Volume 25, Number 4, December 2024 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

ISSN 1234-7655

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, *December 2024*
Vol. 25, No. 4, 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, *December 2024*
Vol. 25, No. 4, pp. V–VI

Policy Statement

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

***

Abstracting and Indexing Databases

Statistics in Transition new series is currently covered in:

BASE – Bielefeld Academic Search Engine

CEEOL – Central and Eastern European Online Library

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

CNKI Scholar (China National Knowledge Infrastructure)

CNPIEC – cnpLINKer

CORE

Current Index to Statistics

Dimensions

DOAJ (Directory of Open Access Journals)

EconPapers

EconStore

Electronic Journals Library

Elsevier – Scopus

Genamics JournalSeek

Google Scholar

Index Copernicus

J-Gate

JournalGuide

JournalTOCs

Keepers Registry

MIAR

Microsoft Academic

OpenAIRE

ProQuest – Summon

Publons

QOAM (Quality Open Access Market)

ReadCube

RePec

SCImago Journal & Country Rank

TDNet

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

Ulrichsweb & Ulrich's Periodicals Directory

WanFang Data

WorldCat (OCLC)

Zenodo

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. VII–XI

From the Editor

With this issue of Statistics in Transition new series, which we present to our Readers as the last one in 2024, we begin the first year of a new decade of its publication – the fourth since its launch in 1993. With 10 articles by authors from 8 countries – Slovakia, Poland, Ghana, USA, India, Nigeria, Bangladesh, and Iraq – this issue demonstrates the continuity in serving a broad, international community of experts in various fields of statistics. Both as an academic discipline and as application-oriented fact-based knowledge useful for solving problems in various areas of decision-making and other practical activities.

Wishing you a Happy New Year, we hope that a sense of true professional satisfaction will accompany our Readers while reading this issue, as well as all subsequent ones in the upcoming 2025.

Research articles

The first paper by **Erik Šoltés**, **Silvia Komara**, **Martina Košíková**, and **Tatiana Šoltésová** entitled ***Comparison of household work intensity in Slovakia and Czechia through least squares means analysis based on GLM*** addresses the problem of assessing the impact of relevant factors and their interactions on the *work intensity* (WI) of households in Slovakia and Czechia. For this purpose, general linear models, contrast analysis and estimates of marginal means are employed. The presented analyses are based on the EU-SILC 2021 data and carried out for Slovakia and Czechia. The paper reveals the common and different features of these countries in terms of WI of households. Particular attention is devoted to the identification of the profiles of persons at high risk of living in QJ households. The paper provides estimates of the marginal means of WI households for employed, unemployed and disabled persons with different education, age and from different types of households in Slovakia and in Czechia.

Tomasz Panek, **Jan Zwierzchowsk**i, and **Jan Kroszka** in the article ***The impact of the COVID-19 pandemic on the financial situation of people aged 50+ based on SHARE data*** describe the changes in the financial situation of households of people aged 50+ during the COVID-19 pandemic. The authors evaluate the outcome of the introduced national policies, the EU countries' economic performance, labor market conditions and the individual characteristics of the financial situation of the members

of the examined households. To achieve this goal, an original synthetic index was constructed to measure the changes in the overall financial situation of the surveyed group of households. This index combines various indicators, including income, subjective income assessment, the use of savings to finance current consumption and the postponement of bill payments, allowing a comprehensive evaluation of the shifts in the financial status of the 50+ population during the pandemic. In addition, the study aimed at examining how the age of respondents is interlinked with the changes in their financial situation using data from the Survey of Health, Ageing, and Retirement in Europe (SHARE). The study's findings show that during the pandemic, the changes in the financial situation of households with people aged 50+ varied across the selected countries.

In the next paper, ***Modelling Tinnitus Functional Index reduction using supervised machine learning algorithms***, prepared by **Edmund Fosu Agyemang**, the reduction in the Tinnitus Functional Index (TFI) is attempted utilizing supervised machine learning algorithms, focusing primarily on Ordinary Least Squares (OLS), K-Nearest Neighbor (KNN), Ridge, and Lasso regressions. The analysis highlighted Group, ISI, and SWLS as significant predictors of TFI reduction, identified through the best subset selection and confirmed by both forward and backward selection criteria in the OLS regression. Notably, the shrinkage methods, Ridge and Lasso regressions, demonstrated superior performance compared to OLS and KNN, with the Ridge regression presenting the smallest test mean square error (MSE). This finding establishes the Ridge regression as the best model for analyzing our Tinnitus dataset relative to the other methods. This research highlights the potential of supervised machine learning algorithms in advancing personalized Tinnitus treatment, reflecting broader trends in the field as evidenced by studies in the literature.

Housila P. Singh's, **Rajesh Tailor's**, and **Priyanka Malviya's** paper entitled ***Efficient use of auxiliary information in estimating finite population variance in sample surveys*** discusses the problem of estimating the finite population variance of the study variable y using information on the known population variance of the auxiliary variable x in sample surveys. The bias and mean squared error of the suggested class of estimators up to the first order of approximation was obtained. Preference regions were derived under which the suggested class of estimators is more efficient than the usual unbiased estimator, Das' and Tripathi's estimators (1980), Isaki's ratio estimator (1983), Singh's et al. estimator (1973, and Gupta's and Shabbir's estimator (2007). An empirical study as well as simulation study were carried out in support of the presented study.

Sebastian Wójcik's paper ***AMUSE: Analysis of mobility using simultaneous equations. Present population of refugees in Poland*** describes the significant influx of Ukrainian refugees into Poland following the escalation of the conflict in Ukraine after

February 2022. It highlights the challenges in tracking refugee movements using traditional statistical and administrative data sources due to problems associated with timeliness and spatial granularity. As a result, official statistics are turning to big data sources, such as mobile network operator (MNO) data, to supplement existing data. The paper focuses on utilizing synthetic MNO daily data from SIM cards issued to Ukrainian refugees by a Polish MNO. It proposes AMUSE model: mobility model for data de-duplication and a simple estimator for estimating the present refugee population based on aggregated signaling data over time and areas. Further research shall be focused on including data variability over time into modelling.

The article ***Forecasting under-five child mortality in Bangladesh: progress towards the SDGs target by 2030*** by **Sacchidanand Majumder** and **Soma Chowdhury Biswas** explores the under-five child mortality trends and estimates projection by 2030 to make progress towards achieving the SDGs target regarding the under-five child mortality in Bangladesh. Child mortality is a crucial indicator of a nation's socio-economic advancement and the well-being of mothers, reflecting the overall quality of life within a society. The yearly dataset regarding mortality among children aged five and under (per 1,000 live births) in Bangladesh employed in this study was collected from the World Bank Databank (https://data.worldbank.org/indicator) for the 1972–2022 period. The selection of the best-fitted model for the purpose of forecasting was between the ARIMA model and the Double Exponential Smoothing Holt's Method. Compared with the ARIMA (1,2,1) model, the Double Exponential Smoothing Holt's method proved the best-fitted model for forecasting the under-five child mortality in the future. The results show that under-five child mortality in Bangladesh is an annually declining trend. The average under-five child mortality is forecasted to drop by one during the 2023–2035 period. Thus, the predicted value of under-five child mortality would be 26 in 2025 and 22 in 2030; the national target is 27 (per 1,000 live births) in 2025 and the SDGs target (25 deaths per 1,000 live births).

Kuntal Bera and **M. Z. Anis** in the paper ***On some statistical properties of a stationary Gaussian process in the presence of measurement errors*** discuss some statistical properties, including the mean and variance of a stationary Gaussian process when observed data are affected by measurement errors. As a special case, the authors debate a stationary autoregressive process of order one with Gaussian white noise where measurement error follows an independent Gaussian distribution. To estimate some inferential results based on the collected sample, sometimes the mean and variance of the sample are required to be estimated. The results of this study show that the theoretical values based on the obtained results and the estimated sample values are reasonably close. Here, the authors have considered a particular case of a stationary Gaussian process, namely an AR(1) process. But there are more stationary Gaussian processes other than the AR(1) process. Also, in this paper, it is assumed that the

measurement error follows an independent Gaussian distribution but there are some situations where the measurement error does not follow Gaussian distribution.

The next paper, by **Arkadiusz Derkacz,** entitled ***A method of estimating the Return on Housing Investment (ROHI)*** presents a method for estimating the profitability level of housing investments. Market practice shows that the profitability of this type of investment is influenced by specific determinants that are absent in the classical approach to profitability analysis. The most commonly used method is the Return on Equity (ROE) ratio, which is dedicated to enterprises. However, housing investments are becoming an increasingly popular form of investment among private individuals. This makes the classical ROE method proved suboptimal for such investments ventures (i.e. those that involve the purchase of a residential property and its subsequent rental to third parties). In this context,-an attempt was made to develop a method that would allow to estimate the profitability level of this type of investment. It was found that the ROHI method enables the estimation of the profitability level, taking into consideration the most important determinants characteristic of this type of investment.

Other articles

XXXXI Multivariate Statistical Analysis 2023, Lodz, Poland. Conference Papers

Piotr B. Nowak's paper ***Estimation of the Cox model with grouped lifetimes*** presents how random numbers can be used to transform grouped lifetimes into a pseudo-complete sample. The aim of the study is to investigate the Fisher consistency of the partial likelihood estimator of the regression parameters in the Cox model based on the restored sample. It has been proven that for elliptical-type distributional assumptions about explanatory variables the estimators of the regression parameters in the Cox model based on the pseudo-complete sample are consistent up to a scaling factor. A simulation study illustrates the asymptotic properties of the estimates. In addition, real data case analysis is presented. The importance of the discussed problem is due to the fact that initial data are often aggregated and then classical methods based on the assumption of continuity of the dependent variable are limited. Therefore, the presented randomization method can also be used in other regression models, where a dependent variable is grouped.

Research Communicates and Letters

Ahmed Mahdi Salih and **Murtadha Mansour Abdullah** in their article entitled ***Comparison between classical and Bayesian estimation with joint Jeffrey's prior to Weibull distribution parameters in the presence of large sample conditions*** proposed a comparison of Weibull distribution parameters under large sample conditions. Three methods for estimating parameters: Maximum Likelihood Estimator, Moment Estimator, and Bayesian Estimator with a non-informative prior (weak prior with

minimal influence) are being evaluated. The classical estimation methods of Weibull distribution parameters have been chosen by the authors to the study, including the maximum likelihood estimator and moments estimation (ME). These methods were compared with the Bayesian estimation method (BE) with Jeffrey's prior function. The authors validated the proposed study via simulation using both small and large samples. To determine the best estimation method, mean square errors (MSEs) were used. The simulation findings suggest that maximum likelihood estimators are reasonably effective when using small sample sizes; in cases where the sample size is larger, the BE performs more effectively for both scale and shape parameters of the Weibull distribution function.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 1–26, https://doi.org/10.59139/stattrans-2024-001
Received – 01.05.2024; accepted – 18.07.2024

Comparison of household work intensity in Slovakia and Czechia through least squares means analysis based on GLM

Erik Šoltés[1], **Silvia Komara**[2], **Martina Košíková**[3], **Tatiana Šoltésová**[4]

Abstract

The work intensity (WI) of a household is primarily monitored in order to identify (quasi-)jobless (QJ) households. QJ households are those households whose members use less than 20% of their work potential. Individuals in such households, together with income-poor and severely materially and socially deprived persons are included in the Europe 2030 Strategy as socially excluded who need to be targeted by social policies.
The aim of the paper is to assess the impact of relevant factors and their interactions on the WI of households in Slovakia and Czechia. For this purpose, general linear models, contrast analysis and estimates of marginal means are employed. The presented analyses are based on the EU-SILC 2021 survey and carried out separately for Slovakia and Czechia. The paper reveals the common and different features of these countries in terms of the WI of households. Particular attention is devoted to the identification of the profiles of persons at high risk of living in QJ households.

Key words: work intensity, (quasi-)joblessness, general linear model, marginal means, contrast analysis.

1. Introduction

The work intensity (WI) of households is an indicator that is monitored within the Europe 2030 strategy. In addition to income poverty and material and social deprivation, special attention is directed towards the population residing in households

[1] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, Slovakia. E-mail: erik.soltes@euba.sk. ORCID: https://orcid.org/0000-0001-8570-6536.
[2] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, Slovakia. E-mail: silvia.komara@euba.sk. ORCID: https://orcid.org/0000-0001-6641-7456.
[3] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, Slovakia. E-mail: martina.kosikova@euba.sk. ORCID: https://orcid.org/0009-0003-4429-8398.
[4] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, Slovakia. E-mail: tatiana.soltesova@euba.sk. ORCID: https://orcid.org/0000-0002-0953-2519.

with very low work intensity (VLWI). Such households are commonly referred to as (quasi-)jobless (QJ) households.

Not only unemployment but also (quasi-)joblessness (QJ) is associated with poverty and social exclusion. According to Eurostat (2023a), in 2021, the at-risk-of-poverty (AROP) rate in the EU for the segment of the population living in QJ households was 62.3%, whereas in the segment of the population residing in households with very high WI, it was 5.4%. For persons living in households with very low or very high WI, the AROP rates in Czechia were 59.0% and 3.0%, respectively, and in Slovakia, they were 74.7% and 5.5%, respectively. Within the EU, Slovakia recorded the 7th highest AROP rate among QJ households. Eurostat noted higher rates only in the Baltic countries, the Netherlands, Sweden, and Croatia.

