in the case of records ($m \to -1,\ k = 1$), and in view of (5) and (12), we have

$$\begin{aligned}\mu^{j_1,j_2}_{r,s,n,-1,1} + \theta\mu^{j_1,j_2+1}_{r,s,n,-1,1} &= \frac{2\theta}{(r-1)!(s-r-1)!}\int_0^\infty\int_x^\infty x^{j_1}y^{j_2}[-\ln\bar F(x)]^{r-1}\\ &\quad\times[-\ln\bar F(y)+\ln\bar F(x)]^{s-r-1}\bar F(y)\,dydx\\ &+ \frac{\theta^2}{(r-1)!(s-r-1)!}\int_0^\infty\int_x^\infty x^{j_1}y^{j_2+1}[-\ln\bar F(x)]^{r-1}\\ &\quad\times[-\ln\bar F(y)+\ln\bar F(x)]^{s-r-1}\bar F(y)\,dydx\,.\end{aligned} \tag{18}$$

Simplifying the integration (18) by parts, we may consider the term $[-\ln\bar F(y) + \ln\bar F(x)]^{s-r-1}$ as representing the parts of the integrand. The theorem is thus proven.

3. Exact moment of the BHE distribution using *gos*

This section presents the closed form expressions for a single moment of the BHE distribution based on *gos*, represented in terms of an exponential integral function.

Theorem 4. For $j \in \mathbb{N}_0$, and $1 \leq r \leq n$. The single moment of the BHE distribution based on gos is expressed as:

$$\mu^{j}_{r,n,\tilde{m},k} = c_{r-1} \sum_{i=1}^{r} \sum_{i_1=0}^{j} (-1)^{i_1+j} a_i(r) \frac{e^{\gamma_i}}{\theta^j} \binom{j}{i_1} \left[E_{\gamma_i - i_1 + 1}(\gamma_i) + E_{\gamma_i - i_1}(\gamma_i) \right], \tag{19}$$

provided that $\gamma_i - i_1 \in \mathbb{N}_0$ and $E_n(x) = \int_1^{\infty} \frac{e^{-xt}}{t^n}\, dt\,, \quad (n = 0, 1, \ldots, \quad \Re(x) > 0)$ (Gradshteyn and Ryzhik, 2014, p. xxxv).

Proof : In light of (2), (10), and (11), we get

$$\mu^{j}_{r,n,\tilde{m},k} = c_{r-1} \sum_{i=1}^{r} a_i(r) \int_0^{\infty} x^j \left[\frac{e^{-\theta x}}{1+\theta x} \right]^{\gamma_i - 1} \theta e^{-\theta x} \frac{(2+\theta x)}{(1+\theta x)^2}\, dx\,.$$

Substituting $1 + \theta x = t \implies \theta dx = dt$, we have

$$\begin{aligned} \mu^{j}_{r,n,\tilde{m},k} &= c_{r-1} \sum_{i=1}^{r} a_i(r) \int_1^{\infty} \left(\frac{t-1}{\theta} \right)^j \left[\frac{e^{-(t-1)}}{t} \right]^{\gamma_i - 1} \frac{e^{-(t-1)}(1+t)}{t^2}\, dt \\ &= c_{r-1} \sum_{i=1}^{r} \sum_{i_1=0}^{j} (-1)^{i_1+j} a_i(r) \frac{e^{\gamma_i}}{\theta^j} \binom{j}{i_1} \left[\int_1^{\infty} \frac{e^{-t\gamma_i}}{t^{\gamma_i - i_1 + 1}}\, dt + \int_1^{\infty} \frac{e^{-t\gamma_i}}{t^{\gamma_i - i_1}}\, dt \right]. \end{aligned} \tag{20}$$

The theorem has been proved using the result of (Gradshteyn and Ryzhik, 2014, p. xxxv).

Corollary 1. In the special case of m-gos, where $m_1 = m_2 = \ldots = m \neq -1$, the single moment of the BHE distribution based on m-gos is expressed as:

$$\begin{aligned} \mu^{j}_{r,n,m,k} = \frac{c_{r-1}}{(m+1)^{r-1}(r-1)!} \sum_{i=0}^{r-1} \sum_{i_1=0}^{j} (-1)^{i+i_1+j} \frac{e^{\gamma_{r-i}}}{\theta^j} \binom{j}{i_1} \\ \times \left[E_{\gamma_{r-i} - i_1 + 1}(\gamma_{r-i}) + E_{\gamma_{r-i} - i_1}(\gamma_{r-i}) \right]. \end{aligned}$$

Remark 1. The single moment of the BHE distribution based on OS ($m = 0,\ k = 1$) is given by

$$\begin{aligned} \mu^{j}_{r,n,0,1} = \frac{c_{r-1}}{(m+1)^{r-1}(r-1)!} \sum_{i=0}^{r-1} \sum_{i_1=0}^{j} (-1)^{i+i_1+j} \frac{e^{(n-r+i+1)}}{\theta^j} \binom{j}{i_1} \\ \times \left[E_{n-r+i-i_1+2}(n-r+i+1) + E_{n-r+i-i_1+1}(n-r+i+1) \right]. \end{aligned}$$

Remark 2. Putting $m_i = R_i \in \mathbb{N}_0,\ k = R_n + 1$ *and* $\gamma_i = n - i + 1 + \sum_{i=r}^{n} R_i$ in (19), then the single moment of the BHE distribution based on progressive Type II censored OS is given by

$$\mu_{r:m:n}^{(R_1,R_2,\ldots,R_n)^{(j)}} = c_{r-1} \sum_{i=1}^{r} \sum_{i_1=0}^{j} (-1)^{i_1+j} a_i(r) \frac{e^{\gamma_i}}{\theta^j} \binom{j}{i_1} \times \left[E_{\gamma_i - i_1 + 1}(\gamma_i) + E_{\gamma_i - i_1}(\gamma_i) \right].$$

4. Maximum Likelihood Estimation

In this section, we have derived the MLE of the parameter of BHE distribution based on *gos*.

Let $X(1,n,\widetilde{m},k), X(2,n,\widetilde{m},k), \ldots, X(n,n,\widetilde{m},k)$ be the n *gos* from the BHE distribution having *pdf* given in (9). To simplify, consider the realization of these n *gos* as $\underline{\mathbf{x}} = (x_1, x_2, \ldots, x_n)$. The likelihood function of the BHE distribution is then expressed as follows:

$$L(\theta|\underline{\mathbf{x}}) = k\theta^n \left(\prod_{j=1}^{n-1} \gamma_j \right) \left[\prod_{i=1}^{n-1} \left\{ \frac{e^{-\theta x_i}}{1+\theta x_i} \right\}^{m_i+1} \right] \left[\prod_{i=1}^{n} \frac{2+\theta x_i}{(1+\theta x_i)} \right] \left[\frac{e^{-\theta x_n}}{(1+\theta x_n)} \right]^k. \tag{21}$$

Thus, the log-likelihood function of the BHE distribution derived from *gos* can be expressed as follows:

$$\begin{aligned} l(\theta|\underline{\mathbf{x}}) = \ln k + n \ln \theta + \sum_{j=1}^{n-1} \ln \gamma_j + \sum_{i=1}^{n-1} (m_i+1) \ln \left\{ \frac{e^{-\theta x_i}}{1+\theta x_i} \right\} \\ + \sum_{i=1}^{n} \ln \left\{ \frac{2+\theta x_i}{(1+\theta x_i)} \right\} + k \ln \left\{ \frac{e^{-\theta x_n}}{1+\theta x_n} \right\}. \end{aligned} \tag{22}$$

Differentiating (22) *w.r.t.* θ and setting it equal to zero, we have

$$\frac{n}{\theta} - \sum_{i=1}^{n-1} \frac{(m_i+1)(2x_i+\theta x_i^2)}{(1+\theta x_i)} - \sum_{i=1}^{n} \frac{x_i}{(2+\theta x_i)(1+\theta x_i)} - \frac{k(2x_n+\theta x_n^2)}{(1+\theta x_n)} = 0. \tag{23}$$

Further, in case of records $(m \to -1,\ k = 1)$, the log-likelihood function based on first n records, $\underline{\mathbf{r}} = (r_1, \ldots, r_n)$ is given by

$$l(\theta|\underline{\mathbf{r}}) = n \ln \theta - \theta r_n - \ln(1+\theta r_n) + \sum_{i=1}^{n} \ln \left(\frac{2+\theta r_i}{1+\theta r_i} \right). \tag{24}$$

Differentiating (24) with respect to θ and setting it equal to zero, we get

$$\frac{n}{\theta} - r_n - \frac{r_n}{1+\theta r_n} - \sum_{i=1}^{n} \frac{r_i}{(1+\theta r_i)(1+2\theta r_i)} = 0 \tag{25}$$

Since Equations (23) and (25) are nonlinear, a direct solution for θ cannot be achieved. Consequently, the R software (R Core Team et al., 2013) and the ***nleqslv*** (Hasselman and Hasselman, 2018) package are employed to compute the MLE of θ. The MLE's convergence is examined using Newton-Raphson technique.

4.1. Asymptotic Confidence Interval

In this subsection, we have derived the asymptotic confidence interval (ACI) for the parameter $\hat{\theta}$ of the BHE distribution based on *gos*.

The observed Fisher information for the BHE distribution is specified as follows:

$$I_0(\hat{\theta}) = -\frac{\partial^2 l(\theta|\underline{\mathbf{x}})}{\partial \theta^2}|_{\theta=\hat{\theta}},$$

The second-order derivative of the log-likelihood function *w.r.t.* the parameter θ can be expressed as

$$\frac{\partial^2 l(\theta|\underline{\mathbf{x}})}{\partial \theta^2} = -\frac{n}{\theta^2} + \sum_{i=1}^{n-1} \frac{(m_i+1)x_i^2}{(1+\theta x_i)^2} + \sum_{i=1}^{n} \frac{(3x_i^2+2\theta x_i^3)}{(1+\theta x_i)^2(2+\theta x_i)^2} + \frac{kx_n^2}{(1+\theta x_n)^2}.$$

By utilizing the Fisher information, the asymptotic variance of the MLE of the parameter for the BHE distribution can be obtained as

$$Var_0(\hat{\theta}) = \frac{1}{I_0(\hat{\theta})}.$$

Under certain regularity conditions, the sampling distribution of $\frac{(\hat{\theta}-\theta)}{\sqrt{Var_0(\hat{\theta})}}$ can be approximated utilizing a standard normal distribution. The large sample CI also referred to as the ACI for θ, is given by $[\hat{\theta}_L, \hat{\theta}_U] = \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{Var_0(\hat{\theta})}$.

In the case of the lower bound, $\hat{\theta}_L$ can be negative. In such cases, $\hat{\theta}_L$ is determined by $\hat{\theta}_L = max\,(0, \hat{\theta}_L)$. We can also use the simulation to obtain the average length (AL) and the coverage probability (CP).