The VLWI rate, along with the AROP rate and the rate of severe material and social deprivation (SMSD), constitute the composite indicator AROPE (at risk of poverty or social exclusion). This indicator and its component rates are significantly influenced by individuals' economic activity status, with the unemployed and disabled individuals being particularly vulnerable (Eurostat 2023b). While the VLWI rate among the population with a disability was 180% (11.9 p.p.) higher than among the population without limitations in the EU, in Slovakia, it was 320% (8.6 p.p.) higher, and in Czechia, it was as much as 570% (14.2 p.p.) higher for persons with disabilities.

The mentioned facts motivated us to analyze WI in Slovakia and Czechia from the perspective of different economic activity statuses. This paper is not limited only to the VLWI but focuses on the WI index. This index is a target continuous numerical variable, which we analyzed depending on various factors through the analysis of marginal means and contrast analysis, which are based on a general linear model. The aim of the paper is to assess the pure influence (by fixing the influence of other relevant factors) of the most fundamental socio-economic and socio-demographic factors on WI in Slovakia and Czechia and subsequently compare outcomes between these two countries. The following research objectives are also oriented on the most relevant factors:

- to assess whether the impact of factors on WI is different or the same for different statuses of economic activity,
- to identify categories with no significant differences and identify categories or clusters where demonstrable differences in WI exist,
- to estimate the marginal means of WI for individual groups of persons and identify risk groups of persons in terms of QJ.

2. Literature review

According to Treanor (2018), employment can be a pathway out of poverty if it involves stable and quality work. When adult household members have unstable

employment, it is reflected in the household work intensity. Households with low WI struggle to earn an adequate annual income, which escalates their risk of poverty and material deprivation. In this paper, we do not deal only with VLWI, as other levels of WI can also be associated with poverty and social exclusion. E.g., Kis and Gábos (2016) showed that in the new member states of the EU not only low and very low household WI is positively associated with a higher risk of consistent poverty, but also medium WI. Fabrizi and Mussida (2020) showed that higher WI of Italian households with children significantly reduces the probabilities to fall into poverty and social exclusion. High utilization of the labour potential of households is therefore a key factor in combating poverty and social exclusion.

Low WI is correlated not only with income poverty but also with material deprivation. Duiella and Turrini (2014) and similarly García-Gómez et al. (2021) found relationship between poverty, material deprivation and WI which became stronger in countries most severely hit by the economic crisis in the period 2008–2014. Based on the above, we suppose that this relationship will intensify also due to the inflation, and war in Ukraine. Verbunt and Guio (2019) confirmed that WI is very effective in explaining within-country differences in the risk of income poverty/material deprivation in some CEE countries. In Slovakia, following the financial and economic crisis, unemployed persons living in households with a high and medium WI had markedly higher chances to move to employment, as compared to the unemployed in households with low WI (Gerbery and Miklošovic 2020). Fabrizi a Mussida (2020) found that living in a work-poor household is associated with living in consistent poverty (people at consistent poverty are those who are both at-risk-of-poverty and simultaneously experiencing enforced deprivation).

In addition, low WI of households has a negative impact on children and young people and on their social exclusion in the future. Analyses conducted by Treanor and Troncoso (2022) revealed that children in Scotland, whose parents consistently maintain high and medium WI, exhibit some of the lowest scores for both conduct and emotional problems. This suggests that stability in income and employment positively influences children's mental wellbeing. Guio and Vandenbroucke (2019) found out that QJ is an important driver of child deprivation in Belgium, even when income is controlled for. A similar conclusion was also reached by Regan and Maître (2020), when they demonstrated that during the period of 2008–2018, children in work-intensive households (households categorized as having high or very high WI) had a low AROP rate as well as a low deprivation rate. As noted by Hallaert et al. (2023), the proportion of children living in QJ households increased during the COVID-19 pandemic.

In selecting the explanatory factors, we relied on the results of our previous research and the works of other researchers. These factors include economic activity status, education, household type, age, marital status, health condition, region, and degree of

urbanization. We focused especially on the impact of the first four factors on WI, whose significant influence on poverty and social exclusion has been confirmed, e.g. by the studies mentioned below. Verbunt and Guio (2019) concluded that education plays a much bigger role in explaining poverty and social exclusion in economically weaker countries. According to Nieuwenhuis and Maldonado (2018), the risks of poverty among single-parent families are significantly higher than among complete families. Watson et al. (2015) revealed that across different household types, living in a jobless household was associated with a reduced probability that a non-employed person would make a transition into employment. Hofmarcher (2021) found strong effects of an additional year of education in reducing the likelihood of living in poverty. Education reduced not only the likelihood of relative income poverty but also reduced the likelihood of lacking essential household necessities and living in a household with weak labor market attachment.

3. Research methods

In this paper, we proceed from two separated general linear models (GLMs), one for Slovakia and the other for Czechia, based on which the influence of categorical factors and their interactions on a continuous numerical response variable characterizing WI will be assessed. In terms of interpreting the results, it is important to note that in our research, we used factors with fixed effects (Searle and Gruber 2017), and for categorical factors, we used indicator (dummy) coding (Darlington and Hayes 2016). The interaction was based on the crossed classification structure (Littell et al. 2010).

The general linear model can be written in matrix form as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

The matrix $\boldsymbol{X}$ is of non-full-rank, and a generalized inverse method is used to estimate the vector of parameters $\boldsymbol{\beta}$, the result of which is an estimate:

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y} \tag{2}$$

where the matrix $(\boldsymbol{X}^T\boldsymbol{X})^-$ is a generalized inverse matrix that must satisfy at least the first of the Penrose conditions (Searle and Gruber 2017). The estimation of the vector of parameters $\boldsymbol{\beta}$ obtained by the generalized inverse method is not unique, but there is a group of linear functions of the parameters, which we refer to as estimable functions $\boldsymbol{L\beta}$ (Elswick et al. 1991), for which there is a single solution (in more detail, e.g. Agresti 2015 and Littell et al. 2010).

As the aim of the paper is, among other things, to assess between which categories of relevant factors there is a significant difference, the subject of our interest will be the

testing of general linear hypotheses. To verify the general linear hypothesis (see more detail in, e.g. McFarquhar 2016 or Searle and Gruber 2017)

$$H_0: \quad \boldsymbol{L\beta} = \mathbf{0} \tag{3}$$

the following test statistic is used:

$$F = \frac{\frac{(\boldsymbol{Lb})^T \cdot \left[\boldsymbol{L}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-}\boldsymbol{L}^T\right]^{-1} \cdot (\boldsymbol{Lb})}{l}}{\frac{SSE}{n-p}} \tag{4}$$

where l *is* the number of independent rows of the matrix $\boldsymbol{L}$, the *SSE* is sum of squared residuals, n is the sample size, p is the number of parameters of the GLM. We reject the null hypothesis if the value of the test statistic satisfies the inequality:

$$F > F_{1-\alpha}(l\,;\ n-p) \tag{5}$$

The test mentioned above is used to verify simple hypotheses (if $l = 1$) and to simultaneously test multiple hypotheses (if $l \geq 2$). To verify simple hypotheses, of course, a t-test is also used, or alternatively, an interval estimate is constructed as well (see, e.g. Kuznetsova et al. 2017; Littell et al. 2010; and Westfall and Tobias 2007).

The analyses presented in the paper are based on unbalanced data, while we assess the impact of several effects. In such a situation, group arithmetic means do not provide an adequate picture of the response of the target variable for the particular factor because they do not take into account other effects, which may lead to the Simpson paradox (Wang et al. 2018). Cai (2014) states that if the data are unbalanced, arithmetic means are not appropriate because they do not consider that not all factors have the same chance of influencing the target variable. In such cases, it is appropriate to estimate the marginal means, which are based on the model (in our case on the GLM). The marginal mean is also referred to as the LS-mean (Least Squares mean; SAS Institute Inc. 1997) or the EM-mean (Estimated Marginal mean; Searle et al. 1980). LS-means are predicted means that are calculated from the fitted model and are adjusted appropriately for any other variable (Suzuki et al. 2019).

In this paper, we use marginal mean analysis using the LSMEANS statement. In addition, we employ contrast analysis (Dean et al. 2017; Schad et al. 2020) using the CONTRAST statement within PROC GLM (SAS Institute Inc. 2018) in the SAS programming language. Through contrast analysis we test the equality of marginal means of the target variable for individuals belonging to different groups determined by the interaction of two categorical factors.

For simplicity, let us consider the two-way classification model (see Little et al. 2010)

$$y_{ijk} = \underbrace{\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\mu_{ij}} + \varepsilon_{ijk} \tag{6}$$

where y_{ijk} is the k-th observed score for the (i, j)-th cell, α_i is the effect of the i-th $(i = 1,\ 2,\ \ldots, a)$ level of factor A, β_j is the effect of the j-th $(j = 1,\ 2,\ \ldots, b)$ level of factor B, $(\alpha\beta)_{ij}$ is the interaction effect for the i-th level of factor A and j-th level of factor B, μ is the overall mean, μ_{ij} is the marginal mean for the i-th level of factor A and j-th level of factor B, ε_{ijk} is the random error associated with individual observations.

Model (6) can be rewritten in the form of a regression model as follows:

$$y_{ij} = \mu + \sum_{i=1}^{a} \alpha_i X_{Ai} + \sum_{j=1}^{b} \beta_j X_{Bj} + \sum_{j=1}^{b} \sum_{i=1}^{a} (\alpha\beta)_{ij} X_{Ai} X_{Bj} + \varepsilon_{ij} \qquad (7)$$

where μ is the intercept, X_{Ai} is the dummy variable for i-*th* level of factor A, X_{Bj} is the dummy variable for j-*th* level of factor B and α_i, β_j, $(\alpha\beta)_{ij}$ are regression coefficients.

Let us assume that our interest lies in verifying the equality of the following marginal means:

$$H_0 : \mu_{21} = \mu_{22} = \mu_{24} = \mu_{2b} \qquad (8)$$

That is, we test the equality of the marginal means at the second level of factor A for the first, second, fourth, and last level of factor B. This hypothesis regarding the equality of 4 means needs to be reformulated into the form of 3 null hypotheses, each of which is a linear combination of the respective means. The mentioned null hypothesis can be expressed, for example, through these 3 null hypotheses:

$$H_0 : \mu_{21} = \mu_{22} \wedge H_0 : \mu\,(\mu_{21},\ \mu_{22}) = \mu_{24} \wedge H_0 : \mu\,(\mu_{21},\ \mu_{22},\ \mu_{24}) = \mu_{2b}$$

which we rewrite into linear combinations:

$$\mu_{21} - \mu_{22} = 0 \qquad \frac{1}{2}\mu_{21} + \frac{1}{2}\mu_{22} - \mu_{24} = 0 \qquad \frac{1}{3}\mu_{21} + \frac{1}{3}\mu_{22} + \frac{1}{3}\mu_{24} - \mu_{2b} = 0 \qquad (9)$$

Based on these linear combinations, we then determine the elements of a contrast matrix **L**, which determines the general linear hypothesis (3). The elements of matrix **L** are obtained by rewriting each marginal mean μ_{ij} in hypotheses (9) according to equation (6). The coefficients associated with the effects α_i, β_j and $(\alpha\beta)_{ij}$ are the sought-after elements of the contrast matrix **L**. A simpler way to determine these elements is to utilize a contingency table in the form of Table 1.

Table 1: Coefficients of a contrast matrix **L** in the case of factor interaction

Factor α	**Factor β**						**Sum**
	1	**2**	**3**	**4**	⋯	***b***	
1	l_{11}	l_{12}	l_{13}	l_{14}	⋯	l_{1b}	$l_{1\bullet}$
2	l_{21}	l_{22}	l_{23}	l_{24}	⋯	l_{2b}	$l_{2\bullet}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
a	l_{a1}	l_{a2}	l_{a3}	l_{a4}	⋯	l_{ab}	$l_{a\bullet}$
Sum	$l_{\bullet 1}$	$l_{\bullet 2}$	$l_{\bullet 3}$	$l_{\bullet 4}$	⋯	$l_{\bullet b}$	$l_{\bullet\bullet}$

Thus, to test the null hypothesis (8), we will simultaneously test hypotheses (9). Let us now show how we determine the coefficients for the last hypothesis in (9). We write the coefficients of the linear combination into the cells of Table 1 and do their summation in the sum row and sum column, as indicated in Table 2.

In the total row and total column in Table 2, there are coefficients for the factor A and factor B, respectively. The coefficients for interaction $A \cdot B$ are listed in the field of Table 2. Similarly, we could determine the coefficients for the other 2 partial hypotheses.