5. Bayesian Estimation

The prior information for the parameter of the BHE distribution is taken into account. This informative prior has a two-parameter gamma distribution. Therefore, the *pdf* of the prior distribution of the parameter is given by

$$\pi(\theta) = \frac{b_1^{a_1}}{\Gamma a_1} \theta^{a_1-1} e^{-b_1\theta}, \quad a_1, b_1, \theta > 0. \tag{26}$$

We have used two loss functions: the squared error (SE) loss function (symmetric) and the linear exponential (LINEX) loss function (asymmetric). The LINEX loss function combines linear and exponential components to model asymmetric penalties for over- and underestimation in the statistical decision theory and risk analysis. $L_2(\delta,\Theta) = \exp(v(\delta - \Theta)) - v(\delta - \Theta) - 1$, where δ is the true value, Θ is the estimate, and v controls for asymmetry, which penalizes overestimation more heavily as v increases. This function is useful in scenarios with unequal consequences for over- and underestimation, such as inventory management and financial risk assessment, and has applications in Bayesian estimation, forecasting, and decision-making under uncertainty. These loss functions add flexibility to our results and provide applicability in many real-world scenarios. The symmetric loss function, or SE loss function, is selected for its ability to penalize both underestimation and overestimation equally, which is beneficial in most cases. Nevertheless, in several scenarios, a positive loss can be more serious than a negative loss and vice versa. In such cases, the use of an asymmetric loss function is recommended. The present study examined two distinct loss functions: symmetrical (SE) and asymmetrical (LINEX).

The mathematical representation of these loss functions and the associated Bayesian estimators is as follows:

$$L_1(\delta,\theta) = (\delta - \theta)^2, \quad \theta > 0.$$

In the case of the SE loss function, the Bayes estimator is equal to the posterior mean represented by (δ_{SE}). The definition of the LINEX loss function is as follows:

$$L_2(\delta,\theta) = \exp(v(\delta - \theta)) - v(\delta - \theta) - 1, \quad v \neq 0$$

and the corresponding Bayes estimator as

$$\delta_{LINEX} = -\frac{1}{v}\ln(E(\exp(-v\theta))).$$

In view of (21) and (26), the posterior density function of θ given $\underline{\mathbf{x}}$ is given by

$$\pi(\theta|\underline{\mathbf{x}}) = \frac{1}{A}P^{\star}(\theta;\underline{\mathbf{x}}),$$

where

$$P^{\star}(\theta;\underline{\mathbf{x}}) = \theta^{n+a_1-1}e^{-b_1\theta}\left[\prod_{i=1}^{n-1}\left\{\frac{e^{-\theta x_i}}{1+\theta x_i}\right\}^{m_i+1}\right]\left[\prod_{i=1}^{n}\frac{2+\theta x_i}{(1+\theta x_i)}\right]\left[\frac{e^{-\theta x_n}}{(1+\theta x_n)}\right]^{k}.$$

In case of records ($m \to -1,\ k = 1$),

$$P^{\star}(\theta;\underline{\mathbf{r}}) = \theta^{n+a_1-1}\frac{e^{-\theta(b_1+r_n)}}{(1+\theta r_n)}\prod_{i=1}^{n}\left(\frac{2+\theta r_i}{1+\theta r_i}\right)$$

and $A = \int_0^{\infty} P^{\star}(\theta;\underline{\mathbf{x}})\,d\theta$. Then, under the SE loss function, the Bayes estimate of any function of the parameter, say $g(\theta)$, becomes

$$\hat{g}_{SE} = E(g(\theta)|\underline{\mathbf{x}}) = \frac{1}{A}\int_0^\infty g(\theta)P^\star(\theta;\underline{\mathbf{x}})\,d\theta\ . \tag{27}$$

The aforesaid equation is the ratio of two integrals, and its explicit solution is not available. Therefore, we adopt the Metropolis- Hastings (M-H) algorithm to compute the approximate Bayes estimate of the parameter of the BHE distribution.

5.1. Metropolis- Hastings Algorithm

In this subsection, random samples are generated from the posterior density using the M-H algorithm. We used these generated samples to get the parameter's Bayes estimator and Bayesian interval estimates. The modified M-H algorithm is used here. For example, see Dey and Pradhan (2014). Let $N(\theta,\sigma_\theta)$ denote the normal distribution. The expression for M-H algorithm is presented as follows:

1. Set an initial value of θ_0.
2. For $j = 1,\ldots,N$, repeat the following steps:

- Set $\theta = \theta_{j-1}$.
- Generate a candidate value η from $N(\theta^\star,\sigma_\theta)$, where $\theta^\star = \ln(\theta)$ and generate u from *U(0,1)*.
- Set $\theta^{'} = \exp(\eta)$.
- Calculate $\delta = min\left\{1, \frac{\pi(\theta^{'}|\underline{\mathbf{x}})}{\pi(\theta|\underline{\mathbf{x}})}\right\}$.
- If $u \leq \delta$, then set $\theta_j = \theta^{'}$ otherwise set $\theta_j = \theta_{j-1}$.

The σ_θ selection is a significant aspect of the acceptance rate. The initial B (burn-in period) generated samples may be discarded.

Let us now assume that the set of samples generated by the M-H mentioned above algorithm is $\{\theta_l, l = 1,\ldots,L\}$, where $L = N^{'} - B$). Then, under the SE and LINEX loss functions, the estimated Bayes estimate of parameter of the BHE distribution can be determined by $\hat{\theta}_{SE} = \frac{1}{L}\sum_{l=1}^{L}\theta_l$, and $\hat{\theta}_{SE} = \frac{1}{L}\left(\sum_{l=1}^{L}\exp(-v\theta_l)\right)$, respectively.

6. Generation of Random Samples

In this section, we have discussed two methods of generating random samples from BHE distribution based on *gos*.

Algorithm 1: Generation algorithm using the result of Cramer (2003, p. 30). The *gos* can alternatively be defined as

$X(r,n,\widetilde{m},k) = F^{-1}\left(1-\prod_{j=1}^{r}B_j\right) = \bar{F}^{-1}\left(\prod_{j=1}^{r}B_j\right)$, where $B_1,B_2,\ldots,B_r$ are independent power function distributed *rv's* with distribution function $P(B_j \leq x) = x^{\gamma_j},\ x \in (0,1)$, and $1 \leq j < n$.

The result above can be utilized to develop an efficient algorithm for the generation of *gos* random samples based on the BHE distribution as follows:

1. Generate n independent and identically distributed (*iid*) *rv's* uniform random samples, $U_1, U_2, , \ldots, U_n$.

2. Transform $Y_j = U_j^{\frac{1}{\gamma_j}};\ j = 1,2,\ldots,n$. This implies that Y_j follows $B(\gamma_j, 1)$ or power function distribution with exponent γ_j.

   Since
   $$x = \frac{1}{\theta}\left(W_0\left(\frac{e}{1-u}\right) - 1\right),$$
   where, W_0 denotes the principal branch of lambert W function.

3. Compute $U(r,n,\widetilde{m},k) = 1 - \prod_{i=1}^{r} Y_i;\ r = 1,2,\ldots,n$.
   Since $X(r,n,\widetilde{m},k) = F^{-1}\left(1 - \prod_{i=1}^{r} Y_i\right)$, using the quantile function of the BHE distribution, we get
   $$X(r,n,\widetilde{m},k) = \frac{1}{\theta}\left(W_0\left(\frac{e}{\prod_{i=1}^{r} Y_i}\right) - 1\right).$$

Algorithm 2: Let $X^{\star}(r,n,\widetilde{m},k);\ r = 1,2,\ldots,n$ be the *gos* based on the *cdf* $F(x) = 1 - e^{-x}$ i.e., $exp(1)$. Using the normalized spacing of *gos*, (Kamps, 1995, p. 81). $X^{\star}(r,n,\widetilde{m},k) = \sum_{i=1}^{r} \frac{Y_i}{\gamma_i}$, where $Y_1, Y_2, \ldots, Y_n$ are *iid rv's* follow exp(1). Now adopting the technique of Arnold et al. (1992), which is as follows: Let $X^{\star}(r,n,\widetilde{m},k);\ r = 1,2,\ldots,n$ be the *gos* based on continuous *cdf* F then $H(X) = -\ln[1 - F(X)]$ has standard exponential distribution and $H(X)$ is monotonic function of X and $X^{\star}(r,n,\widetilde{m},k)$ be the *gos* based on exp(1) then

$$\begin{aligned} X(r,n,\widetilde{m},k) =& F^{-1}(1 - e^{-X^{\star}(r,n,\widetilde{m},k)}) \\ =& F^{-1}\left(1 - e^{-\Sigma_{i=1}^{r} \frac{Y_i}{\gamma_i}}\right) \\ =& \frac{1}{\theta}\left[W_0\left(e^{1+\Sigma_{i=1}^{n} \frac{Y_i}{\gamma_i}}\right) - 1\right]. \end{aligned} \tag{28}$$

Putting $m_i = 0$ and $k = 1$ i.e., $\gamma_i = n - i + 1$ in the aforesaid algorithm, we shall get a random sample of n OS based on the BHE distribution. Similarly, if we put $m \to -1,\ k \in \mathbb{N}$ i.e., $\gamma_i = k$ in the above-discussed algorithm, we shall generate the first n k^{th} records of BHE distribution. Put $m_i = R_i \in \mathbb{N}_0,\ k = R_n + 1$ *and* $\gamma_i = n - i + 1 + \sum_{i=r}^{n} R_i$ in the above algorithm described above, we shall get a random sample for progressive Type II censored OS. Similarly, if we put $k = \beta_n,\ m_i = (n - i + 1)\alpha_i - (n - i)\alpha_{i+1} - 1$ *and* $\gamma_i = (n - i + 1)\beta_i; \beta_1, \ldots, \beta_n > 0$ in the above-discussed algorithm, we shall get a random sample of sequential OS.

6.1. Bootstrap Method

In this subsection we have discussed the percentile bootstrap method, denoted boot-p, which is used to construct an approximate confidence interval (CI) for θ using the following algorithm.

Step 1: First generate *n-gos* from the BHE distribution, then calculate the MLE of the parameter of the BHE distribution, say $\hat{\theta}$.

Step 2: Utilizing $\hat{\theta}$ derived from Step 1, generate a random sample of *n-gos* drawn from the BHE distribution, referred to as a bootstrap sample.

Step 3: Based on the bootstrap sample that is obtained in Step 2, compute the corresponding MLE $\hat{\theta}^{\star}$ of θ.

Step 4: Repeat Steps (2) and (3) B-times in order to obtain $\{\theta_1^{\star}, \theta_2^{\star}, \ldots, \theta_B^{\star}\}$.

Step 5 Arrange $\{\theta_1^{\star}, \theta_2^{\star}, \ldots, \theta_B^{\star}\}$ in ascending order and obtain $\{\theta_{(1)}^{\star}, \theta_{(2)}^{\star}, \ldots, \theta_{(B)}^{\star}\}$.

Step 6 The approximate $100(1-\alpha)\%$ boot-p CI for θ is defined as: $\left(\theta^{\star}_{(B\frac{\alpha}{2})}, \theta^{\star}_{(B(1-\frac{\alpha}{2}))}\right)$.

6.2. Simulation Study Based on MLEs

The performance of the estimator can be measured by the mean squared error (MSE) of the MLEs of the scale parameter θ of the BHE distribution using OS (records). Here are the steps to calculate the MSEs:

1. Here, first generate n OS (records) from the BHE distribution.

2. The MLEs of the BHE distribution's parameters are calculated using the OS (records) generated in Step 1.

3. Total 6000 iterations, are used to calculate the MLEs of the distribution parameters. Each iteration of the BHE distribution yields a new set of OS (records) from which the MLEs are calculated.

The MSE values are calculated considering numerous n values for OS (records) configurations. Furthermore, the BHE distribution's parameter value is specified for fixed value of θ. The MSEs can be seen in Figures 3-4 and shown in the graph. The most significant conclusion from Figures 3-4 is that the MSE values decrease as the sample size (n) increases. This pattern signifies that higher OS (records) improve the accuracy and precision of the estimators. The MSE decreases with increase in the sample size, suggesting that the ML estimators are getting closer to the true values of the parameter θ.

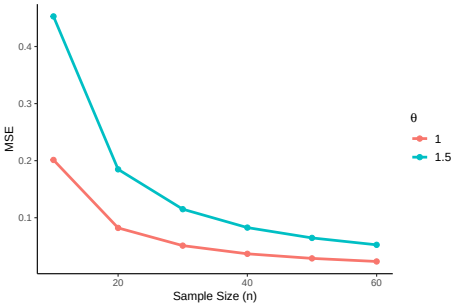

Figure 3. The MSE plot for $\theta = 1$ and $\theta = 1.5$ based on OS

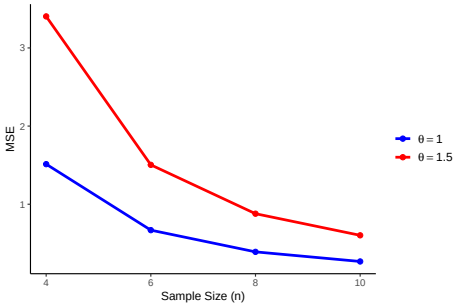

Figure 4. The MSE plot for $\theta = 1$ and $\theta = 1.5$ based on record values

6.3. Simulation Study Based on MCMC

The following subsection presents the results of a simulation study that demonstrates the performance of the estimators obtained in this manuscript. To achieve this, the method proposed by Wang and Shi (2013) is used to generate several sample sizes of upper record values from the BHE distribution. To evaluate the performance of the estimator, the MSE of the MLE of the parameter of the BHE distribution is assessed in the context of classical estimation. We conducted a Monte Carlo study for 6000 replications in the context of Bayesian analysis and present the results in Table 5. We employ the following algorithm for the calculation, as illustrated in the accompanying tables.

1. First we generate samples of upper record values from the BHE distribution discussed in section 6 (algorithm 1-2).

2. Evaluate the Bayes estimators of the unknown parameter of the BHE distribution under a symmetric loss function, i.e., the SE loss function, and a symmetric loss function, i.e., the LINEX loss function, using the generated samples and the method discussed in Section 6.

3. Repeat the above steps $N = 6000$ times and $K = 1000$ times (burn-in samples) to obtain the Bayes estimators of the parameter of the BHE distribution.

4. Now use the following expression to calculate the MSE

   $MSE(\hat{\theta}) = \frac{1}{N-K}\sum_{i=1}^{N-K}(\hat{\theta}_i - \theta)^2$, where θ is the true value of the parameter.

The simulation studies are carried out using ***R-software*** with 6000 iterations. The proposed MLE and Bayesian estimates of the MSE are calculated. The results of the point estimation are summarized in Table 3 with true parameter values set at $\theta = 1$ and $\theta = 1.5$, respectively, for sample sizes $n = 4, 6$, and 10. Additionally, the performances of the proposed classical CIs and Bayesian CrIs were assessed based on the AL and CP. The ALs and CPs for the 90%, 95%, and 99% ACIs, boot-p CIs, and CrIs for θ are presented in Table 4, for sample sizes $n = 4, 6, 8$, and 10.

The Bayes estimates perform better than the MLE in terms of MSE, as shown in Table 3. Furthermore, it can be observed that the informative Bayes estimates under both the SE loss and LINEX loss functions perform better than the MLE in terms of MSEs. Clearly, the MSEs of the proposed estimates decrease as n increases.

Regarding interval estimation, Table 4 present the ALs and CPs of the ACIs, boot-p CIs, and Bayesian CrIs for $\theta = 1.5$. Evidence suggests that informative Bayesian CrIs outperform both ACI and boot-p CIs in terms of the AL and CP. Notably, the CPs of Bayesian CrIs typically align well with their nominal levels, thus offering a dependable and robust inference method. The data generally indicate that informative CrIs are the most valid approach, as they yield the highest simulated CPs compared with CIs derived from classical methods.

Table 1. Means and variances of *m-gos* and OS from the BHE distribution ($\theta = 3$, $k = 2$, $m = 1$)

		m-gos										OS									
n		$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	$r=8$	$r=9$	$r=10$	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$	$r=8$	$r=9$	$r=10$
1	Mean	0.0924										0.1988									
	Variance	0.0102										0.0502									
2	Mean	0.0441	0.1408									0.0924	0.3051								
	Variance	0.0022	0.0134									0.0102	0.0676								
3	Mean	0.0289	0.0745	0.1739								0.0598	0.1578	0.3788							
	Variance	0.0009	0.0032	0.0152								0.0040	0.0158	0.0772							
4	Mean	0.0215	0.0512	0.0979	0.1992							0.0441	0.1069	0.2087	0.4355						
	Variance	0.0004	0.0015	0.0040	0.0163							0.0022	0.0069	0.0196	0.0836						
5	Mean	0.0171	0.0390	0.0693	0.1170	0.2198						0.0349	0.0809	0.1458	0.2505	0.4818					
	Variance	0.0003	0.0009	0.0019	0.0045	0.0172						0.0013	0.0038	0.0089	0.0224	0.0882					
6	Mean	0.0142	0.0316	0.0539	0.0848	0.1331	0.2371					0.0289	0.0651	0.1125	0.1792	0.2862	0.5209				
	Variance	0.0002	0.0005	0.0011	0.0022	0.0049	0.0179					0.0009	0.0024	0.0050	0.0105	0.0246	0.0918				
7	Mean	0.0121	0.0265	0.0442	0.0669	0.0981	0.1471	0.2522				0.0246	0.0544	0.0916	0.1402	0.2084	0.3173	0.5548			
	Variance	0.0002	0.0004	0.0007	0.0013	0.0024	0.0053	0.0183				0.0007	0.0016	0.0033	0.0061	0.0118	0.0263	0.0946			
8	Mean	0.0106	0.0229	0.0375	0.0554	0.0784	0.1100	0.1595	0.2654			0.0215	0.0468	0.0774	0.1154	0.1651	0.2345	0.3449	0.5848		
	Variance	0.0001	0.0003	0.0005	0.0008	0.0015	0.0026	0.0055	0.0188			0.0004	0.0012	0.0023	0.0040	0.0069	0.0128	0.0277	0.0969		
9	Mean	0.0094	0.0201	0.0326	0.0473	0.0655	0.0887	0.1206	0.1706	0.2773		0.0190	0.0410	0.0669	0.0982	0.1370	0.1875	0.2580	0.3697	0.6117	
	Variance	0.0001	0.0002	0.0003	0.0007	0.0010	0.0016	0.0028	0.0057	0.0191		0.0003	0.0009	0.0017	0.0029	0.0046	0.0077	0.0137	0.0289	0.0989	
10	Mean	0.0084	0.0179	0.0288	0.0414	0.0563	0.0746	0.0981	0.1303	0.1806	0.2880	0.0171	0.0365	0.0590	0.0855	0.1173	0.1568	0.2080	0.2794	0.3923	0.6360
	Variance	0.0000	0.0002	0.0003	0.0005	0.0007	0.0011	0.0017	0.0029	0.0059	0.0195	0.0003	0.0008	0.0013	0.0021	0.0032	0.0051	0.0083	0.0145	0.0300	0.1007

Table 2. Means and variances of the BHE distribution based on progressive Type II censored OS ($\theta = 3$)

N	n	$(R_1,\ldots,R_n)$	$r=1$		$r=2$		$r=3$		$r=4$		$r=5$		$r=6$	
			$E(X)$	$Var(X)$	$E(X)$	$Var(X)$	$E(X)$	$Var(X)$	$E(X)$	$Var(X)$	$E(X)$	$Var(X)$	$E(X)$	$Var(X)$
8	3	(0,1,4)	0.6452	0.0005	0.6199	0.0012	0.5830	0.0027	-	-	-	-	-	-
		(3,0,2)	0.6452	0.0005	0.5999	0.0028	0.5357	0.0077	-	-	-	-	-	-
		(4,0,1)	0.6452	0.0005	0.5839	0.0048	0.4839	0.0170	-	-	-	-	-	-
		(2,0,3)	0.6452	0.0005	0.6093	0.0020	0.5623	0.0045	-	-	-	-	-	-
		(2,1,2)	0.6452	0.0005	0.6093	0.0020	0.5457	0.0067	-	-	-	-	-	-
		(1,2,2)	0.6452	0.0005	0.6155	0.0015	0.5522	0.0062	-	-	-	-	-	-
12	4	(3,1,1,3)	0.6525	0.0002	0.6306	0.0007	0.6004	0.0017	0.5529	0.0044	-	-	-	-
		(4,2,1,1)	0.6525	0.0002	0.6274	0.0009	0.5812	0.0033	0.4809	0.0153	-	-	-	-
		(1,2,1,4)	0.6525	0.0002	0.6351	0.0005	0.6095	0.0012	0.5721	0.0029	-	-	-	-
		(5,3,0,0)	0.6525	0.0002	0.6231	0.0011	0.5263	0.0122	0.3071	0.0723	-	-	-	-
		(2,3,1,2)	0.6525	0.0002	0.6331	0.0006	0.5967	0.0021	0.5323	0.0070	-	-	-	-
		(1,2,2,3)	0.6525	0.0002	0.6351	0.0005	0.6095	0.0012	0.5624	0.0037	-	-	-	-
16	6	(0,1,7,0,1,1)	0.6554	0.0001	0.6446	0.0003	0.6312	0.0004	0.5947	0.0019	0.5469	0.0046	0.4439	0.0173
		(0,2,0,1,4,3)	0.6554	0.0001	0.6446	0.0003	0.6300	0.0005	0.6138	0.0007	0.5935	0.0012	0.5456	0.0038
		(3,2,1,3,0,1)	0.6554	0.0001	0.6417	0.0003	0.6221	0.0007	0.5960	0.0016	0.5315	0.0063	0.4274	0.0194
		(1,2,0,5,1,1)	0.6554	0.0001	0.6438	0.0002	0.6278	0.0006	0.6099	0.0009	0.5628	0.0034	0.4609	0.0158
		(1,3,4,0,1,1)	0.6554	0.0001	0.6438	0.0002	0.6262	0.0006	0.5895	0.0021	0.5415	0.0048	0.4381	0.0176
		(0,5,1,2,2,0)	0.6554	0.0001	0.6446	0.0003	0.6250	0.0007	0.5991	0.0014	0.5515	0.0040	0.3347	0.0615
		(1,3,4,0,2,0)	0.6554	0.0001	0.6438	0.0002	0.6262	0.0006	0.5895	0.0021	0.5415	0.0048	0.3234	0.0629

In this aspect, the means and variances for a sample size of $n = 1:1:10$ are figured out in Table 1 for both OSs and *m-gos*, respectively. For progressive Type II censored OS and different choices of $(R_1, R_2, \ldots, R_n)$, a pre-specified fixed number of failures (n) is taken to be $3, 4,$ *and* 6. For $r = 1:1:6$ and $N = 8:4:16$, we obtained the means of the progressive Type II censored OS as in Table 2.