Table 2: Coefficients for the CONTRAST statement to test the null hypothesis $H_0 : \frac{1}{3}\mu_{21} + \frac{1}{3}\mu_{22} + \frac{1}{3}\mu_{24} - \mu_{2b} = 0$

Factor α	**Factor β**						**Sum**
	1	**2**	**3**	**4**	⋯	***b***	
1	0	0	0	0	⋯	0	**0**
2	$\frac{1}{3}$	$\frac{1}{3}$	0	$\frac{1}{3}$	⋯	-1	**0**
3	0	0	0	0	⋯	0	**0**
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
a	0	0	0	0	0	0	**0**
Sum	$\frac{1}{3}$	$\frac{1}{3}$	**0**	$\frac{1}{3}$	⋯	**-1**	**0**

The coefficients determined this way can be used in the CONTRAST statement, while the relevant variable and its associated coefficients for the individual partial hypotheses of simultaneous testing are separated by a comma. In our case the CONTRAST statement has the following syntax:

```
CONTRAST '21=22=24=2b'
         β 1 -1 α*β 0 … 0 1 -1,
         β 0.5 0.5 0 -1 α*β 0 … 0 0.5 0.5 0 -1,
         β 0.33333 0.33333 0 0.33333 … -1
         α*β 0 … 0 0.33333 0.33333 0 0.33333 … -1;
```

Let us note that:

- in all three partial hypotheses (9), the coefficients for factor A are equal to zero, so factor A is not given in the CONTRAST statement;
- for the interaction $A \cdot B$, the coefficients are initially 0 ... 0, representing the 0 coefficients given in the first row of the array of Table 2;
- the trailing zero coefficients for factor A, for factor B and for the interaction $A \cdot B$ need not be specified in the CONTRAST statement (we have not specified them);
- if factor A or factor B itself is not included in model (1), and only the interaction of these factors is included, then factor A and factor B, respectively, will not be listed in the CONTRAST statement.

To estimate the marginal means of the target variable, we employ ESTIMATE statements, whose construction is similar to that of CONTRAST statements. When estimating the marginal means, in addition to the effects, the overall mean is also included (see formula (6)), and its coefficient is the total sum $l_{\bullet\bullet}$ from Table 1. The procedures in SAS used in this paper are largely universal and are also used in other software and open-source systems (see Lenth 2016; Tabachnick and Fidell 2013).

4. Database

The paper analyses WI in Slovakia (SK) and Czechia (CZ) using a general linear model with explanatory variables listed in Table 3, while it should be emphasized that we focused only on persons whose economic activity status was employed, unemployed or disabled. The analyses are based on the EU-SILC 2021 database provided by the Statistical Office of the Slovak Republic (SO SR) and the Czech Statistical Office (CZSO). The presented analyzes are based on data on household members from Slovakia (6,277 persons) and Czechia (7,749 persons), while retired, students, housepersons, other inactive persons were not included.

Table 3: Description of input explanatory variables

Original variables (EU-SILC) – categories and description	Names of new dummy variables
HT – Household type	**HT**
Single-person household	1A_0Ch
Single parent household with at least 1 dependent child	1A_1^{+}Ch
2 adults' household, at least 1 aged 65+	2A(1^{+}R)
2 adults' household without dependent children	2A_0Ch
2 adults' household with 1 dependent child	2A_1Ch
2 adults' household with 3+ dependent children	2A_3^{+}Ch
Other households without dependent children	Other_0Ch
Other households with dependent children	Other_1^{+}Ch
2 adults' household with 2 dependent children	2A_2Ch
PE041 – The highest level of education achieved (ISCED)	**EDUCATION**
Pre-primary (ISCED 0)	ISCED 0-2
Primary (ISCED 1)	
Lower secondary (ISCED 2)	
Upper secondary (ISCED 3)	ISCED 3-5
Post-secondary (not tertiary) (ISCED 4)	
Short cycle of tertiary education (ISCED 5)	
Bachelor or equivalent (ISCED 6)	ISCED 6-8
Master's or equivalent (ISCED 7)	
Doctorate or equivalent (ISCED 8)	

Table 3: Description of input explanatory variables (cont.)

Original variables (EU-SILC) – categories and description	Names of new dummy variables
RX010 – Age	**AGE**
Age at the end of income reference period	-30
	30-40
	40-50
	50^{+}
PH010 – General health	**HEALTH**
Very bad	Bad
Bad	
Fair	Fair
Good	Good
Very good	
PB190 – Marital Status	**MARITAL STATUS**
Single	Single
Married	Married
Widowed	Widowed
Divorced	Divorced
RB090 – Gender	**GENDER**
Male	M
Female	F
DB100 – Degree of urbanisation	**URBANISATION**
Thinly populated area	Sparse
Intermediate area	Intermediate
Densely populated area	Dense

Source: own work based on EU-SILC data and European Commission (2021)

The definition of the target variable WI is given by the methodology used by Eurostat. Based on it, WI of a household is the ratio of the total number of months that working-age household members (aged less than 65 years) have worked during the income reference year and the total number of months the same household members theoretically could have worked in the same period.

We work with a continuous numerical variable WI according to the mentioned definition, but we also interpret the results of our analyses in relation to the level of WI: very low (VLWI – household working time was equal to or less than 20% of the full potential, QJ household), low (LWI), medium (MWI), high (HWI), which are defined for households' WI from intervals [20%; 45%), [45%; 55%], (55%; 85%] and very high (VHWI – household working time was more than 85% of the full potential).

5. General linear model for work intensity

5.1. Regressor selection and model estimation

Utilizing the stepwise regression method (Agresti 2015), we included the regressors listed in Table 4 into the model. Naturally, WI is fundamentally influenced by economic activity. The EA variable alone explains approximately 45% of the WI variability within each examined country. The impact of the other variables listed in Table 3 also proved to be significant. These variables affect WI differently for various statuses of economic activity, confirming the significance of individual interactions (Table 4). By incorporating these interactions, we were able to substantially increase the explained variability of WI to more than 50% (approximately 54% in Slovakia and around 53% in Czechia).

Table 4: Fundamental analysis of the models for WI in Slovakia and Czechia

R-Square		Coeff Var		Root MSE		WI Mean	
SK	CZ	SK	CZ	SK	CZ	SK	CZ
0.5395	0.5290	25.5575	18.9727	0.2020	0.1668	0.7904	0.8794

Source	DF	Partial R-Square		Type III SS		F Value		Pr > F	
		SK	CZ	SK	CZ	SK	CZ	SK	CZ
EA	2	0.4416	0.4515	19.6849	12.9115	241.21	231.90	<.0001	<.0001
EA×HT	24	0.0461	0.0463	15.5075	13.9110	15.84	20.82	<.0001	<.0001
EA×Education	6	0.0280	0.0037	14.1802	1.0287	57.92	6.16	<.0001	<.0001
EA×Age	9	0.0116	0.0123	6.5344	4.8655	17.79	19.42	<.0001	<.0001
EA×Gender	3	0.0024	0.0090	1.3544	4.3760	11.06	52.40	<.0001	<.0001
EA×Urbanisation	6	0.0048	0.0020	2.6302	0.8924	10.74	5.34	<.0001	<.0001
EA×Health	6	0.0027	0.0017	1.3963	0.8501	5.70	5.09	<.0001	<.0001
EA×Marital_status	9	0.0023	0.0026	1.0853	1.1868	2.96	4.74	0.0017	<.0001

Source: EU-SILC 2021, own processing in SAS EG.

Following the EA variable, WI is most strongly influenced by household type (HT). This factor, in interaction with EA, contributes to explaining approximately 4.6% of the variability of WI. These findings hold true for both countries. The contribution of further interactions is below 3% and varies between Slovakia and Czechia. The next most significant interaction is EA×Education in Slovakia and EA×Age in Czechia, yielding contributions of 2.8% and 1.2%, respectively. The contribution of each of the other interactions listed in Table 4 is less than 1%.

Table 5: Estimation of model parameters for WI in Slovakia and Czechia

Parameter		SK			CZ		
EA category		**D**	**U**	**E**	**D**	**U**	**E**
Intercept		0.896***	0.896***	0.896***	0.893***	0.893***	0.893***
EA intercept		-0.087	-0.139**	0.000	-0.405***	-0.347***	0.000
HT	**1A_0Ch**	-0.475***	-0.145**	0.091***	-0.353***	-0.140***	0.091***
	1A_1⁺Ch	-0.414***	-0.256***	0.032	-0.271***	-0.142***	0.053***
	2A(1⁺R)	-0.443***	-0.301***	0.104***	-0.319***	-0.265***	0.079***
	2A_0Ch	-0.151***	-0.138***	0.025**	-0.049	-0.036	0.057***
	2A_1Ch	-0.068	-0.082*	-0.002	-0.116*	0.061	0.010
	2A_3⁺Ch	-0.025	-0.169***	-0.064***	-0.153*	-0.127**	-0.052***
	Other_0Ch	-0.101*	-0.065*	0.004	-0.040	0.050	0.036***
	Other_1⁺Ch	-0.092	-0.114***	-0.059***	0.172**	-0.001	-0.017*
	2A_2Ch	0.000	0.000	0.000	0.000	0.000	0.000
EDU	**ISCED 0-2**	-0.084	-0.361***	-0.183***	0.069	-0.100**	-0.009*
	ISCED 3-5	-0.005	-0.090**	-0.006	0.041	0.049	0.001
	ISCED 6-8	0.000	0.000	0.000	0.000	0.000	0.000
Age	**30-40**	-0.251***	-0.095***	-0.012	-0.093**	-0.217***	0.007
	40-50	-0.207***	-0.163***	0.067***	-0.183***	-0.111***	0.056***
	50+	-0.176***	-0.102***	0.029***	-0.006	-0.151***	0.025***
	-30	0.000	0.000	0.000	0.000	0.000	0.000
Health	**Bad**	-0.077	0.000	-0.053***	-0.125***	-0.005	-0.035***
	Fair	-0.091	-0.033	-0.036***	-0.092**	-0.066**	-0.018***
	Good	0.000	0.000	0.000	0.000	0.000	0.000
Sex	**M**	0.005	-0.072***	-0.025***	0.089***	-0.011	-0.048***
	F	0.000	0.000	0.000	0.000	0.000	0.000
Urban.	**Intermediate**	-0.095***	0.050	-0.024***	-0.005	0.130***	0.000
	Sparse	-0.033	0.119***	-0.040***	0.025	0.118***	0.010*
	Dense	0.000	0.000	0.000	0.000	0.000	0.000
MS	**Divorced**	0.035	-0.127***	-0.006	0.058*	-0.086**	-0.005
	Never_married	-0.050*	-0.086***	0.003	-0.104***	-0.111***	0.006
	Widowed	0.030	0.064	0.012	0.111*	0.189	-0.015
	Married	0.000	0.000	0.000	0.000	0.000	0.000

***p < 0.01; **0.01< p < 0.05; *0.05< p < 0.1, level of significance

Source: EU-SILC 2021, own processing in SAS EG.

Table 5 illustrates the effects of individual factors (such as HT, Education, Age, and so on) on WI of households, wherein the person with the respective economic activity status (D - disabled, U - unemployed, E - employed) resides. Examining the HT factor, for disabled and unemployed persons, WI is lowest in single-person households

(1A_0Ch), households of 1 adult with at least 1 child (1A_1$^+$Ch), and households of 2 adults, at least 1 of whom is aged 65$^+$ (2A(1$^+$R)). If an adult with disabled status resides in these types of households, WI is generally more than 40 p.p. lower in Slovakia and about 30 p.p. lower in Czechia than in households of 2 adults with 2 children. In the case of unemployed persons, these differences are considerably smaller but still significant, even at a significance level of 0.01. Now turning our attention to the effect of education, the models reveal that low-educated individuals (ISCED 0–2) live in households with the lowest WI. This claim is statistically supportable for the unemployed and employed, but not for the disabled. This feature is also common to Slovakia and Czechia. If a disabled or unemployed person is under the age of 30, they live in a household with a significantly higher WI than older people. This phenomenon may be related to the fact that, for relatively young people, unemployment may be temporary, or the disability is such that the person is able to be employed. There are other intriguing findings arising from Table 5, but these will not be elaborated upon in this section, as the subsequent analyses in later sections of the paper will unveil them more comprehensively.

In the subsequent sections of the paper, in addition to examining the influence of economic activity itself, our focus will shift towards quantifying the effects of the other three most significant factors: household type, education, and age.

5.2. Analysis of LS Means and Contrast Analysis under GLM

In this section, we will employ an analysis of marginal means to identify significant differences in WI among pairs of categories within the household type (HT) factor, while controlling for the influence of other factors. Given the interaction of this factor with economic activity, this comparison will be conducted separately for each economic activity. Contrast analysis will be employed to form clusters of household types ensuring no significant difference between the household types within a cluster in terms of WI, while demonstrating a notable difference between the clusters. For these clusters, we will estimate the mean of WI, providing insight into the impact of the factor on WI, and enabling the identification of the most vulnerable individuals in terms of QJ.

5.2.1. Analysis of the impact of interaction *EA*×*HT*

The *p*-values matrix for LS-means equality tests (Table 6) indicates whether various types of households with a disabled person exhibit different WI. Let us narrow our focus to Slovakia, where the *p*-values are displayed above the main diagonal. We see that there are insignificant differences between the pairs of the first three types of households (1A_0Ch, 1A_1$^+$Ch, 2A(1$^+$R)) and between the pairs of the last 5 types of households (2A_1Ch, 2A_2Ch, 2A_3$^+$Ch, Other_0Ch, Oth_1$^+$Ch).

Table 6: Comparison of LS-means of WI for effect *EA×HT* for the disabled in Slovakia (above the diagonal) and in Czechia (below the diagonal)

Least Squares Means for effect EA×HT Pr > |t| for H0: LSMean(i)=LSMean(j)									
i/j	**1A 0Ch**	**1A 1⁺Ch**	**2A (1⁺R)**	**2A 0Ch**	**2A 1Ch**	**2A 3⁺Ch**	**Other 0Ch**	**Other 1⁺Ch**	**2A 2Ch**
1A_0Ch		0.4580	0.5143	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
1A_1⁺Ch	0.1027		0.7205	0.0007	<.0001	0.0020	<.0001	<.0001	<.0001
2A(1⁺R)	0.4135	0.3766		<.0001	<.0001	0.0001	<.0001	<.0001	<.0001
2A_0Ch	<.0001	<.0001	<.0001		0.1025	0.0224	0.0726	0.0605	0.0086
2A_1Ch	<.0001	0.0057	<.0001	0.0846		0.6974	0.5302	0.6570	0.3288
2A_3⁺Ch	0.0046	0.1281	0.0213	0.1142	0.5916		0.4636	0.5228	0.8303
O_0Ch	<.0001	<.0001	<.0001	0.7661	0.0611	0.0888		0.7813	0.0809
O_1⁺Ch	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001		0.1227
2A_2Ch	<.0001	0.0004	<.0001	0.4435	0.0936	0.0749	0.5367	0.0122	

Source: EU-SILC 2021, own processing in SAS EG.

The findings led us to assume that disabled persons living in the first three types of households have, on average, the same WI in their households ($H_0 : \mu_{11} = \mu_{12} = \mu_{13}$). This assumption was tested through a simultaneous test of two null hypotheses:

$$H_0 : \mu_{11} = \mu_{12} \wedge H_0 : \mu\,(\mu_{11},\ \mu_{12}) = \mu_{13}$$

To accomplish this, we employed the CONTRAST statement in the SAS programming language, as outlined in Section 3. The results are presented in the first row of Table 7, showing the test statistic value (4) and the corresponding p-value. Similarly, for disabled persons of the last 5 household types, we conducted tests for the equality of marginal means: $H_0 : \mu_{15} = \mu_{16} = \mu_{17} = \mu_{18} = \mu_{19}$.

Table 7: Equality tests for LS-Means of WI for the disabled from selected types of households in Slovakia

Contrast	**DF**	**Contrast SS**	**MS**	**F Value**	**Pr > F**
(EA=D)×HT 11=12=13	2	0.0297	0.0148	0.36	0.6953
(EA=D)×HT 15=16=17=18=19	4	0.1451	0.0363	0.89	0.4695
(EA=D)×HT 11=12=13 vs 15=16=17=18=19	1	3.4742	3.4742	85.14	<.0001
(EA=D)×HT 11=12=13 vs 14	1	2.3674	2.3674	58.02	<.0001
(EA=D)×HT 15=16=17=18=19 vs 14	1	0.3413	0.3413	8.36	0.0038

Source: EU-SILC 2021, own processing in SAS EG.

The contrast analysis (Table 7) did not reject the assumption that, in Slovakia, disabled persons living in household types 1A_0Ch, 1A_1⁺Ch, and 2A(1⁺R), have on average the same WI ($p = 0.6953$) under the condition of ceteris paribus. A similar conclusion was reached for household types 2A_1Ch, 2A_2Ch, 2A_3⁺Ch, Other_0Ch,

Oth_1⁺Ch ($p = 0.4695$). Moreover, these 2 clusters of household types have significantly different means of WI from each other ($p < 0.0001$) and are also significantly different from the fourth household type 2A_0Ch ($p < 0.0001$ and $p = 0.0038$). In this way, we have managed to create 3 distinct clusters of household types for disabled persons in Slovakia from the perspective of WI. To estimate the WI mean in the resulting household clusters, we used the ESTIMATE statement in the SAS programming language. The results are presented in Table 8.

Table 8: The estimate of $\mu(\mu_{11}, \mu_{12}, \mu_{13})$, μ_{14} and $\mu(\mu_{15}, \mu_{16}, \mu_{17}, \mu_{18}, \mu_{19})$ for the *EA×HT* interaction (disabled in Slovakia)

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(EA=D)×HT 11=12=13	0.0854	0.0385	2.22	0.0268
(EA=D)×HT 14	0.3784	0.0359	10.55	<.0001
(EA=D)×HT 15=16=17=18=19	0.4722	0.0377	12.53	<.0001

Source: EU-SILC 2021, own processing in SAS EG.

We followed a similar procedure for the other two economic activity statuses (U – unemployed and E – employed), both for Slovakia and Czechia. For the analyzed three economic activity statuses (D, U, E), the resulting clusters are presented separately for Slovakia and Czechia in Table 9. For each cluster in the table, the estimated mean of WI is calculated along with its standard error (SE) and *p*-value for the test of equality of marginal means of WI across the household types included in the cluster.

Table 9: Estimates of the marginal means of work intensity in clusters of household types for disabled (D), unemployed (U), and employed (E) in Slovakia and Czechia.

EA	SK				CZ		
	Cluster	**Estimate (SE) [p-value]**	**HT**		**Cluster**	**Estimate (SE) [p-value]**	**HT**
D	**1**	0.085 (0.038) [0.6953]	1A_0Ch 1A_1⁺Ch 2A(1⁺R)		**1**	0.120 (0.035) [0.2494]	1A_0Ch 1A_1⁺Ch 2A(1⁺R)
	2	0.378 (0.036) [-]	2A_0Ch		**2**	0.300 (0.045) [0.5916]	2A_1Ch 2A_3⁺Ch
	3	0.472 (0.038) [0.4695]	2A_1Ch 2A_3⁺Ch Other_0Ch Other_1⁺Ch 2A_2Ch		**3**	0.404 (0.036) [0.7360]	2A_0Ch Other_0Ch 2A_2Ch
	-	-	-		**4**	0.606 (0.041) [-]	Other_1⁺Ch

Table 9: Estimates of the marginal means of work intensity in clusters of household types for disabled (D), unemployed (U), and employed (E) in Slovakia and Czechia (cont.)

EA	SK				CZ		
	Cluster	**Estimate (SE) [p-value]**	**HT**		**Cluster**	**Estimate (SE) [p-value]**	**HT**
U	**1**	0.210 (0.043) [0.5453]	1A_1^{+}Ch 2A(1^{+}R)		**1**	0.196 (0.078) [-]	2A(1^{+}R)
	2	0.348 (0.025) [0.4293]	1A_0Ch 2A_0Ch 2A_3^{+}Ch Other_1^{+}Ch		**2**	0.325 (0.053) [0.9713]	1A_0Ch 1A_1^{+}Ch 2A_3^{+}Ch
	3	0.440 (0.028) [0.1637]	2A_1Ch Other_0Ch 2A_2Ch		**3**	0.476 (0.048) [0.2734]	2A_0Ch 2A_1Ch Other_0Ch Other_1^{+}Ch 2A_2Ch
E	**1**	0.731 (0.012) [0.7826]	2A_3^{+}Ch Other_1^{+}Ch		**1**	0.818 (0.012) [-]	2A_3^{+}Ch
					2	0.853 (0.010) [-]	Other_1^{+}Ch
	2	0.793 (0.010) [0.8761]	2A_1Ch Other_0Ch 2A_2Ch		**3**	0.875 (0.007) [0.1499]	2A_1Ch 2A_2Ch
					4	0.906 (0.009) [-]	Other_0Ch
	3	0.821 (0.012) [0.7325]	1A_1^{+}Ch 2A_0Ch		**5**	0.925 (0.008) [0.6878]	1A_1^{+}Ch 2A_0Ch
	4	0.890 (0.013) [0.4478]	1A_0Ch 2A(1^{+}R)		**6**	0.955 (0.009) [0.3364]	1A_0Ch 2A(1^{+}R)

Source: EU-SILC 2021, own processing in SAS EG.

For the disabled, we revealed the lowest WI in the cluster of household types 1A_0Ch, 1A_1^{+}Ch, and 2A(1^{+}R), which corresponds to households with at most 1 adult in the productive age. The identical cluster emerged for both countries. In Czechia (11.96%), we estimated the respective marginal mean to be almost 3.5 p.p. higher than in Slovakia. However, in both countries, it was at a very low level (below the 20%), which identifies QJ. If a disabled person lives in a household with a higher number of adults, the risk of QJ naturally decreases, as confirmed by our analyses. While in Czechia, households with 2 adults and a disabled person typically have low WI, in Slovakia, this is slightly higher. The 3rd cluster for disabled persons in Slovakia, which includes households with 2 adults with children, has the mean of WI at the middle level (see also Figure 1). For the disabled, the mean of WI is convincingly high only for households of type "other" with at least 1 child (Other_1^{+}Ch), and only in Czechia. Our analysis indicated that the disabled in Czechia generally live

in households with lower WI. This finding aligns with official statistics, which state that in 2021, the QJ rate for the disabled was 16.7% in Czechia and 11.3% in Slovakia, despite the fact that the overall QJ rate in the population was lower in Czechia (2.5%) than in Slovakia (2.7%).

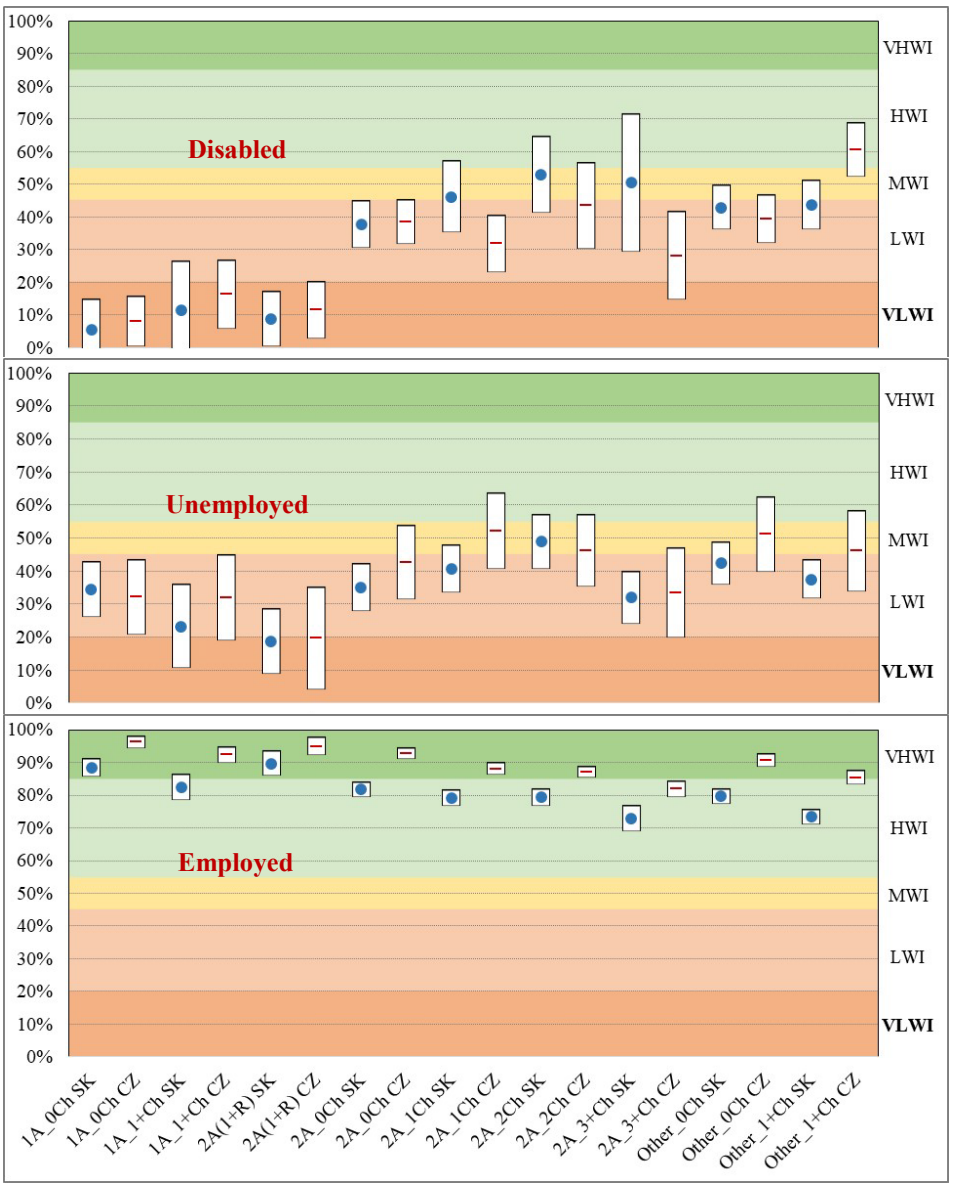


Figure 1: Interval estimates (95%) of LS-Means of WI for *EA*×*HT* interaction (for the disabled at the top, unemployed in the middle, employed at the bottom)

Source: EU-SILC 2021, own processing in SAS EG.