For fixed sample size n, increasing r increases the mean and the variance. For any specific value of r, the increase in n causes a decrease in the variance; this is true universally. Similarly, for the progressive censored OS of Type II, the increase in r increases the mean for fixed N and n. However, there is no definite pattern for a fixed r and varying N and n. The means depend on the selection of $(R_1, R_2, \ldots, R_n)$. Similarly, we may interpret in case of the product moments of *m-gos*, and OS.

Table 3. MSEs of MLE and Bayes estimators of θ for different θ values (rep = 6000, burn-in = 1000, $a_1 = 1$, $b_1 = 1.5$)

	$\theta = 1$ (MCMC)					$\theta = 1.5$ (MCMC)				
n	MSE(MLE)	SE loss	LINEX loss			MSE(MLE)	SE loss	LINEX loss		
			$v = -0.5$	$v = 1$	$v = 1.5$			$v = -0.5$	$v = 1$	$v = 1.5$
4	1.5128	0.3006	0.2198	0.2656	0.2792	3.4034	0.7605	0.5830	0.7265	0.7667
6	0.6685	0.1960	0.1064	0.1455	0.1578	1.5038	0.5106	0.3194	0.4473	0.4846
8	0.3906	0.1306	0.0301	0.0573	0.0666	0.8788	0.3165	0.1233	0.2179	0.2484
10	0.2677	0.0964	0.0129	0.0292	0.0354	0.6022	0.2513	0.0567	0.1305	0.1564

Table 4. Summary of the ALs and CPs of 90%, 95%, and 99% ACIs, boot-p CIs, and Bayesian CrIs of $\theta = 1.5$ ($a_1 = 1$, $b_1 = 1.5$)

	90%						**95%**						**99%**					
n	ACI		boot-p		CrI		ACI		boot-p		CrI		ACI		boot-p		CrI	
	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP	AL	CP
4	4.2268	0.93	6.1820	0.85	1.9780	0.91	5.0365	0.96	8.7556	0.89	2.3832	0.96	6.6191	0.98	14.2162	0.96	3.1999	0.99
6	2.9870	0.92	3.9340	0.87	1.8223	0.91	3.5593	0.95	5.2419	0.91	2.1868	0.96	4.6776	0.99	8.5819	0.95	2.9102	0.99
8	2.4275	0.91	3.0280	0.86	1.6832	0.91	2.8926	0.96	3.8772	0.90	2.0153	0.96	3.8015	0.99	5.8542	0.97	2.6703	0.99
10	2.0623	0.91	2.5026	0.88	1.5573	0.90	2.4573	0.96	3.0749	0.93	1.8587	0.96	3.2295	0.99	4.5861	0.97	2.4445	0.99

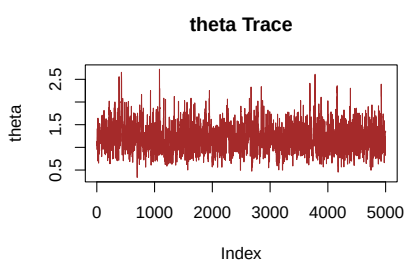

Figure 5. Trace plot for θ

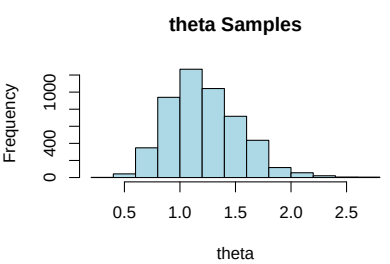

Figure 6. Histogram for θ

7. Application to the Real Data

In this section, we analyze two real data sets to illustrate the application of our model. The first data set was originally reported by Caramanis et al. (1983) and later analyzed by Altun and Hamedani (2018) and Alqasem et al. (2014), while the second data set was studied by Kabdwal et al. (2024). The first data set consists of 23 sample values measuring unit capacity factors using the "SC16" algorithm. The second data set comprises failure times of 23 secondary reactor pumps. The data sets are as follows:

Data set I: 0.853, 0.759, 0.866, 0.809, 0.717, 0.544, 0.492, 0.403, 0.344, 0.213,0.116, 0.116, 0.092, 0.070, 0.059, 0.048, 0.036, 0.029, 0.021, 0.014, 0.011, 0.008, 0.006.

Data set II: 2.160, 0.150, 4.082, 0.746, 0.358, 0.199, 0.402, 0.101, 0.605, 0.954, 1.359, 0.273, 0.491, 3.465, 0.070, 6.560, 1.060, 0.062, 4.992, 0.614, 5.320, 0.347, 1.921.

The BHE distribution is fitted to both data sets. For data set I, the Kolmogorov-Smirnov (K-S) distance and associated p-value at a fixed value of $\theta = 2.5$ were 0.26087 and 0.4218, respectively. For data set II, the K-S distance and p-value at a fixed value of $\theta = 1$ were 0.17391 and 0.888, respectively. These results suggest that the BHE distribution provides an appropriate fit for both data sets. Furthermore, three upper record values were identified in data set-2: 2.160, 4.082, and 6.560.

Using these records, the MLEs, 95% ACIs, Bayes estimates, and their corresponding 95% Bayesian CrIs were calculated for parameter θ. Table 5 show the sensitivity of the Bayesian estimate, providing a summary of the point and interval estimates derived from both classical and Bayesian methodologies, respectively.

Table 5. Estimates and 95% Confidence/Credible Intervals of θ based on the real datasets

Dataset	MLE	Bayes Estimates (MCMC)				95% Confidence Intervals for θ	
		SE	LINEX			ACI	CrI
			$v = -0.5$	$v = 1$	$v = 1.5$		
Data set I	1.3056	0.9588	1.0577	0.8217	0.7710	(-0.4981, 3.1092)	(0.1793, 2.4419)
Data set II	0.2843	0.3570	0.3666	0.3393	0.3312	(-0.0964, 0.6651)	(0.0844, 0.8162)

8. Conclusions

The manuscript presents the recursive relations for the single and product moments of the BHE distribution based on *gos* and records have been derived. Further, the closed form expression for the single moment of the BHE distribution based on *gos* is derived. Moreover, the specific case of *gos*, including OS and progressive Type II censored OS, is addressed to derive the explicit expression for the single moment of the BHE distribution. Based on *gos* and their submodels like OS, and progressive Type II censored OS, the BHE distribution's mean and variance are computed. The MLE of the BHE distribution's parameter are derived based on *gos*. The Bayes estimator of the parameter of the BHE distribution is derived using the Metropolis-Hastings algorithm. Two methods are discussed for the generation of the

random sample from the BHE distribution based on *gos*. The simulation study demonstrates the validity of these findings. Finally, the results are verified using two real data sets.

Acknowledgements

The author would like to thank the Chief Editor and the anonymous reviewers for their fruitful suggestions, which helped us to improve the manuscript.

References

Abouelmagd, T. H. M., (2018). The Logarithmic Burr-Hatke Exponential Distribution for Modeling Reliability and Medical Data. *International Journal of Statistics and Probability*, 7(5), pp. 73–85.

Afify, A. Z., Aljohani, H. M., Alghamdi, A. S., Gemeay, A. M. and Sarg, A. M., (2021). A New Two-Parameter Burr-Hatke Distribution: Properties and Bayesian and Non-Bayesian Inference with Applications. *Journal of Mathematics*, 2021(1), pp. 1061083.

Alawady, M. A., Barakat, H. M., Xiong, S. and Abd Elgawad, M. A., (2022). Concomitants of Generalized Order Statistics from Iterated Farlie-Gumbel-Morgenstern Type Bivariate Distribution. *Communications in Statistics-Theory and Methods*, 51(16), pp. 5488–5504.

Almuqrin, M. A., (2023). A New Flexible Distribution with Applications to Engineering Data. *Alexandria Engineering Journal*, 69, pp. 371–382.

Alqasem, O. A., Mohammed, H. S., Elshahhat, A., Hussein, H. A. and Jaheen, Z. F., (2024). Bayesian and Empirical-Bayesian Evaluations of the Topp-Leone Lifetime Model Using Lower Record Statistics. *AIP Advances*, 14(4).

Altun, E. and Hamedani, G. G., (2018). The Log-Xgamma Distribution with Inference and Application. *Journal de la Société Française de Statistique*, 159(3), pp. 40–55.

Aly, A. E., (2023). Predictive Inference of Dual Generalized Order Statistics from the Inverse Weibull Distribution. *Statistical Papers*, 64(1), pp. 139–160.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N., (1992). A First Course in Order Statistics. SIAM.

Arshad, M., Azhad, Q. J., Gupta, N. and Pathak, A. K., (2023). Bayesian Inference of Unit-Gompertz Distribution Based on Dual Generalized Order Statistics. *Communications in Statistics-Simulation and Computation*, 52(8), pp. 3657–3675.

Barco, K. V. P., Mazucheli, J. and Janeiro, V., (2017). The Inverse Power Lindley Distribution. *Communications in Statistics-Simulation and Computation*, 46(8), pp. 6308–6323.

Bhatti, F. A., Hamedani, G. G., Najibi, S. M. and Ahmad, M., (2021). On the Extended Chen Distribution: Development, Properties, Characterizations and Applications. *Annals of Data Science*, 8, pp. 159–180.

Caramanis, M., Stremel, J., Fleck, W. and Daniel, S., (1983). Probabilistic Production Costing: An Investigation of Alternative Algorithms. *International Journal of Electrical Power & Energy Systems*, 5(2), pp. 75–86.

Cramer, E., (2003). Contributions to Generalized Order Statistics. PhD Thesis.

Dey, S., Pradhan, B., (2014). Generalized Inverted Exponential Distribution under Hybrid Censoring. *Statistical Methodology*, 18, pp. 101–114.

El-Morshedy, M., Eliwa, M. S. and Altun, E., (2020). Discrete Burr-Hatke Distribution with Properties, Estimation Methods and Regression Model. *IEEE Access*, 8, pp. 74359–74370.

El-Morshedy, M., Aljohani, H. M., Eliwa, M. S., Nassar, M., Shakhatreh, M. K. and Afify, A. Z., (2021). The Exponentiated Burr-Hatke Distribution and Its Discrete Version: Reliability Properties with CSALT Model, Inference and Applications. *Mathematics*, 9(18), pp. 1–26.