The better situation in the entire population in Czechia compared to Slovakia is confirmed by a significantly higher mean of WI for employed persons. Employed individuals live in households with the highest WI when it comes to childless households with at most one adult in the productive age (1A_0Ch and 2A(1$^+$R)), for which we estimated a very high WI. While in Slovakia it was almost 89%, in Czechia it was 6.5 p.p. higher. For employed persons, the mean of WI in all corresponding clusters in Slovakia was significantly lower than in Czechia. However, even in households with the lowest WI (2A_3$^+$Ch and Other_1$^+$Ch), the mean WI for employed persons was at least at a high level (HWI). The composition of clusters of household types for employed persons was very similar in Slovakia and Czechia (see Table 9).

For unemployed persons, the similarity between Slovakia and Czechia in the composition of clusters was lower. Also for this reason, we provide point and interval estimates of the mean of WI in graphical form for individual household types rather than for their clusters (Figure 1). Unemployed persons faced the highest risk of QJ when they lived in households with 2 adults, at least 1 of whom was 65$^+$. We estimated the mean of their WI to be on the borderline between VLWI and LWI. Among the other household types, those with a single adult were the most at risk in both countries, similar to the case with the disabled (1A_0Ch and 1A_1$^+$Ch). For these households, LWI was typical. However, this degree of WI was characteristic for most household types for unemployed individuals both in Slovakia and in Czechia. An unemployed person from a household with 2 adults and 2 children mostly lived in a household MWI. For unemployed persons, the MWI was also typical for Czech households of 2 adults and 1 child, 2 adults without children and other types of households without or with children (Other_0Ch, Other_1$^+$Ch).

5.2.2. Analysis of the impact of interaction *EA*×*Education*

Not surprisingly, higher WI is associated with higher education, but this does not apply to all statuses of economic activity (Figure 2).

In 2021, in both countries, the education of a disabled individual did not have a significant impact on WI of households, as confirmed by Figure 2 and the non-significant parameters (Table 5). However, this does not hold true for the impact of education on unemployed person. This influence is significant in both countries, but it is greater in Slovakia. For Slovakia, we found that while low-educated unemployed persons mostly live in QJ households, for unemployed persons with tertiary education, households with a medium WI are characteristic. In Czechia, even low-educated unemployed persons mostly do not live in QJ households because their households achieve a low (and not very low) level of WI. For unemployed persons with higher education (ISCED 3-5 or ISCED 6-8), households in Czechia are characterized by WI at the border between a low (LWI) and a medium level (MWI).

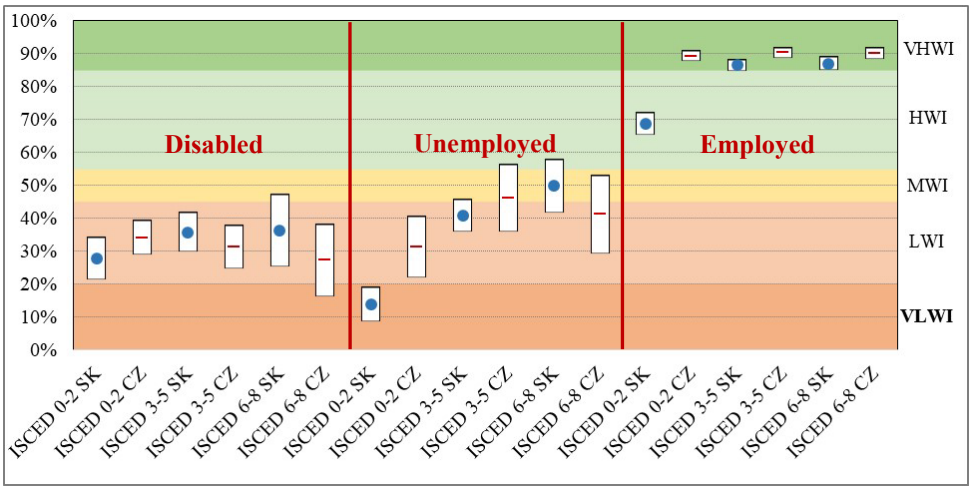


Figure 2: Interval estimates (95%) of LS-Means of WI for *EA× Education* interaction

Source: EU-SILC 2021, own processing in SAS EG.

As unemployed individuals with higher education (at least ISCED 3) have a mean WI at a low to medium level, they appear to experience shorter periods of unemployment and/or have another adult in the household (spouse or partner) who is employed. However, the status of economic activity "unemployed" have such a significant effect on WI that education cannot sufficiently compensate for this and therefore persons who are unemployed and whose education is tertiary generally live in households with significantly lower WI as employed persons. This applies even to the employed persons in Slovakia with ISCED 0-2 education, for whom the average WI is "only" at a high level (HWI).

5.2.3. Analysis of the impact of interaction *EA×Age*

In 2021, disabled and unemployed persons aged under 30 lived in households with higher WI compared to older people (Figure 3). In this context, we are referring to WI levels bordering between low and medium, and in the case of disabled individuals in Slovakia and the unemployed in Czechia, reaching up to the high level. For disabled persons and the unemployed in age categories above 30 years (age categories 30–40, 40–50, 50^{+}), households with a low level of WI were typical, and there were no significant differences among these age categories.

When specifically examining disabled and unemployed persons, there were no significant differences in WI between Slovakia and Czechia within individual age categories. However, this is not the case for the employed. For this economic activity status, we found demonstrably higher WI in Czechia than in Slovakia in all age categories.

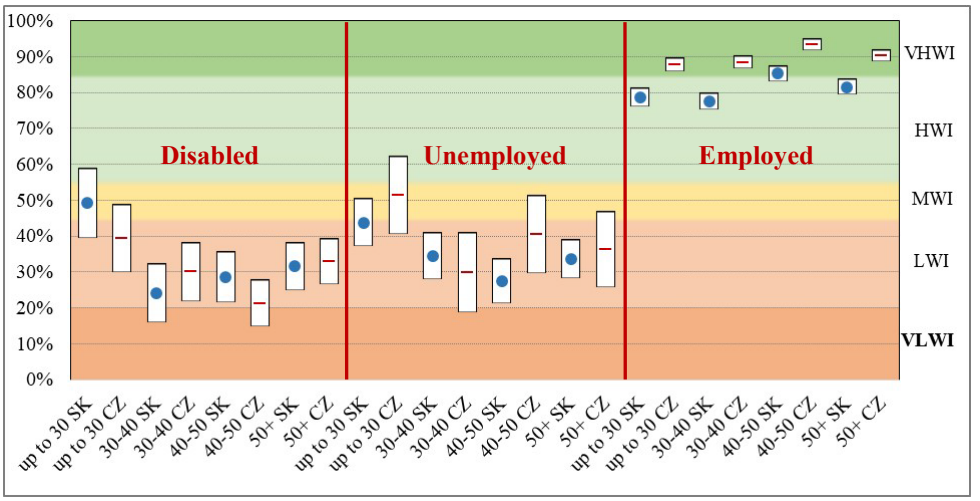


Figure 3: Interval estimates (95%) of LS-Means of WI for *EA*×*Age* interaction

Source: EU-SILC 2021, own processing in SAS EG.

6. Conclusion and discussion

The type of household a person lives in, their education, age, gender, health conditions, marital status, as well as the degree of urbanization in their residential area are factors that significantly influenced WI in Slovakia and Czechia in 2021, but this impact varied across different economic activity statuses. These factors are considered very important by other researchers (Watson et al. 2015; Horemans et al. 2018; Verbunt and Guio 2019) when assessing employment/unemployment and atypical employment in relation to social exclusion.

The general linear models, in which we considered the interactions of these factors with the status of economic activity, explained the variability of WI to about 53%. The status of economic activity, household type, education, and age participated the most in this explained variability. In this paper, we focused on assessing the net impact (i.e. by fixing the influence of other relevant factors) of household type, education and age of the person on WI of the household, considering only employed, unemployed and disabled persons and abstracting from other statuses. The last two mentioned groups of persons belong among the most vulnerable groups in terms of WI, as well as in terms of social exclusion. In addition to economic activity, van der Zwan and de Beer (2021) also used the above factors (age, education, and household type) as control variables, applying them to the assessment of employment for disability and non-disability persons.

The paper provides estimates of the marginal means of WI households for employed, unemployed and disabled persons with different education, age and from

different types of households in Slovakia and in Czechia. Given the large number of household types, we identified, for each economic activity status, between which types of households there are no significant differences from the perspective of WI. On the basis of simultaneous testing of WI marginal means, clusters of household types were created that were largely similar, but not identical, in Slovakia and Czechia.

Our analyses confirmed that in 2021, in Czechia and Slovakia, disabled and unemployed persons lived in households with significantly lower WI compared to the employed, which is natural based on their main economic activity status. According to Calegari et al. (2022) households with disabled members experience a greater risk of social exclusion compared to households without disabled members. Our research even showed that in 2021 in Czechia and Slovakia households with one adult person (with or without children) had demonstrably lower WI if it concerned person with a disability (very low WI) compared to an unemployed person (low WI). The WI mean for disabled persons can be assumed to be around 50% (medium WI) only exceptionally and applies to persons under the age of 30 years or to persons living in complete households, while these findings concern only Slovakia and not Czechia.

Following the above finding, let us focus on the effect of age. For the disabled in Czechia, the age category "under the age of 30 years" is also the least risky, although the WI mean was approximately 10 p.p. lower compared to Slovakia, corresponding to low WI. This age category is also the least risky for unemployed persons. Conversely, for persons with an employed status, the age category under the age of 30 years along with the 30–40 category exhibited the lowest WI. However, in the case of employed persons, we are discussing high (Slovakia) or very high (Czechia) WI.

While WI increases with higher education levels, this trend does not hold for persons with disabilities. In 2021, in Slovakia, both unemployed and employed persons with low education lived in households that had WI lower by one level compared to individuals with higher education. In Slovakia, low-educated unemployed persons generally showed QJ. In the case of Czechia, the impact of education on WI was significantly lower, where low-educated unemployed persons lived in households with low WI. In Czechia, employed individuals at every educational level generally lived in households with very high WI. In Slovakia, this WI level was achieved by employed persons with education at the ISCED 3–5 and ISCED 6–8 level. However, those employed with low education (ISCED 0-2) lived in households that utilized their work potential at around 70% (high WI). Education has been identified as a risk factor for poverty and social exclusion in other EU countries as well, as demonstrated in studies by authors such as Dudek and Szczesny (2021), Filandri and Struffolino (2019), and Hofmarcher (2021). Our findings indicate that, in terms of WI, education plays a crucial role for unemployed persons, but not for disabled persons. This finding supports the conclusion of Mussida and Sciulli (2024), who state that higher education is

a protective factor against poverty, in addition to being positively associated with work intensity/labor market participation, while it exerts a less clear role on disability.

In terms of household type, disabled and unemployed persons in both countries faced the highest risk of QJ in households of 2 adults, at least 1 aged 65^+ and in incomplete households such as single-person households and households of 1 adult with at least 1 child. However, this conclusion does not hold for employed individuals, because if they lived in the above 3 household types [1A_0Ch, 2A(1^+R), 1A_1^+Ch], then on average, their households achieved the highest WI (very high WI). The only exception was the household type of 1 adult with at least 1 child in Slovakia, for which high WI was typical. For employed persons in Slovakia, the average WI of other household types was also at a high level, with the lowest (73%) observed for households of 2 adults with at least 3 children. This type of household also recorded the lowest average WI in Czechia (82% for employed persons). For other household types with employed persons in Czechia, we estimated very high WI.

If we exclude the most risky household types [1A_0Ch, 2A(1^+R), 1A_1^+Ch] for the unemployed and disabled statuses, the remaining household types have either low or medium WI, which is valid for both countries. If unemployed persons lived in complete households, they had the lowest WI (32% in Slovakia, 33% in Czechia), if it was a household of 2 adults with at least 3 children. This type of household was also the riskiest of the complete households for disabled in Czechia (WI of 28%). Disabled persons living in complete households with children had a significantly higher WI mean in Slovakia compared to Czechia, reaching a medium level.

To summarize the previous findings for the economic activity status "employed", it can be concluded that individuals with education at the ISCED 0-2 level, within the age categories of up to 30 years and 30–40 years, and those living in households with 2 adults and at least 3 children are the most at-risk based on the WI perspective. This is consistent with the conclusions of Filandri and Struffolino (2019), who found that risk factors in terms of in-work poverty include young age, a low level of education, and households with a high number of children. Their findings about the riskiness of households with a small number of earners can be confirmed in our study only for disabled and unemployed persons.

Our analyses also confirmed that households in Slovakia have, on average, lower WI than in Czechia. This phenomenon was confirmed for employed persons in all household types, as well as across all educational groups and age categories. There are a few exceptions in the case of unemployed persons, such as for education at the ISCED 6–8 level, where our analyses did not confirm this phenomenon. In the case of disabled persons, we cannot confirm the mentioned phenomenon at all, as in many compared categories, we observed comparable results for both countries and in certain categories, such as complete households with children, we identified a significantly higher WI

in Slovakia than in Czechia. However, our conclusions align with the findings of Eurostat (2023b), which indicate that people in Czechia with a disability have a relatively low probability of living in QJ households but a difference between people with a disability and people with no disability is large since in 2021 Czechia had the largest relative gap.