Gradshteyn, I. S. and Ryzhik, I. M., (2014). Table of Integrals, Series, and Products. Academic Press.

Hasselman, B., (2018). Package 'nleqslv'. *R Package Version*, 3(2).

Jodrá, P., (2010). Computer Generation of Random Variables with Lindley or Poisson-Lindley Distribution via the Lambert W Function. *Mathematics and Computers in Simulation*, 81(4), pp. 851–859.

Jodrá, P., Arshad, M., (2022). An Intermediate Muth Distribution with Increasing Failure Rate. *Communications in Statistics-Theory and Methods*, 51(23), pp. 8310–8327.

Kabdwal, N. C., Azhad, Q. J. and Hora, R., (2024). Statistical Inference of the Exponentiated Exponential Distribution Based on Progressive Type II Censoring with Optimal Scheme. *International Journal of System Assurance Engineering and Management*, 15(8), pp. 3833–3853.

Kamps, U., (1995). A Concept of Generalized Order Statistics. *Journal of Statistical Planning and Inference*, 48(1), pp. 1–23.

Kamps, U., Cramer, E., (2001). On Distributions of Generalized Order Statistics. *Statistics*, 35(3), pp. 269–280.

Khan, A. H., Khan, M. J. S., (2012). On Ratio and Inverse Moment of Generalized Order Statistics from Burr Distribution. *Pakistan Journal of Statistics*, 28(1), pp. 59–68.

Khan, M. J. S., Ansari, F., Azhad, Q. J. and Kabdwal, N. C., (2024). Moments and Inferences of Inverted Topp-Leone Distribution Based on Record Values. *International Journal of System Assurance Engineering and Management*, 15(6), pp. 2623–2633.

Khan, M. J. S., Iqrar, S., (2019). On Moments of Dual Generalized Order Statistics from Topp-Leone Distribution. *Communications in Statistics-Theory and Methods*, 48(3), pp. 479–492.

Khan, M. J. S., Khatoon, B., (2020). Statistical Inferences of $R = P(X < Y)$ for Exponential Distribution Based on Generalized Order Statistics. *Annals of Data Science*, 7(3), pp. 525–545.

Khan, M. J. S., Sharma, A. and Iqrar, S., (2020). On Moments of Lindley Distribution Based on Generalized Order Statistics. *American Journal of Mathematical and Management Sciences*, 39(3), pp. 214–233.

Khan, M. J. S., Sharma, A. and Sharma, A., (2018). Shannon Entropy and Characterization of Nadarajah and Haghighi Distribution Based on Generalized Order Statistics. *Journal of Statistics: Advances in Theory and Applications*, 19(1), pp. 43–69.

Kumar, D., (2017). Relations of Dagum Distribution Based on Dual Generalized Order Statistics. *Journal of Applied Mathematics & Informatics*, 35(5-6), pp. 477–493.

Pandey, A., Singh, R. P. and Tyagi, A., (2022). An Inferential Study of Discrete Burr-Hatke Exponential Distribution under Complete and Censored Data. *Reliability: Theory & Applications*, 17(4), pp. 109–122.

R Core Team, (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Ragab, I. E., Khan, S., Alsadat, N., Jamal, F., Hamedani, G. G. and Elgarhy, M., (2024). Modeling to COVID-19 and Cancer Data: Using the Generalized Burr-Hatke Model. *Journal of Radiation Research and Applied Sciences*, 17(3), pp. 1–18.

Shirozhan, M., Mamode Khan, N. A. and Bakouch, H. S., (2023). An INAR (1) Time Series Model via a Modified Discrete Burr-Hatke with Medical Applications. *Iranian Journal of Science*, 47(1), pp. 121–136.

Wang, L., Shi, Y., (2013). Reliability Analysis of a Class of Exponential Distribution under Record Values. *Journal of Computational and Applied Mathematics*, 239, pp. 367–379.

Yadav, A. S., Altun, E. and Yousof, H. M., (2021). Burr-Hatke Exponential Distribution: A Decreasing Failure Rate Model, Statistical Inference and Applications. *Annals of Data Science*, 8(2), pp. 241–260.

Yousof, H. M., Altun, E., Ramires, T. G., Alizadeh, M. and Rasekhi, M., (2018). A New Family of Distributions with Properties, Regression Models and Applications. *Journal of Statistics and Management Systems*, 21(1), pp. 163–188.

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 229–241, https://doi.org/10.59139/stattrans-2026-035
Received – 23.09.2025; accepted – 09.04.2026

Rethinking equivalence scales in the measurement of extreme poverty: evidence from Poland

Hanna Dudek[1]

Abstract

This paper examines the measurement of extreme poverty in Poland, focusing on the role of equivalence scales. The key research question is whether the original OECD equivalence scale, currently used by Statistics Poland, is consistent with the scale effects implied by the construction of the subsistence minimum across different household types. Using a one-parameter power equivalence scale, extreme poverty rates are recalculated under alternative assumptions based on anonymized microdata from the 2023 Household Budget Survey. The results indicate that the original OECD scale shows stronger economies than the one embedded in the subsistence minimum values, leading to substantial differences in the estimated extreme poverty rate. These findings underline the importance of a careful selection of equivalence scales in official poverty statistics, as methodological choices directly influence poverty assessment and the targeting of social policy.

Key words: equivalence scale, extreme poverty, subsistence minimum, household composition, poverty measurement.

1. Introduction

Equivalence scales are analytical instruments used to compare the economic situation of households that differ in size and demographic composition. They reflect how a household's structure affects its cost of living (Statistics Poland, 2025). Simple per-capita measures of disposable income or total expenditure are inadequate for this purpose, as they ignore economies of scale arising from shared housing, joint household management, and collective consumption. To illustrate, consider a single-person household and a four-person household, each with a monthly disposable income of PLN 5,000 per person. At first glance, a four-person household might appear to be in a stronger financial position. A per-capita comparison would suggest identical living standards. However, due to economies of scale in consumption, the larger household

[1] Warsaw University of Life Sciences, Warsaw, Poland. E-mail: hanna_dudek@sggw.edu.pl.
ORCID: https://orcid.org/0000-0001-8261-2745

does not need to spend four times as much as the single-person household to achieve the same level of well-being.

Equivalence scales provide a more accurate basis for comparing the economic situation of households with different size and compositions. They are used in interpersonal and interhousehold comparisons of well-being in the measurement of social welfare, economic inequality and poverty (Lewbel & Pendakur, 2008). Formally, an equivalence scale is a measure of the cost of living of a household with a given size and demographic composition relative to that of a reference household, assuming both households attain the same level of utility or standard of living (Lewbel & Pendakur, 2008). In most applications, a one-person household serves as the reference unit and is assigned a scale value of one. Nevertheless, some studies adopt multi-person households as the reference, depending on the research objective (Koulovatianos et al., 2005).

While the importance of equivalence scales is widely acknowledged, their selection remains a major challenge. Different scales can lead to markedly different poverty rates and inequality measures (Coulter et al., 1992b; De Vos & Zaidi, 1997; Dudel et al., 2021). This issue is particularly salient in the measurement of extreme poverty, where poverty thresholds are defined in absolute terms and closely linked to basic needs.

This study addresses the research question of whether the original OECD equivalence scale, applied by Statistics Poland (GUS) in estimating extreme poverty, is appropriate. Specifically, it examines whether the scale effect implied by this scale is consistent with that embedded in the construction of the subsistence minimum across different household types. Using anonymized data from the 2023 Household Budget Survey, equivalence scales are approximated by a one-parameter power function, and extreme poverty rates in Poland are recalculated under alternative parameter values.

The paper contributes to the literature by enhancing the methodological foundations of extreme poverty measurement and by providing empirical evidence relevant for official statistics and social policy in Poland.

2. Equivalence scales

Various approaches to estimating equivalence scales can be broadly classified into expert-based (normative) and empirical (statistical) scales (Dudel et al., 2021; Szulc, 2006). Expert-based scales rely on normative judgments, typically formulated by expert panels or policy institutions, regarding the relative needs of different household types. Empirical scales, by contrast, are derived from statistical analyses of household expenditure or income data, aiming to capture actual consumption patterns and economies of scale.

Each approach has distinct advantages and limitations. Expert-based scales are transparent and easy to apply, making them well-suited for standardized applications

in official statistics. However, they risk oversimplifying household needs and overlooking heterogeneity in consumption patterns. Empirical scales, by contrast, offer greater flexibility and context-specific relevance but require high-quality data and strong identifying assumptions, making them methodologically more demanding. Table 1 summarizes the main features of both approaches.

Table 1. Key characteristics of expert-based and empirical equivalence scales

Characteristic	Expert-based scales	Empirical scales
Data Requirements	Low; base on expert knowledge or policy conventions	High; require detailed, high-quality household data
Flexibility	Low; usually uniform across populations and contexts	High; can be tailored to specific populations, or time periods
Complexity	Low; easy to apply and understand	High; require advanced statistical methods, such as econometric modeling, and strong identification strategies
Policy relevance	Strong for standardized comparisons and official statistics	Strong for context-specific analysis and nuanced poverty assessment
Limitations	Risk of oversimplifying needs and ignoring household heterogeneity	Methodologically complex; sensitive to model assumptions about behavior, preferences, and resource allocation.

Source: own work.

In expert-based approaches, household needs are defined by specifying minimum requirements for goods and services (Schröder, 2004). The most widely used scales of this type include the LIS scale, the original OECD scale, and the modified OECD scale.

In these approaches, household needs are typically assessed by experts who specify the minimum requirements for goods and services necessary to attain a given standard of living (Schröder, 2004). The most well-known and frequently applied scales of this type include the LIS (Luxembourg Income Study) scale, the original OECD (Organization for Economic Co-operation and Development) scale, and the modified OECD scale.

The LIS scale is defined as

$$S_{LIS} = \sqrt{N}, \tag{1}$$

where N is the number of persons in the household. The original OECD scale is given by

$$S_{70/50} = 1 + 0.7(a - 1) + 0.5c, \tag{2}$$

while the modified OECD scale takes the form

$$S_{50/30} = 1 + 0.5(a - 1) + 0.3c, \tag{3}$$

where *a* denotes the number of adults in the household and *c* is the number of children under the age of 14.

Expert-based equivalence scales are widely applied in official statistics. For instance, Statistics Poland (GUS) uses the original OECD equivalence scale for expenditure-based poverty measurement, whereas Eurostat relies on the modified OECD scale within its income-based poverty measurement framework. The widespread use of OECD scales reflects their practicality and suitability for standardized comparisons across populations and over time. In this respect, the OECD equivalence scales play an important role in ensuring international comparability, which constitutes a legitimate and key objective of official statistics.

Alongside expert-based approaches, the literature proposes a wide range of methods for estimating empirical equivalence scales (Biewen & Juhasz, 2017; Bishop et al., 2014; Coulter et al., 1992b; Donaldson & Pendakur, 2004; Dudel et al., 2021; Kot, 2015; Schröder, 2004; Szulc, 2009). Despite this extensive methodological development, scholars consistently emphasize that no universally accepted method exists for deriving a "correct" equivalence scale (Coulter et al., 1992a; Kot, 2015). As a result, researchers typically rely on a range of complementary approaches. In the Polish context, empirical scales have been estimated using diverse methods by Kot (2023), Kalbarczyk-Stęclik et al. (2017), Morawski (2018), Szulc (2014), and Ulman (2012), among others.

The literature indicates that equivalence scales depend on household income or total expenditure levels. Cross-country studies (Bishop et al., 2014; Kalbarczyk-Stęclik et al., 2017; Kot, 2023; Mysíková et al., 2022) find that economies of scale tend to be larger in Western European countries than in Southern or Central and Eastern European countries. More generally, equivalence scales tend to be lower in wealthier countries than in poorer ones (Calvi et al., 2023). This pattern reflects differences in consumption structures: in poorer countries, where food accounts for a larger share of household budgets, economies of scale are more limited because food is primarily a private good (Calvi et al., 2023; Deaton & Paxson, 1998).

Empirical analyses conducted within individual countries (e.g. Garbuszus et al., 2021; Kot, 2023; Pendakur, 1999) further demonstrate that the so-called independence-of-base (IB) assumption is typically rejected. This assumption posits that equivalence scale values are invariant with respect to household welfare, typically measured by income or consumption. In practice, empirical results indicate that equivalence scale values decline as income or expenditure rises (Donaldson & Pendakur, 2004; Kot, 2023). Consequently, applying income-independent scales such as the OECD scales may lead to biased assessments of poverty.

Against this background, the remainder of this paper examines equivalence scales for extremely poor households in Poland in 2023. As extreme poverty in Poland is defined in relation to the minimum subsistence level, the following section explores this concept.

3. The subsistence minimum

In Poland, minimum subsistence levels are used as thresholds for measuring extreme poverty. These levels are defined using the basic needs method, also referred to as the "basket" method, which specifies the minimum level of consumption necessary for survival. The subsistence minimum includes only fundamental existential needs such as food, clothing, hygiene, and housing, along with the educational requirements of school-age children (Deniszczuk & Sajkiewicz, 1997). This threshold, also referred to as the biological minimum, marks the point at which consumption becomes so limited that it poses a biological risk to human life and impairs psychophysical development (Kurowski, 2022). Accordingly, it covers only those expenditures that are strictly essential for sustaining life.

Minimum subsistence levels are calculated by the Institute of Labor and Social Affairs (IPiSS) for six types of employee households and two types of retired households. The annual average values established for 2023 are presented in Table 2.

Table 2. The minimum subsistence levels (average monthly values in 2023, in PLN)

Specification	Employee households (household sizes and types)						Retired households (sizes and types)	
	1 (M+W)/2	2 M+W	3 M+W+YC	3 M+W+OC	4 M+W+YC+OC	5 M+W+YC+2OC	1 (M+W)/2	2 M+W
Food	377.93	755.86	1077.60	1204.23	1525.97	1974.34	328.21	656.43
Housing	398.80	564.55	813.84	813.84	1069.69	1322.16	398.80	564.55
Education	0	0	6.44	80.93	87.37	168.30	0	0
Clothing and footwear	35.35	65.85	98.08	102.51	134.73	171.39	35.35	65.85
Healthcare	15.10	26.10	61.73	42.00	76.61	92.33	23.88	43.66
Personal hygiene	30.95	61.33	76.30	84.63	99.63	123.13	27.17	53.78
Other expenses	42.91	73.68	106.70	116.41	149.70	192.58	40.67	69.21
Total	901.04	1547.38	2240.69	2444.55	3143.69	4044.23	854.08	1453.49

Source: IPiSS (2024). Symbols: M – man aged 25-60, W – woman aged 25-60, (M + W)/2 – average expenditure for a household consisting of a man and a woman, YC – younger child aged 4-6, OC – older child aged 13-15. In the case of retired households, the symbols M and W denote men and women over 60, respectively.

An important feature of the minimum subsistence structure is the dominant role of food expenditures. In nearly all household types, with the exception of single-person retired households, food accounts for approximately 40–50% of total spending. As food is a necessary good and largely non-shareable, it is relatively insensitive to economies of scale. Consequently, a high proportion of food in total expenditure indicates that extremely poor households have limited scope to benefit from economies of scale.

To quantify the economies of scale associated with the subsistence minimum, the corresponding equivalence scale is approximated using the method described in Section 4.

4. Methods and data

The equivalence scale corresponding to the subsistence minimum levels is approximated using a one-parameter power scale of the form

$$S = N^e, \tag{4}$$

where e denotes the elasticity of scale with respect to the household size (Ayala & Jurado, 2011; Coulter et al., 1992a; Hunter et al., 2003), and N is the household size, defined as the number of individuals in the household.

By construction, e=0 corresponds to the case of maximum economies of scale, while e=1 means no economies of scale, with income or expenditure calculated on a per capita basis. Higher values of parameter e therefore indicate lower economies of scale, reflecting a smaller degree of resource sharing within the household (Buhmann et al., 1988).

As shown in the literature, any arbitrary equivalence scale S can be closely approximated by a single-parameter power scale (Buhmann et al., 1988; Coulter et al., 1992a). Specifically, if the scale is expressed as in formula (4), taking logarithms yields

$$ln(S) = e \cdot ln\,(N), \tag{5}$$

which allows the parameter e to be estimated using a linear regression model without an intercept, with $ln(S)$ as the dependent variable and $ln(N)$ as the independent variable. Buhmann et al. (1988), analyzing 34 equivalence scales across multiple countries, reported that the correlation coefficients between $ln(S)$ and $ln(N)$ were close to unity. Their findings indicate that any arbitrary equivalence scale can be effectively approximated by the one-parameter power scale. In the present study, the same approach is applied to estimate the parameter e corresponding to the equivalence scales under examination.

Next, equivalized total expenditures (Y^{ekw}) are computed as

$$Y^{ekw} = \frac{Y}{S}, \tag{6}$$

where Y is total household expenditure and S is the equivalence scale.

On the basis of equivalized total expenditures, poverty rates are subsequently calculated. The reference household is assumed to be a one-person employee household. A household was classified as poor if its equivalized expenditure fell below the established poverty threshold.

The poverty rate H (headcount ratio), defined as the percentage of individuals living below the poverty line, is calculated as

$$H = \frac{\sum_{i=1}^{n} U_i \cdot N_i \cdot weight_i}{\sum_{i=1}^{n} N_i \cdot weight_i}, \tag{7}$$

where U_i is a binary variable equal to 1 if the i-th household is classified as poor, and 0 otherwise, N_i denotes the number of individuals in the i-th household, and $weight_i$ is the survey weight assigned to the i-th household.

In this study, the survey weights (*MN_kal*) are derived from the 2021 National Census of Population and Housing, conducted by Statistics Poland. As documented in Statistics Poland (2024a), these weights are obtained through calibration procedures that adjust the sample to match the population's age and gender distributions.

The poverty rates are calculated using data from the 2023 Household Budget Survey (HBS), a nationwide survey conducted annually by Statistics Poland on the basis of a random sample of households. In 2023, the HBS covered a total of 28,089 households.

5. Results and discussion

To assess the impact of household composition on basic living costs, Table 3 reports equivalence scales derived from the minimum subsistence levels for various household types, as listed in the last row of Table 2. The reference household is a one-person employee household, for which the equivalence scale is set to one. For all remaining household types, the equivalence scale values are calculated by dividing their respective minimum subsistence levels by 901.04, which corresponds to the minimum subsistence level of the reference household. This approach enables a straightforward comparison of how differing household sizes and compositions affect the resources required to attain the minimum standard of living.

Table 3. Equivalence scale implied by minimum subsistence levels

Household type	Demographic composition	Minimum subsistence level [PLN]	Equivalence scale (S)	House-hold size (N)
1-person employee household	(M+W)/2	901.04	1	1
2-person employee household	M+W	1547.38	1.72	2
3-person employee household	M+W+YC	2240.69	2.49	3

Table 3. Equivalence scale implied by minimum subsistence levels (cont.)

Household type	Demographic composition	Minimum subsistence level [PLN]	Equivalence scale (S)	House-hold size (N)
3-person employee household	M+W+OC	2444.55	2.71	3
4-person employee household	M+W+YC+OC	3143.69	3.49	4
5-person employee household	M+W+YC+2OC	4044.23	4.49	5
1-person retired households	(M+W)/2	854.08	0.95	1
2-person retired households	M+W	1453.49	1.61	2

Source: own work based on IPiSS (2024). The symbols M, W, YC, and OC are defined in Table 2.

Using the data from the last two columns of Table 3, the elasticity of scale parameter with respect to the household size is estimated by means of a linear regression model without an intercept, where *ln(S)* is the dependent variable and *ln(N)* is the explanatory variable. For comparison, the parameter *e* is estimated for the original OECD scale among extremely poor households, using HBS microdata with survey weights and assuming a reference household poverty threshold of PLN 901.04. The estimation results for the elasticity of scales with respect to the household size are reported in Table 4.

Table 4. Estimation results for the elasticity of the scale parameter

Elasticity of scale	Estimate	Standard error	t-statistics	95% confidence interval		R-squared
				lower limit	upper limit	
Implied by minimum subsistence	0.882	0.026	33.944	0.820	0.943	0.994
Original OECD scale	0.790	0.001	620.430	0.788	0.793	0.998

Source: own work.

The results indicate that the original OECD scale is associated with a lower value of the elasticity parameter e, corresponding to stronger economies of scale. This finding means that the original OECD scale assumes a greater degree of resource sharing within households than is suggested by the structure of the subsistence minimum.

To illustrate the consequences of this difference, simulations are conducted using the Stata package and the 2023 HBS microdata for a range of values of the scale elasticity parameter. In all simulations, the extreme poverty threshold is set at PLN 901.04, corresponding to the minimum subsistence level of a one-person employee household. The resulting relationship between the elasticity parameter and the extreme poverty rate is depicted in Figure 1.

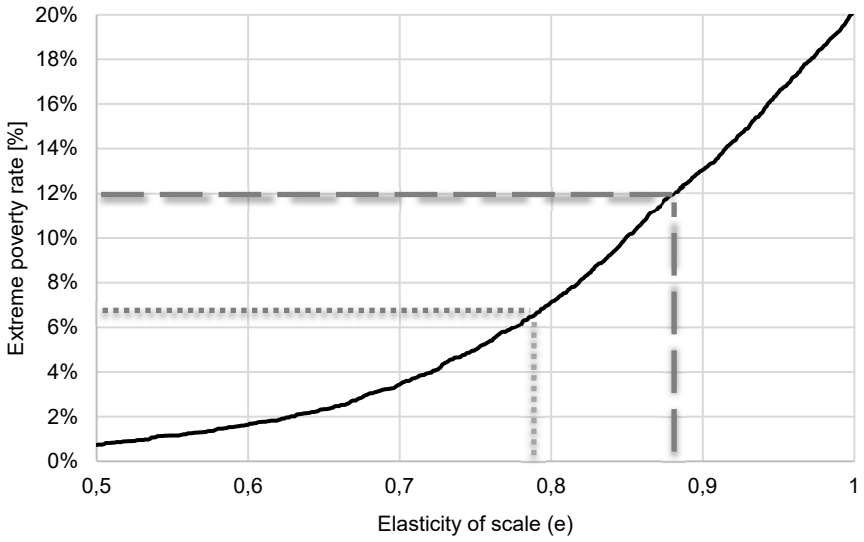

Figure 1. The impact of scale elasticity on the extreme poverty rate.