We are convinced that the paper fills a gap in research on poverty and social exclusion that mostly focuses on income poverty and material deprivation but rarely on QJ and even rarer on WI as continuous variable or on all levels of WI. Our research provides only a partial answer to the question of which population groups to target with social measures. This is because WI is only one of several aspects that determine social exclusion and although it is correlated with other dimensions (income poverty, material deprivation), it does not have a complete overlap with them. The form and intensity of social support are also a question, as demonstrated by Lehwess-Litzmann and Nicaise (2020), who found that the generosity of social benefits is negatively connected to the speed at which households increase their WI. A topic for further research is whether this relationship holds equally for households with unemployed members as well as for households with disabled members.

Between Slovakia and Czechia, we have uncovered some differences in WI but there were numerous similarities that could stem from the shared historical development and geopolitical and cultural similarities of these countries. While we hold the belief that many of these conclusions may be applicable to CEE countries, further research is necessary to verify this assertion.

Acknowledgements

This work was supported by the VEGA project The Impact of Inflation on Poverty and Social Exclusion in Slovakia and the EU (No. 1/0285/24).

References

Agresti, A., (2015). Foundations of linear and generalized linear models. John Wiley & Sons.

Cai, W., (2014). Making comparisons fair: how LS-means unify the analysis of linear models. *SAS Institute Inc. Paper. SA*, S060-2014. [online] https://support.sas.com/resources/papers/proceedings14/SAS060-2014.pdf. [Accessed on 18 October 2023].

Calegari, E., Fabrizi, E. and Mussida, C., (2022). Disability and work intensity in Italian households. *Review of Economics of the Household*, pp. 1–20. https://doi.org/10.1007/s11150-021-09559-6.

Dean, A., Voss, D. and Draguljić, D., (2017). Design and Analysis of Experiments. Springer, New York.

Dudek, H., Szczesny, W., (2021). Multidimensional material deprivation in Poland: a focus on changes in 2015–2017. *Quality & Quantity*, 55(2), pp. 741–763. https://doi.org/10.1007/s11135-020-01024-3.

Duiella, M., Turrini, A., (2014). Poverty developments in the EU after the crisis: a look at main drivers. *ECFIN Economic Brief*, 31, pp. 1-10. https://doi.org/10.2765/72447

Elswick Jr, R. K., Gennings, C., Chinchilli, V. M. and Dawson, K. S., (1991). A simple approach for finding estimable functions in linear models. *The American Statistician*, 45(1), pp. 51–53.

Eurostat, (2023a). Risk of poverty decreases as work intensity increases. [online] https://ec.europa.eu/eurostat/web/products-eurostat-news/w/ddn-20230227-1. [Accessed on 27 October 2023].

Eurostat, (2023b). Disability statistics - poverty and income inequalities. [online] https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Disability_statistics_-_poverty_and_income_inequalities#At_risk_of_poverty_or_social_exclusion. [Accessed on 27 October 2023].

European Commission, (2021). Methodological guidelines and description of EU-SILC target variables. 2021 operation (Version 4_09/12/2020). [online] https://circabc.europa.eu/sd/a/f8853fb3-58b3-43ce-b4c6-a81fe68f2e50/Methodological%20guidelines%202021%20operation%20v4%2009.12.2020.pdf. [Accessed on 20 October 2023].

Fabrizi, E., Mussida, C., (2020). Assessing poverty persistence in households with children. *Journal of Economic Inequality*, 18(4), pp. 551–569. https://doi.org/10.1007/ s10888-020-09455-6.

Filandri, M., Struffolino, E., (2019). Individual and household in-work poverty in Europe: understanding the role of labor market characteristics. *European Societies*, 21(1), pp. 130–157. https://doi.org/10.1080/14616696.2018.1536800.

García-Gómez, C., Pérez, A. and Prieto-Alaiz, M., (2021). Copula-based analysis of multivariate dependence patterns between dimensions of poverty in Europe. *Review of Income and Wealth*, 67(1), pp. 165–195. https://doi.org/10.1111/roiw.12461.

Gerbery, D., Miklošovic, T., (2020). Labour Market Transitions and Their Determinants in Slovakia: Path from Crisis to Recovery 1. *Ekonomicky Casopis*, 68(7), pp. 651–676. https://doi.org/10.1111/roiw.12461.

Guio, A. C., Vandenbroucke, F., (2019). Poverty and child deprivation in Belgium. A comparison of risk factors in the three regions and neighbouring countries. (March 1, 2019). *King Baudouin Foundation*.

Hallaert, J. J., Vassileva, I. and Chen, T., (2023). Rising Child Poverty in Europe: Mitigating the Scarring from the COVID-19 Pandemic. *International Monetary Fund*, Working Paper No. WP/2023/134.

Hofmarcher, T., (2021). The effect of education on poverty: A European perspective. Economics of Education Review, 83, 102124. https://doi.org/10.1016/j.econedurev.2021.102124.

Horemans, J., (2016). Polarisation of non-standard employment in Europe: Exploring a missing piece of the inequality puzzle. *Social Indicators Research*, 125, pp. 171–189. https://doi.org/10.1007/s11205-014-0834-0.

Horemans, J., Lohmann, H. and Marx, I., (2018). Atypical employment and in-work poverty. *Handbook on in-work poverty*, pp. 146–170.

Kis, B. A., Gábos, A., (2016). Consistent poverty across the EU. *Corvinus Journal of Sociology and Social Policy*, 7(2), pp. 3–27. https://doi.org/10.14267/CJSSP.2016.02.01

Kuznetsova, A., Brockhoff, P. B. and Christensen, R. H. B., (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, 82(13). https://doi.org/10.18637/jss.v082.i13.

Lehwess-Litzmann, R. and Nicaise, I., (2020). Surprisingly small: Effects of "generous" social benefits on re-employment of (quasi-) jobless households. *Journal of International and Comparative Social Policy*, 36(1), pp. 76–91. https://doi.org/10.1017/ics.2020.1

Lenth, R. V., (2016). Least-squares means: the R package lsmeans. *Journal of statistical software*, 69, pp. 1-33. https://doi.org/10.18637/jss.v069.i01.

Littell, R. C., Stroup, W. W. and Freund, R. J., (2010). SAS for Linear Models. 4th ed. Cary, NC: SAS Institute Inc.

McFarquhar, M., (2016). Testable hypotheses for unbalanced neuroimaging data. Frontiers in neuroscience, 10, 270. https://doi.org/10.3389/fnins.2016.00270.

Mussida, C., Sciulli, D., (2024). Poverty, work intensity, and disability: evidence from European countries. *The European Journal of Health Economics*, pp. 1–20. https://doi.org/10.1007/s10198-024-01679-x.

Nieuwenhuis, R., Maldonado, L., (2018). The triple bind of single-parent families: Resources, employment and policies to improve well-being. Policy Press. https://doi.org/10.2307/j.ctt2204rvq.

Regan, M., Maître, B., (2020). Child poverty in Ireland and the pandemic recession (No. 2021/4). Budget Perspectives. https://doi.org/10.26504/bp202104.

SAS Institute Inc., (1997). SAS Technical Report R-103, Least-squares means in the fixed-effects general models. NC: SAS Institute Inc.

SAS Institute Inc., (2018). SAS/STAT® 15.1 User's Guide. The GLM Procedure. Cary SAS/STAT® 15.1, NC: SAS Institute Inc.

Schad, D. J., Vasishth, S., Hohenstein, S. and Kliegl, R., (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of memory and language*, 110, 104038. https://doi.org/10.1016/j.jml.2019.104038.

Searle, S. R., Gruber, M. H. J., (2017). Linear Models, 2nd ed., John Wiley & Sons.

Searle, S. R., Speed, F. M. and Milliken, G. A., (1980). Population marginal means in the linear model: an alternative to least squares means. *The American Statistician*, 34(4), pp. 216–221.

Suzuki, M., Taniguchi, T., Furihata, R., Yoshita, K., Arai, Y., Yoshiike, N. and Uchiyama, M., (2019). Seasonal changes in sleep duration and sleep problems: A prospective study in Japanese community residents. *PLoS One*, 14(4), e0215345. https://doi.org/10.1371/journal.pone.0215345.

Tabachnick, B. G., Fidell, L. S. and Ullman, J. B., (2013). *Using multivariate statistics*, Vol. 6, pp. 497–516. Boston, MA: pearson.

Treanor, M. C., (2018). Income poverty, material deprivation and lone parenthood. *The triple bind of single-parent families*, pp. 81-100. https://doi.org/10.2307/j.ctt2204rvq.10.

Treanor, M., Troncoso, P., (2022). Poverty, parental work intensity and child emotional and conduct problems. *Social Science & Medicine*, 312, p. 115373. https://doi.org/10.1016/j.socscimed.2022.115373.

van der Zwan, R., de Beer, P., (2021). The disability employment gap in European countries: What is the role of labour market policy?. *Journal of European Social Policy*, 31(4), pp. 473–486. https://doi.org/10.1177/09589287211002435.

Verbunt, P., Guio, A. C., (2019). Explaining differences within and between countries in the risk of income poverty and severe material deprivation: Comparing single

and multilevel analyses. *Social Indicators Research*, 144, pp. 827–868. https://doi.org/10.1007/s11205-018-2021-1.

Wang, B., Wu, P., Kwan, B., Tu, M. X. and Feng, Ch., (2018). Simpson's paradox: examples. *Shanghai archives of psychiatry*, 30(2), p. 139. https://doi.org/10.11919/j.issn.1002-0829.218026.

Watson, D., Maître, B. and Russell, H., (2015). Transitions into and out of Household Joblessness, 2004 to 2014 (No. 5). ESRI Social Inclusion Report.

Westfall, P. H., Tobias, R. D., (2007). Multiple testing of general contrasts: Truncated closure and the extended Shaffer–Royen method. *Journal of the American Statistical Association*, 102(478), pp. 487–494. https://doi.org/10.1198/016214506000001338.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 27–50, https://doi.org/10.59139/stattrans-2024-002
Received – 05.10.2023; accepted – 26.08.2024

The impact of the COVID-19 pandemic on the financial situation of people aged 50+ based on SHARE data

Tomasz Panek[1], **Jan Zwierzchowski**[2], **Jan Kroszka**[3]

Abstract

The COVID-19 pandemic led to the lockdown of economies, resulting in the deterioration of the financial situation of numerous households. To support economies and societies, governments implemented various measures involving job protection and financial support. This study aims to assess the changes in the financial situation of households of people aged 50+ during the pandemic. We evaluate the outcome of the introduced national policies, the EU countries' economic performance, labor market conditions and the individual characteristics of the financial situation of the members of the examined households. To achieve this goal, an original synthetic index was constructed to measure the changes in the overall financial situation of the surveyed group of households. This index combines various indicators, including income, subjective income assessment, the use of savings to finance current consumption and the postponement of bill payments, allowing a comprehensive evaluation of the shifts in the financial status of the 50+ population during the pandemic. Additionally, the study aims to examine how the age of respondents is interlinked with the changes in their financial situation.

Data from the Survey of Health, Ageing, and Retirement in Europe (SHARE), including the SHARE Corona Telephone Survey conducted during the first and the second wave of the COVID-19 pandemic were utilized for the analysis. The study's findings show that during the pandemic, the changes in the financial situation of households with people aged 50+ varied across the selected countries. Furthermore, they reveal that both the response to the consequences of the COVID-19 pandemic of a given country and its overall development level, as well as the characteristics of the respondents had a diverse impact on the financial situation and ability to cope with the economic risks faced by individuals aged 50+ during the studied period. These findings can serve as a basis for the design of targeted government interventions aiming to mitigate the negative impact of the pandemic on the material situation of this vulnerable age group.

Key words: COVID-19 pandemic, financial situation changes, older population, SHARE.

[1] Warsaw School of Economics, Poland. E-mail: tompa@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-1034-7222.

[2] Warsaw School of Economics, Poland. E-mail: jzwier@sgh.waw.pl. ORCID: https://orcid.org/0000-0003-3355-1290.

[3] Warsaw School of Economics, Poland. E-mail: jk82095@student.sgh.waw.pl. ORCID: https://orcid.org/0009-0009-9389-5248.

1. Introduction

The COVID-19 pandemic has had a profound impact on societies across the world (Mahler et al., 2022; Castaneda Aguilaret et al., 2022), particularly on those aged 50 or over. In 2020, governments implemented solutions to prevent the spread of the virus, leading to extensive measures being taken, including lockdowns and restrictions on economic activity (Bastini & Stoevska, 2021). The labor market suffered, and many economically active people experienced a reduction in income (Fana et al., 2020; Weber, 2021). The situation was dire for households with limited savings. These households were often forced to make significant reductions in their expenditures, such as postponing rent, loan, and credit repayments, and prioritizing other obligations. To mitigate the negative effects of economic restrictions states introduced various protective measures aimed at companies, local governments, and individuals (Gentilini, 2020; Baptista et al., 2021; Cantillon et al., 2021; Echebarria Fernández, 2021).

After a year, the pandemic continued to impact societies worldwide, leading to further waves of the virus and ongoing government action, including restrictions on social contacts and labor market activities. The financial situation of many households continued to deteriorate. Meanwhile, states have continued various protection action to combat the fall in household income (Abramowska-Kmon et al., 2023).