Source: own work based on HBS data. Note: the dotted grey line corresponds to the elasticity of scale for the original OECD scale, while the dashed line corresponds to the elasticity of scale implied by the minimum subsistence levels (see the values in the second column of Table 4).

Figure 1 demonstrates that the estimated extreme poverty rate is highly sensitive to the choice of equivalence scale. Higher values of the parameter *e* reduce the equivalized total expenditures of multi-person households, thereby increasing the risk that these households fall below the poverty threshold. As a result, higher elasticity of scale values is associated with higher poverty rates. Specifically, when the equivalence scale implied by the minimum subsistence level is applied (e=0.882), the extreme poverty rate is approximately 12%. In contrast, Statistics Poland reports an extreme poverty rate of 6.6% for 2023 (Statistics Poland, 2024b). This estimate corresponds to the application of the original OECD scale, which is associated with an elasticity parameter of e=0.790 (see Table 4).

These findings are consistent with earlier research demonstrating that poverty estimates are sensitive to the choice of an equivalence scale (Buhmann et al., 1988; De Vos & Zaidi, 1997; Éltető & Havasi, 2002). However, the present study focuses specifically on the data and methodological framework used to assess extreme poverty in Poland. It therefore provides new evidence on the importance of applying an appropriate equivalence scale when measuring extreme poverty in this context.

The analysis, based on transparent and replicable tools, shows that the original OECD equivalence scale implies stronger economies of scale than the scale derived from minimum subsistence levels. This conclusion is supported by the use of a one-

parameter power scale, which proves to be a very good approximation of the equivalence scales examined.

The study's findings should serve as a stimulus for discussion among experts in the field of poverty measurement and official poverty statistics in Poland.

6. Conclusions

This study provides new insights into the measurement of extreme poverty in Poland. By explicitly linking subsistence minimum values to the implied scale elasticity, it proposes a transparent and empirically grounded framework for examining the consistency of equivalence scales used in this context. The results indicate that the equivalence scale currently applied by Statistics Poland assumes stronger economies of scale than those embedded in the construction of the subsistence minimum. As a result, the extreme poverty rates reported in official statistics are likely underestimated, as the applied scale assumes a relatively high degree of intra-household resource sharing.

From a policy perspective, the results highlight the importance of aligning equivalence scales with the actual living costs faced by households at the margin of subsistence. Since alternative assumptions about household economies of scale yield substantially different estimates of extreme poverty rates, adopting an equivalence scale that more accurately reflects the consumption patterns of extremely poor households can improve both the accuracy and the credibility of official poverty statistics in Poland. In turn, this would enhance their usefulness for the design and targeting of social policy interventions.

References

Ayala, L., Jurado, A., (2011). Pro-poor economic growth, inequality and fiscal policy: The case of Spanish regions. *Regional Studies*, 45(1), pp. 103–121. https://doi.org/10.1080/00343400903173209.

Biewen, M., Juhasz, A., (2017). Direct estimation of equivalence scales and more evidence on independence of base. *Oxford Bulletin of Economics and Statistics*, 79, pp. 875–905. https://doi.org/10.1111/obes.12166.

Bishop, J. A., Grodner, A., Liu, H. and Ahamdanech-Zarco, I., (2014). Subjective poverty equivalence scales for Euro Zone countries. *Journal of Economic Inequality*, 12(2), pp. 265–278. https://doi.org/10.1007/s10888-013-9254-7.

Buhmann, B., Rainwater, L., Schmaus, G., and Smeeding, T. M., (1988). Equivalence scales, well-being, inequality, and poverty: Sensitivity estimates across ten countries

using the Luxembourg Income Study (LIS) database. *Review of Income and Wealth*, 34(2), pp. 115–142.

Calvi, R., Penglase, J., Tommasi, D. and Wolf, A., (2023). The more the poorer? Resource sharing and scale economies in large families. *Journal of Development Economics*, 160, 102986. https://doi.org/10.1016/j.jdeveco.2022.102986.

Coulter, F. A. E., Cowell, F. A. and Jenkins, S. P., (1992a). Differences in needs and assessment of income distribution. *Bulletin of Economic Research*, vol. 44(2), pp. 77–124. https://doi.org/10.1111/j.1467-8586.1992.tb00538.x.

Coulter F. A. E., Cowell F. A. and Jenkins S. P., (1992b). Equivalence scale relativities and the measurement of inequality and poverty. *Economic Journal*, (414), pp. 1067–1082. https://doi.org/10.2307/2234376.

Deniszczuk, L., Sajkiewicz, B., (1997). The category of minimum subsistence. In: S. Golinowska (ed.), Poverty in Poland. Research findings. Institute of Labor and Social Affairs (in Polish).

de Vos, K., Zaidi, M. A., (1997). Equivalence scale sensitivity of poverty statistics for the member states of the European Community. *Review of Income and Wealth*, 43(3), pp. 319–333. https://doi.org/10.1111/j.1475-4991.1997.tb00222.x.

Deaton, A., Paxson, C., (1998). Economies of scale, household size, and the demand for food. *Journal of Political Economy,* 106(5), pp. 897–930. https://doi.org/10.1086/250035.

Donaldson, D., Pendakur, K., (2004). Equivalent-expenditure functions and expenditure-dependent equivalence scales. *Journal of Public Economics*, 88(1-2), pp. 175–208. https://doi.org/10.1016/S0047-2727(02)00134-2.

Dudel, C., Garbuszus, J., Ott, N. and Werding, M., (2021). Income- (in)dependent equivalence scales and inequality measurement. *German Economic Review*, 22(2), pp. 235–255. https://doi.org/10.1515/ger-2020-0008.

Éltető, Ö., Havasi, É., (2002). Impact of choice of equivalence scale on income inequality and on poverty measures. *Review of Sociology*, 8(2), pp. 137–148. https://doi.org/10.1556/revsoc.8.2002.2.7.

Garbuszus, J. M., Ott, N., Pehle, S. and Werding, M., (2021). Income-dependent equivalence scales: A fresh look at German micro-data. *Journal of Economic Inequality*, 19, pp. 855–873. https://doi.org/10.1007/s10888-021-09494-7.

Hunter, B. H., Kennedy, S. and Smith, D., (2003). Household composition, equivalence scales and the reliability of income distributions: Some evidence for Indigenous and

other Australians. *Economic Record*, 79(244), pp. 70–83. https://doi.org/10.1111/1475-4932.00079.

IPiSS, (2024). *Information on the level and structure of the minimum subsistence level in 2023*. Retrieved from: https://www.ipiss.com.pl/wp-content/uploads/2024/04/ME-srednioroczne-2023.pdf.

Kalbarczyk-Stęclik, M., Miśta, R. and Morawski, L., (2017). Subjective equivalence scale – cross-country and time differences. *International Journal of Social Economics*, 44(8), pp. 1092–1105. https://doi.org/10.1108/IJSE-09-2015-0251.

Koulovatianos, C., Schröder, C. and Schmidt, U., (2005). Properties of equivalence scales in different countries. *Journal of Economics*, 86(1), pp. 19–27. https://doi.org/10.1007/s00712-005-0141-y.

Kot, S. M., (2015). Equivalence scales based on stochastic indifference criterion: The case of Poland. *Statistical Review*, 62, pp. 263–280. https://doi.org/10.5604/01.3001.0014.1753.

Kot, S. M., (2023). *Nonstandard Equivalence Scales and their Applications for European Union Countries*, Gdańsk: Politechnika Gdańska.

Kurowski, P., (2022). The role of the social minimum and subsistence minimum in Poland in a historical perspective. *Social Policy*, XLIX(1), pp. 1–15. (in Polish) https://doi.org/10.5604/01.3001.0016.0915.

Lewbel, A., Pendakur, K., (2008). *Equivalence scales*. In S. N. Durlauf & L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (2nd ed.). Palgrave Macmillan. https://doi.org/10.1057/978-1-349-95121-5_2056-1.

Morawski, L., (2018). Using subjective equivalence scales to analyze poverty in Poland. *Argumenta Oeconomica*, 41(2), pp. 139–160. https://doi.org/10.15611/aoe.2018.2.09.

Mysíková, M., Želinský, T., Garner, T. I. and Fialová, K., (2022). Subjective equivalence scales in Eastern versus Western European countries. *Contemporary Economic Policy*, 40(4), pp. 659–676. https://doi.org/10.1111/coep.12579.

Pendakur, K., (1999). Semiparametric estimates and tests of base-independent equivalence scales. *Journal of Econometrics*, 88(1), pp. 1–40. https://doi.org/10.1016/s0304-4076(98)00020-7.

Schröder, C., (2004). *Variable Income Equivalence Scales. An Empirical Approach*. Physica-Verlag, Heidelberg.

Statistics Poland, (2025) *Concepts Used in Official Statistics: Equivalence Scales*. Retrieved from: https://stat.gov.pl/metainformacje/slownik-pojec/pojecia-stosowane-w-statystyce-publicznej/3204,pojecie.html.

Statistics Poland, (2024a). *Methodological report. Household Budget Survey*. Warsaw.

Statistics Poland, (2024b). *The extent of economic poverty in Poland in 2023*. Retrieved from: https://stat.gov.pl/obszary-tematyczne/warunki-zycia/ubostwo-pomoc-spoleczna/zasieg-ubostwa-ekonomicznego-w-polsce-w-2023-roku,14,11.html.

Szulc, A., (2006). Poverty in Poland during the 1990s: Are the results robust? *Review of Income and Wealth*, 52(3), pp. 423–448. https://doi.org/10.1111/j.1475-4991.2006.00197.x.

Szulc, A., (2009). A matching estimator of household equivalence scales, *Economics Letters*, 103, pp. 81–83. https://doi.org/10.1016/j.econlet.2009.01.027.

Szulc, A., (2014). Empirical versus policy equivalence scales: matching estimation. *Bank i Kredyt*, 1, pp. 37–52.

Ulman, P., (2012). Equivalence scale in terms of Polish households' source of income. *Folia Oeconomica Stetinensia*, 10(2), pp. 114–127. https://doi.org/10.2478/v10031-011-0044-8.

STATISTICS IN TRANSITION new series, September 2026
Vol. 26, No. 3, pp. 243–244

In Memoriam of Professor Czesław Domański

It is with deep sadness that I bid farewell to Professor Czesław Domański, who passed away on July 6, 2026. At the same time, I wish to express—on behalf of myself, the Editorial Board, Associate Editors, and the Editorial Office of the Statistics in Transition new series—our great appreciation and gratitude for his unforgettable contribution to the development of our journal since its inception (he was the last of the journal's founders 33 years ago) and more broadly, to the advancement of statistical sciences; a mission we have jointly pursued over the past few decades.

Professor Czesław Domański was affiliated with the Department of Statistical Methods (as its long-term Head) at the Faculty of Economics and Sociology of the University of Łódź, where he advanced through every stage of his academic career, from assistant to full professor, focusing on mathematical and applied statistics, including econometrics—he served as Vice-Chair of the Committee on Statistics and Econometrics of the Polish Academy of Sciences, from 2008 to 2012—as well as demography and related fields. He was a man of many passions—spanning the academic, organizational, social, and artistic realms (including singing in a choir), which he pursued harmoniously while fulfilling various roles throughout his exceptionally active life.

A hallmark of his scholarly work was an innovative pursuit of truth, the sources of which he sought in data—primarily statistical data – in accordance with the highest ethical principles and scientific ideals. This shaped his (almost mystical) approach to statistics he viewed not merely as a crucial tool for uncovering truth, but above all, as the art of learning from data. He undoubtedly belonged to the elite circle of masters of the craft, embodying every facet of the statistician's profession: from theoretical works—where, among others, he pioneered several topics in nonparametric statistics—through conducting statistical research and teaching the discipline to advising decision-makers, scientific organizations, and public statistics institutions; most notably Statistics Poland through his long-standing membership in its Methodological Commission (since 2009), while collaborating

with the Statistical Office in Łódź—thereby earning recognition as a leader of the "Łódź school of statistics."

Through the prism of a broad methodological framework—a kind of triple helix linking truth, data, and statistics—he viewed both the problems requiring solution and his own areas of responsibility. This applied equally to his academic work—including his tenures as Vice-Rector of the University of Łódź (1993–1996), Dean of the Faculty of Economics and Sociology (1990-1993), and as Director of the Institute of Econometrics and Statistics (1997–2008) and the Institute of Statistics and Demography (2009–2013). And to his role in invigorating the scholarly life of national community of statisticians serving repeatedly as President of the Polish Statistical Association/PSA (years 1994–2005 and 2010–2019).

In this capacity, he fostered the engagement of successive generations of statisticians in collaboration with the PSA and international organizations—most notably the International Statistical Institute, of which he was an Elected Member. He also ensured that deceased statisticians were commemorated through the systematic publication of a regularly updated biographical series. However, his most significant achievement in this area was the establishment of the Jerzy Spława–Neyman Medal (marking the PSA's centennial in 2012), awarded to scholars from around the world for outstanding contributions to statistical science; he was one of its first recipients. His exceptional achievements across many fields of activity have been honored with numerous prestigious medals, orders, and decorations, awarded by both scientific organizations and institutions, and the state.

As we bid farewell to Professor Czesław Domański, we are left with heartening conviction—shared by all the journal's stakeholders, including authors and readers—of the enduring significance of his outstanding role and inputs to a wide range of statistics-related areas; all of this merits both commemoration and continuation.

Professor Włodzimierz Okrasa
Editor-in-Chief
Statistics in Transition new series

STATISTICS IN TRANSITION new series, September 2026
Vol. 27, No. 3, pp. 245–248

About the Authors

Ansari Farhan is an Assistant Professor in the Department of Community Medicine at Dr. S. S. Tantia Medical College, Hospital and Research Centre, Sri Ganganagar, Rajasthan, India. His research interests include order statistics, record values, generalized order statistics, statistical inference, and reliability analysis. He has published research papers and presented his work at national and international conferences.

Berg Emily is an Associate Professor in the Department of Statistics at Iowa State University, where she is affiliated with the Center for Survey Statistics and Methodology.

Chipepa Fastel is a Senior 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 70 research papers. His main areas of interest include distribution theory, survival analysis, biostatistical methods and biometry. He is a member of two editorial boards: Annals of Immunology and Immunotherapy (AII) and International Journal of Mathematics, Statistics and Operations Research.

Choudhury Amit is a statistician and operation researcher at the Department of Statistics, Gauhati University in Guwahati, India. He earned a B.S. (Hons) in Statistics from Cotton College, Guwahati, India, graduating top of his cohort. He then pursued an M.Stat. with specialization in Statistical Quality Control and Operations Research, and a Post-M.Stat. diploma at the Indian Statistical Institute, Kolkata, before completing his Ph.D. at Gauhati University in Queuing Theory. Prof. Choudhury is widely regarded as a passionate teacher and mentor, guiding numerous Ph.D. scholars and PG students. He blends research in queueing systems, Bayesian methods, and health statistics with a strong commitment to teaching and institutional excellence. His publications demonstrate his continued relevance in theoretical and applied statistics, while his leadership roles underpin his dedication to academic mentorship and institution building. His research interests include queuing theory, where he has explored balking, reneging, impatience, and finite-buffer models in single and multi-server systems; statistical inference in both frequentist and Bayesian perspectives as applied to Markovian queues and health outcomes; public-health statistics in the nature of chronic disease modelling. His research impact is evidenced by over 50 publications. He remains committed to advancing statistics education, operations research, and Bayesian methodology while guiding students through his leadership roles.

Dudek Hanna is an Associate Professor in the Department of Econometrics and Statistics at the Warsaw University of Life Sciences (SGGW). Her research interests lie at the intersection of applied econometrics and social statistics, with a particular focus on equivalence scales, poverty measurement, household consumption, food poverty, food insecurity, and material deprivation. She has authored or co-authored more than 130 publications, including articles in peer-reviewed scientific journals and chapters in academic monographs.

Habeeb Ali Salman received B.S., M.S., and the Ph.D. degrees from the University of Baghdad, Baghdad, Iraq. He is currently a full Professor in the Department of Statistics, College of Administration and Economics, University of Sumer. His research interests focus on statistical theory and its applications including the modeling and analysis of linear and nonlinear time series as well as mathematical statistics, econometric theory, and both parametric and non-parametric statistical methods. He has published over 13 research papers in international and local scientific journals and conference proceedings, six of which have appeared in journals indexed in the Scopus database.

Khan M. J. S. is working as Associate Professor at the Department of Statistics and Operations research in Aligarh Muslim University, Aligarh. He has 18 years of teaching experience. He has also worked in as an Assistant Professor at the Department of Applied Statistics, Babasaheb Bhimrao Ambedkar University (A Central University), Lucknow and at Department of Mathematics and Statistics, Banasthali University, Banasthali, Rajasthan. Dr. Mohd. Jahangir Sabbir Khan has published several papers in reputed national and international journals.

Kotlewski Dariusz is an Associate Professor at the Department of Economic Geography, Collegium of Business Administration, Warsaw School of Economics (SGH). Simultaneously, he holds the position of a Consultant in Statistics Poland (GUS). His main areas of interest include: economic growth accounting (including KLEMS), with particular emphasis on its spatial dimension, and the spatial functioning of industries. In addition to purely macroeconomic studies, his research is connected to regional studies, covering the electric power sector, transportation infrastructure, the real estate market, urban and regional development, environmental protection, etc. His most significant achievement is the methodology development and the implementation of the KLEMS productivity accounts in Polish statistics.

Kozłowski Arkadiusz is an Assistant Professor at the Department of Statistics, Faculty of Management, University of Gdańsk. His main area of interest is survey sampling, including probability and non-probability sampling techniques, data weighting, and calibration of weights. He is the head of postgraduate studies Data Analysis – Big Data.

Łukaszonek Wojciech is a teaching assistant at the University of Kalisz and a Ph.D. candidate at the Poznań University of Economics and Business. His research interests

focus on the applications of advanced statistical methods, including multivariate functional data, in the study of local labor markets. He is a co-author of several publications and an active participant in scientific conferences on applied statistics.

Mtemeri Nyika is a researcher affiliated with Avance Clinical in Australia, specializing in clinical trials. Armed with an M.S. in Mathematics and Statistics, he has authored peer-reviewed papers in eye health care that leverage rigorous statistical approaches for clinical applications. His academic pursuits extend to distribution theory, eye health, and survival analysis, alongside exploring AI-driven solutions for image classification and fraud detection.

Nkomo Wilbert is affiliated with the Department of Applied Statistics at Manicaland State University of Applied Sciences in Mutare, Zimbabwe, and the Department of Statistics at the University of South Africa. He holds an M.S. in Biostatistics and Epidemiology and a Ph.D. in Statistics, and has published over 15 research papers in international and national journals and conference proceedings. His research interests center on distribution theory, heavy-tailed distributions, actuarial risk measures, survival analysis, and quantile regression.

Nyakuamba Takesure is working at the Department of Applied Statistics, Manicaland State University of Applied Sciences, Mutare, Zimbabwe. He holds an M.S. in Operations Research and has published five papers. His research interests are in regression analysis, distribution theory, and optimization.

Panichkitkosolkul Wararit is an Associate Professor in the Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand, and is affiliated with the Thammasat University Research Unit in Mathematical Sciences and Applications. His research interests include statistical inference, confidence-interval methodology, time-series analysis, and simulation techniques. He also serves as Editor-in-Chief of Thailand Statistician and as an Associate Editor of Science & Technology Asia.

Pareek Namrata is a Research Scholar in the Department of Statistics at Gauhati University, under the supervision of Dr Amit Choudhury, Professor and Head of the Department. She also serves as an Assistant Professor in the Department of Statistics at Handique Girls' College. She completed her M.S. in Statistics from Gauhati University and her B.S. in Statistics from Cotton University, graduating as a gold medalist at both levels. In recognition of her academic performance, she was conferred the Women's Achiever Award in 2018 for being ranked first among girls in the Faculty of Science, Gauhati University. Her research interests include queueing theory, survey sampling, vital statistics, time series analysis, probability distributions, and testing of hypotheses. She has authored several book chapters in these areas and continues to contribute to

the statistical literature through her ongoing doctoral research on Bayesian inference in queueing systems.

Srisuradetchai Patchanok is an Associate Professor of Statistics in the Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand. He currently serves as Deputy Head of the Department for Research and Partnerships and as Chair of the M.S. Program in Applied Statistics. His research interests include statistical inference, computational statistics and simulation, machine learning, data mining, and statistical modelling with R. He has developed R packages for statistical design and nonparametric analysis.

Szymkowiak Marcin is an Associate Professor at the Department of Statistics, Poznań University of Economics and Business, Deputy Director at the Statistical Office in Poznań, a member of the Main Council of the Polish Statistical Association, and Chair of the Council of the Poznań Branch of the Polish Statistical Association, with many years of experience in statistical data analysis. He specializes in small area estimation, methods for handling non-response, survey sampling, multivariate data analysis, and statistical data integration methods.

Wołyński Waldemar is a Professor at the Faculty of Mathematics and Computer Science at Adam Mickiewicz University in Poznań, Poland. His major research interests focus on various aspects of multivariate statistical analysis. He is an author or co-author of over 70 research papers.

Zhou Zirou is a recent Ph.D. graduate in the Department of Statistics at Iowa State University. The main topics of his work are small area estimations, survey study, and Bayesian analysis.

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:* https://sit.stat.gov.pl/ForAuthors.

- ***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**. 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.**, (**1.1.**, **1.2.** …), **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 https://anglia.libguides.com/harvardctr. 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).