European countries were affected differently by the COVID-19 pandemic, and they introduced different policies (Baptista et al., 2021; ESPN, 2021; Altiparmakis et al., 2021). The impact of these policies and protective measures on the financial situation of households with people aged 50 or over varied. In this article, we assess changes in the financial situation of this group of households[4] in selected European Union countries during the pandemic. Our results are based on the responses from the SHARE Corona 1 and Corona 2 Telephone Surveys that were conducted during the first and second waves of the pandemic (Abramowska-Kmon et al., 2023). The use of harmonized data from the SHARE Corona Telephone Surveys, along with longitudinal information from earlier SHARE Wave 7 (before the COVID-19 outbreak), offers a unique opportunity to assess and interpret variations in the financial situation of households during the pandemic across European countries with different welfare and pension systems. The obtained data reveal the existence of significant variations between countries. The study examines not only the changes in the financial situation of individuals aged 50 or more in European countries during the pandemic but also the influence of country characteristics, such as the magnitude of the pandemic's impact on the economy, the policies introduced in response, and the extent of financial assistance

[4] In this article, whenever we refer to "households," we are s referring to households of individuals aged 50 and older.

supporting the economy and population, as well as the influence of the characteristics of the surveyed individuals. Analyses were conducted for 7 countries representing the following four groups of European countries, with different social care and pension systems, which are relevant to explaining these changes: the Nordic group includes Sweden, the Continental group, which consists of France and Germany, the Mediterranean group, which includes Greece and Spain, the Eastern European group includes Czechia and Poland.

Our contribution to the literature is that we provide cross-country comparisons based on internationally harmonized data, analyzing both particular aspects of the financial situation of older Europeans during the COVID-19 pandemic and changes in this situation due to all aspects jointly. A significant part of our contribution is the analysis of changes in this situation during the second phase of the pandemic in the context of the impact of both country and individual characteristics. Additionally, we examine how the age of respondents is linked with the changes in their financial situation. This study is structured as follows: Section 2 is dedicated to a review of the most recent literature, Section 3 presents main indicators showing differences in country policies, economic performance, and labor market situation during the pandemic, Section 4 discusses the data and methods used in the analyses, and the results, Section 5 presents the discussion of the results and conclusions. The article includes an attachment with definitions of variables used in the empirical analysis.

2. The impact of the COVID-19 pandemic on the financial situation of people 50+: A literature review

The COVID-19 pandemic impact on the economic situation of older adults has received considerable attention in economic research. Several studies have examined the consequences of the pandemic for their situation on a job market. A study conducted by the OECD (OECD, 2020) found that the pandemic has disproportionately affected older workers, with job losses being more concentrated among those aged 50 or more.

Other studies also documented an increase in the unemployment rate among older workers. Celbiş et al. (2023) used machine learning methods to identify groups of older workers particularly vulnerable to job losses. They showed the importance of individual-level factors, such as education, sector of job market and health status. Another study by Theodoropoulos and Voucharas (2022) analyzed the effects of stringency measures on job losses of older workers. They found that stricter government policies were followed by higher number of lay-offs among people 50+, especially those low-educated, older and with health problems. However, authors concluded that it was only a short-term effect.

The pandemic influenced retirement decisions of older workers as well, which was shown by the study of European Central Bank (Botelho, Weißler, 2022). The authors argue that around 175 000 of workers were affected by early retirement due to pandemic. It concerned mostly those in relatively poorer health conditions.

Other studies analyzed the impact of COVID-19 pandemic on financial situation of households. Almeida et al. (2021) used economic forecasts of European Commission to assess this impact. They found that the worsening of income situation may be especially strong for households from lower part of income distribution – older workers are relatively often in this group. This study also indicates that policy interventions of government may have played an important role in reducing the degree of income losses.

Clark et al. (2022) assessed the changes of disposable income of households in 5 European countries between 2020 and 2021 based on COME-HERE survey. Their analysis shows that stringent policies negatively affected the financial situation of such groups as women, young people, and poor workers. However, they proved to be beneficial for retirees.

Di Pietro (2022) analyzed changes in household income during pandemic in Italy. The study was based on panel data from survey conducted by the Bank of Italy in six waves after the outbreak of COVID-19. The author found that income reductions occurred mostly during the first wave of the pandemic. Moreover, these decreases were not correlated with the number of people in the household or with a region of the country.

Several studies have used SHARE datasets to examine the impact of the pandemic on the income situation of people aged 50+ (Celbiş et al., 2023; Theodoropoulos and Voucharas, 2022; Botelho, Weißler, 2022). Chłoń-Domińczak and Holzer-Żelażewska (2022) based their study on data from SHARE Corona Telephone Survey which was conducted in June 2020. They found that the pandemic had a significant negative impact on the financial situation of older adults across European countries, particularly for those who were already economically vulnerable before the pandemic. Their study also showed that more stringent country policies during pandemic led to increased financial difficulties of older adults. However, individual characteristics of respondents – sex, age, household size, educational attainment and economic status – were more important for explaining their financial situation than country ones (including policies).

Bonfatti et al. (2021) showed that individuals below retirement age were particularly affected by COVID-19 crisis, which indicates that social security systems efficiently protected those already retired. They also pointed out to the importance of education and income before the pandemic. The study also found an increase in economic inequalities during the pandemic.

Finally, Rajevska et al. (2022) examined the financial consequences of COVID-19 for older workers in the Baltic States. They showed that the share of people 50+ receiving financial support due to the pandemic differed between analyzed countries. However, their analysis indicates that the labor market participation of older people was not affected by the crisis in 2020.

3. Country policies, economic, and labor market performance during the COVID-19 pandemic

The financial situation of people aged 50+ during the pandemic should be viewed in the context of the economic and labor market changes that occurred during that period. From mid-March 2020, most European countries introduced strict policies, including school closures, workplace closures, and travel bans. Meanwhile, states introduced various protective measures aimed at companies, local governments, and individuals to mitigate the decline in household income. During the second phase of the pandemic, governments continued implementing epidemic restrictions as well as various protective actions. The introduced policies varied across countries, as did the impact of the crisis on national economies.

In the analyses, we capture COVID-19 restriction policy responses by using the Stringency Index. The Stringency Index is a composite measure based on nine response indicators, including: school closures; workplace closures; cancellation of public events; restrictions on public gatherings; closures of public transport; stay-at-home requirements; public information campaigns; restrictions on internal movements; and international travel controls, mean number of confirmed deaths, and duration of stay-at-home requirements (Hale et al. 2021). The sub-indicators were scored and aggregated into a Stringency Index, which reflects the overall government response to the pandemic on a daily basis, with a value between 0 and 100. Notably, in August 2020, the Stringency Index reached its highest values in Spain and Greece (>60), while its lowest values (<40), were recorded in Germany, Czechia and Poland. During the second wave of the pandemic, all surveyed countries except for Germany, Czechia, and Greece, relaxed their pandemic-related restrictions. The largest decreases in the index values were observed in Sweden and Spain, with reductions of 24.8 and 14.8, respectively.

The Economic Support Index measures the government's income support policy for citizens during times of crisis and is composed of two indicators: government income support and household projected debt or contracted relief (Hale et al., 2022). The index is expressed in simply summed scores of the underlying sub-indicators, rescaled to a range from 0 to 100. In August 2020, the Economic Support Index reached values of above 87 in Spain and Greece, while the lowest values were observed in Poland and Germany, slightly above 37. Between August 2020 and August 2021, the Economic

Support Index in Poland and Czechia increased considerably by 25 and 12.5 points, respectively, indicating increased economic support from their governments. In France, however, the index dropped by 25 points, suggesting a decrease in the level of support for households during the second phase of the pandemic. In other countries, the value of this index remained the same.

Government policies prompted different reactions in the economies of the countries studied. Table 1 presents changes in the size of the economy and the labor market in both waves of the pandemic. During the first wave of the pandemic, most of the studied countries experienced GDP declines. However, France stands out as an exception, with a GDP increase of 4.9%. This may suggest the effectiveness of their remedial measures or other factors impacting the economy during that specific period. During the second wave of the pandemic, all analyzed countries experienced a recovery and economic growth. Interestingly Greece recorded one of the most significant GDP declines in the first wave, it posted the highest economic growth during the second wave. This could indicate the country's strong ability to bounce back after a crisis or the effectiveness of economic policies introduced. However, the impact of these indices on GDP changes is more complex and may also be influenced by other factors, such as the structure of the economy, dominant sectors, and global trade conditions.

Table 1: Changes in country characteristics during the pandemic

Country characteristics	Changes in country characteristics						
	Poland	Czechia	Sweden	Germany	France	Spain	Greece
Change in Unemployment Rate (in percentage points)							
Q3 2020-Q1 2020	0.4	0.8	1.9	0.7	0.9	2	0.3
Q3 2021-Q3 2020	-0.3	-0.1	-0.6	-0.6	-0.9	-1.7	-3.1
GDP change (%)							
Q3 2020/Q1 2020	-2.4	-1.8	-0.8	-0.5	4.9	-1.2	-8.1
Q3 2021/Q3 2020	7.1	3.8	4.3	1.8	3.6	4.2	12.4
Stringency Index							
August 2020	39.81	36.11	59.26	56.94	48.15	62.5	61.11
August 2021 - August 2020	-3.88	0.85	-24.79	2.32	-1.94	-14.81	1.46
Economic Support Index							
August 2020	37.5	62.5	62.5	37.5	75	87.5	87.5
August 2021 - August 2020	25	12.5	0	0	-25	0	0

Source: own work based on Eurostat data (unemployment rate, GDP) and Hale et al., 2021 (Economic Support Index, Strigency Index).

During the first wave of the pandemic, an increase in the unemployment rate was observed in all studied countries. The largest increase was recorded in Spain and Sweden, while the smallest increase was observed in Greece. During the second wave of the pandemic, a decrease in the unemployment rate was observed in all studied countries. Interestingly, Greece recorded the most significant drop.

Both analyzed indices did not influence the labor market in a uniform way across countries during both pandemic waves. However, countries which effectively implemented and adjusted their support policies were more effective in safeguarding their labor market. Nevertheless, the dynamics of the unemployment are conditioned by other factors, such as the structure of the economy, labor market flexibility, and previous unemployment levels.

4. How the financial situation of people aged 50+ differed between countries during the pandemic

4.1. Data

The empirical analyses are based on data from the SHARE Corona 1 and Corona 2 Telephone Surveys (Scherpenzeel et al., 2020) and SHARE Wave 8 conducted before the COVID-19 outbreak (Bethmann et al., 2019). To ensure comparability and follow changes in the situation of the same individuals during the pandemic, only those who participated in both telephone surveys were included in the analysis.

Our sample comprises 17,798 individuals from 7 countries, ranging from 1,424 individuals in Sweden to 3,888 individuals in Greece. The results have been generalized using appropriate national-level weighting (Bethmann et al., 2019).

Table 2 presents the sample used in the analysis. The vast majority of respondents were aged 65 and older, with a higher proportion of women. More than half of the respondents lived in two-person households, while less than a third lived alone. However, there were differences in the distributions of the size of older people's households across the studied countries. For example, in Poland, 34.3% of such households consisted of three or more people as compared to only 4.7% in Sweden. Table 2 also highlights the differences in educational backgrounds across countries, with the largest share of people having secondary education in most countries, except for Spain and Greece, where the highest share had education below the secondary level. In Sweden, 37% of respondents had higher education, while only 10.4% of the Spanish respondents attained this level of education. The majority of respondents were retired, ranging from 58.6% in Greece to 90.6% in Czechia. The highest percentages of employed respondents were recorded in Germany (22.4%) and Poland (19.9%).

During the second phase of the pandemic, a significant percentage of households received financial help only in Germany, Czechia, and France, at 55.0%, 35.0%,

and 10.4%, respectively. These data do not correspond with the information presented in Table 1 regarding the Economic Support Index, suggesting that a significant portion of surveyed households may not have qualified for this type of assistance or that the assistance did not reach them.

Table 2: Sample description

Sample characteristics	Number of people							Structure (% of respective total)						
	PL	CZ	SE	DE	FR	ES	EL	PL	CZ	SE	DE	FR	ES	EL
TOTAL*	3184	2759	1424	2135	2178	2230	3888	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Sex														
Men	1406	1063	674	961	884	958	1651	44.2	38.5	45.4	45.0	40.6	43.0	42.5
Women	1778	1696	774	1174	1294	1272	2237	55.8	61.5	54.6	55.0	59.4	57.0	57.5
Household size														
Single	389	713	368	634	635	344	722	12.7	26.9	27.6	23.0	31.2	18.3	21.3
2 people	1526	1551	903	1803	1175	1002	1829	53.0	58.5	67.7	65.5	57.7	53.2	53.9
3 or more people	1055	388	62	316	228	539	843	34.3	14.6	4.7	11.5	11.1	28.5	24.8
Education														
Below secondary**	350	232	220	17	497	1099	1368	11.4	8.7	16.5	0.6	24.4	58.3	40.3
Secondary***	2317	1994	619	1797	959	590	1352	75.5	75.2	46.5	65.3	47.1	31.3	39.8
Higher****	402	426	922	939	582	195	674	13.1	16.1	37.0	34.1	28.5	10.4	19.9
Economic status														
Retired	1977	1890	808	1445	1561	1204	1992	70.8	90.6	83.5	70.9	84.2	67.0	58.6
Employed	555	170	151	457	225	128	520	19.9	8.1	15.6	22.4	12.1	7.1	15.3
Inactive	261	27	9	135	67	465	885	9.3	1.3	0.9	6.7	3.7	25.9	26.1
Age group														
50-59	541	110	54	286	154	57	459	17.6	4.2	4.0	10.4	7.4	2.7	12.2
60-64	597	318	119	506	324	279	555	19.5	12.1	8.8	18.4	15.4	13.1	14.8
65-74	1184	1110	520	1012	807	730	1334	35.6	42.3	38.2	36.8	38.6	34.2	35.6
75 and over	746	1084	667	947	808	1071	1401	27.3	41.4	49	34.4	38.6	50.0	37.4
Receiving financial help in the second wave of the pandemic														
Yes	103	723	38	95	191	68	290	3.7	35.0	3.9	55.0	10.4	3.8	8.8
No	2680	1342	928	78	1646	1723	3081	96.3	65.0	96.1	45.0	89.6	96.2	91.4

Notes: *Differences between TOTAL counts and sums of count of variants of various characteristics are the results of missing answers in survey to certain questions. **ISCED02 ***ISCED 3-4 ****ISCED 5-6.

4.2. Study methodology

In the empirical analysis the following aspects of material situation were accounted for: having savings, subjective feelings regarding the difficulty of making ends meet, the need to use savings to cover daily expenses, and the need to postpone regular payments. We assessed individual aspects of material situation as well as a combined indicator. The synthetic indicator was created using the following formula:

$$I^s = \frac{i_{income} + i_{saving} + i_{use\_savings} + i_{payments}}{4}, \quad (1)$$

where:

i_{income} = 0 if it is difficult for the household to make ends meet, 1 if it is easy,

i_{saving} = 0 if there are no savings, 1 if there are savings,

$i_{use\_savings}$ = 0 if there is a need to use savings for daily expenses, 1 if there is no need,

$i_{payments}$ = 0 if there is a need to postpone regular payments, 1 if there is no need.

The synthetic indicators' values range between zero and one, where one indicates that the financial situation of the household is potentially the most positive achievable, and zero indicates that it is potentially the most negative achievable. A synthetic indicator value below 0.5 can be interpreted as a negative assessment of the overall financial situation. Similarly, a value above 0.5 indicates a positive assessment. To compare the financial situations across countries, synthetic indicators were calculated by aggregating the data for people aged 50+ in each country**.**

The assessment of changes in income between pandemic phases was based on comparisons of the minimum monthly net income during the pandemic waves (see Table A.2 in the Appendix). For the other aspects, a comparative analysis was based on the responses to relevant questions in the SHARE COVID questionnaire. Next, a synthetic indicator of the degree of change was constructed:

$$I^s = \frac{i_{income\_ch} + i_{saving\_ch} + i_{use\_savings\_ch} + i_{payments\_ch}}{4}, \quad (2)$$

where:

i_income_ch = 0 if deteriorated, 0.5 if unchanged, 1 if improved,

i_ savings_ch = 0 if deteriorated, 0.5 if unchanged, 1 if improved,

i_ use_savings_ch = 0 if deteriorated, 0.5 if unchanged, 1 if improved,

i_ payments_ch = 0 if deteriorated, 0.5 if unchanged, 1 if improved.

These synthetic indicators range from zero to one. A value of one indicates the most positive change achievable, while a value of zero indicates the most negative.

Values below 0.5 indicate a negative assessment of changes, whereas values above 0.5 indicate a positive assessment.

To examine the impact of age on the financial situation of households, the distributions of synthetic indicator values were separately analyzed by age for each country using kernel regression (Blundell and Duncan, 1998). This analysis aims to identify a relationship between two variables: the synthetic indicator of overall financial situation/degree of change in financial situation and the age of the individuals surveyed.

4.3. Financial situation of people aged 50+ in the second phase of the pandemic in selected European countries.

Table 3 presents the average values of the four indicators analyzed. Greece stands out with the highest percentage (88.8%) of people aged 50+ finding it difficult to make ends meet, followed by Poland (45.4%). Conversely, Czechia has the lowest percentage (8.6%) of people aged 50+ finding it difficult to make ends meet, with more than 90% indicating that they can manage their finances easily. France has the worst situation among the other countries analyzed, where 22.8% of people aged 50+ find it difficult to make ends meet, which is higher than in other countries, except for Greece and Poland.

In all countries analyzed, a majority of people aged 50+ reported having savings. The percentage is highest in Czechia (93.3%), Greece (91.8%), and France (90.1%). The lowest percentage of households with savings is found in Germany (72.6%).

Among those who face financial difficulties, arrears in the payment of bills and usage of savings to finance current consumption are not only indicators of financial hardship but also strategies to manage income loss during the crisis. The obtained results reveal significant differences between countries in the need to use savings to cover necessary daily expenses. France has the highest percentage (65.8%) of households aged 50+ using savings to meet current needs, followed by Czechia (59.6%), Sweden (54.8%), Germany (53.0%), and Spain (51.3%). Greece, despite having a high percentage of people aged 50+ considering their material situation difficult, has the lowest percentage (19.4%) of people using savings. In Poland, 28% of people aged 50+ declared using savings.

There are considerable differences between countries when it comes to postponing bill payments. Greece has the highest percentage of people aged 50+ postponing bill payments (22.7%), followed by Spain (16.4%). Sweden (3.5%) and Czechia (4.2%) have the lowest percentages.

In summary, during the second phase of the pandemic, the financial situation of people aged 50+ varies considerably across the analyzed countries. While Greece and Poland face greater financial difficulties based on subjective income assessment, countries like Czechia and Sweden have lower percentages of households finding

it difficult to make ends meet. The need to utilize savings and postpone bill payments also differs between countries, revealing distinct financial challenges faced by people aged 50+ during this period.

Table 3: Financial situation of households of people 50+, by its aspects, in the 1st and 2nd waves of the pandemic

Aspects of financial situation	Percentage of households													
	Poland		Czechia		Sweden		Germany		Frane		Spain		Greece	
	1st wave	2nd wave	1st wave	2nd wave	1st wave	2nd wave	1st wave	2nd wave	1st wave	2nd wave	1st wave	2nd wave	1st wave	2nd wave
Subjective income assessment														
difficult	49.6%	45.4%	9.9%	8.6%	8.9%	9.2%	15.5%	14.1%	21.1%	22.8%	27%	14.9%	90.6%	88.8%
not difficult	50.4%	54.6%	90.1%	91.4%	89.1%	90.8%	84.5%	85.9%	78.9%	77.2%	73%	85.1%	9.4%	11.2%
Having savings														
yes		81.7%		93.3%		83.4%		72.6%		90.1%		88.7%		91.8%
no		18.3%		16.7%		16.6%		27.4%		9.9%		11.3%		18.2%
Dipping into saving														
yes	18.2%	28.1%	42%	59.6%	40.9%	54.8%	28.8%	53.0%	34.6%	65.8%	32.9%	51.3%	19.2%	19.4%
no	81.8%	71.9%	58%	40.4%	59.1%	45.2%	71.2%	47.0%	65.4%	34.2%	67.1%	48.7%	80.8%	80.6%
Postponing of the bills payment														
yes	3.1%	10.4%	3.2%	4.2%	6.4%	3.5%	6.4%	11.1%	12.3%	8.7%	9.4%	16.4%	24.2%	22.7%
no	96.9%	89.6%	96.8%	95.8%	93.6%	96.5%	93.6%	88.9%	87.7%	91.3%	90.6%	83.6%	75.8%	77.3%

Source: own analysis based on SHARE.

Table 4 shows the overall average financial situation of households during the second phase of the pandemic. Sweden and Czechia have the highest synthetic indicator values of 0.758. This suggests that older individuals experienced a better overall financial situation in these two countries as compared to other analyzed populations. The lowest value of 0.503 was observed in Greece, indicating the worst households' financial situation.

Table 4: Overall financial situation of households of people aged 50+ in the second wave of the pandemic and its changes during the pandemic

Countries	Synthetic indicator	
	second wave of the pandemic	changes during the pandemic
Poland	0.732	0.493
Czechia	0.758	0.459
Sweden	0.758	0.416
Germany	0.714	0.449
France	0.683	0.451
Spain	0.745	0.584
Greece	0.503	0.466

Source: own analysis based on SHARE.

4.4. Changes in the financial situation of people aged 50+ during the pandemic in selected European countries

In this chapter, we present the changes in the financial situation of individuals aged 50 and over during the pandemic across selected countries. Table 5 provides information on changes of various aspects of the financial situation – both objective (Income, Dipping into Savings, Postponing of Bill Payments) and subjective (Subjective Income Assessment). In all analyzed countries except France and Spain, the majority of households reported a deterioration in equivalent incomes during the second wave of the pandemic. In Sweden, almost two-thirds of those aged 50+ indicated a decrease in income. In Czechia, Greece, Germany, and Poland, the percentage exceeds 50%. In France, more than 40% of people indicated a decline in their household income. Spain stands out, with only 10% of people indicating a decline in their household income during the pandemic.

At the same time, it should be noted that in Spain, more than 80% of households had higher equivalent incomes during the second wave of the pandemic than in its initial wave.

In terms of the subjective assessments, the highest percentage of people who reported an improvement in their households' income situation was observed in Spain (over 30%), followed by Poland (24.7%). The lowest value was observed in Sweden, with only 14.3% of people aged 50+ reporting an improvement in their households' income situation. The greatest deterioration in the subjective assessment of the income situation occurred in the Czech Republic, where 27% of respondents reported a significant deterioration in the ease of making ends meet during the pandemic. The lowest percentage of people who found it more difficult to make ends meet in the second wave of the pandemic than in its first wave was observed in Germany and Sweden, both about 15%.

Table 5: Changes in the financial situation of households of people 50+, by its aspects, during the pandemic

Countries and components of the financial situation	Percentage of households		
	change in situation between pandemic waves		
	deterioration	unchanged	improving
Poland			
Income	55.8%	0%	44.2%
Subjective income assessment	19.8%	55.6%	24.7%
Dipping into saving	18.1%	55.6%	24.7%
Postponing of the bills payment	7.2%	88.3%	4.5%
Czechia			
Income	58.2%	0%	41.8%
Subjective income assessment	27.3%	51.8%	20.9%
Dipping into saving	17.9%	66.9%	15.2%
Postponing of the bills payment	7.1%	92.9%	0%
Sweden			
Income	64.6%	0%	35.4%
Subjective income assessment	15.3%	70.4%	14.3%
Dipping into saving	35.6%	60.9%	3.5%
Postponing of the bills payment	6.4%	86.7%	6.9%
Germany			
Income	50.7%	12.6%	36.7%
Subjective income assessment	15.2%	68.9%	15.9%
Dipping into saving	16.6%	81.3%	2.2%
Postponing of the bills payment	16.5%	79%	4.5%
France			
Income	40.7%	33%	26.4%
Subjective income assessment	18.8%	64.1%	17.1%
Dipping into saving	23.6%	71.8%	4.6%
Postponing of the bills payment	6%	91.6%	2.5%
Spain			
Income	11.3%	7.2%	81.5%
Subjective income assessment	16.9%	51.7%	31.4%
Dipping into saving	18.5%	66.5%	15%
Postponing of the bills payment	11.8%	77.6%	10.6%
Greece			
Income	56.7%	10.9%	32.5%
Subjective income assessment	16.7%	63.6%	19.7%
Dipping into saving	14.2%	79.7%	6%
Postponing of the bills payment	7.3%	80.6%	12.1%

Source: own analysis based on SHARE.

Our analysis shows large differences between countries in the dimension of the need to use savings. The greatest improvement was observed in Czechia, Spain, and Poland, where about 18% of households used savings in the first pandemic phase but not in the second. The greatest deterioration occurred in Sweden, where 35.6% of households began using savings in the second pandemic phase. France also saw a significant 23.6% of older households started using savings. In other countries, it was under 20%, with Greece having the lowest at 14.2%.

The last dimension analyzed was the need to postpone regular payments due to financial distress. The highest deterioration was observed in Germany and Spain, with over 15% and 10% of households having to postpone payments. In other countries, the percentage ranged between 5 and 10%. Greece and Spain had the highest improvement in this dimension, with more than 10% of households no longer needing to postpone payments.

As to the overall financial situation, Spain saw the most improvement (indicator 0.584), while Sweden, Czechia, Germany, France, and Greece experienced declines (indicators 0.416 to 0.466), indicating increased financial challenges for households aged 50+ during the pandemic. In Poland (average indicator value of 0.493) older individuals experienced on average a stable financial situation during the pandemic.

4.5. Assessment of the relationship between age and the financial situation of households of people aged 50+

This section of our study focuses on assessing two aspects of financial conditions among individuals aged 50 and over during the second phase of the pandemic. We first explore the relationship between age and the overall financial situation, identifying how different age groups have been affected across various European countries. We then analyze the changes in these financial situations as the pandemic progressed. Figure 1 shows the relationship of the overall financial situation and age. Interestingly, the course of the regression lines differs considerably between the two countries with the highest values of the synthetic indicator - Spain and Sweden. In Sweden, the overall financial situation tends do decrease with age. In Spain, on the other hand, the worst overall financial situation was experienced by households of the youngest and oldest individuals , i.e. under 65 and over 85.

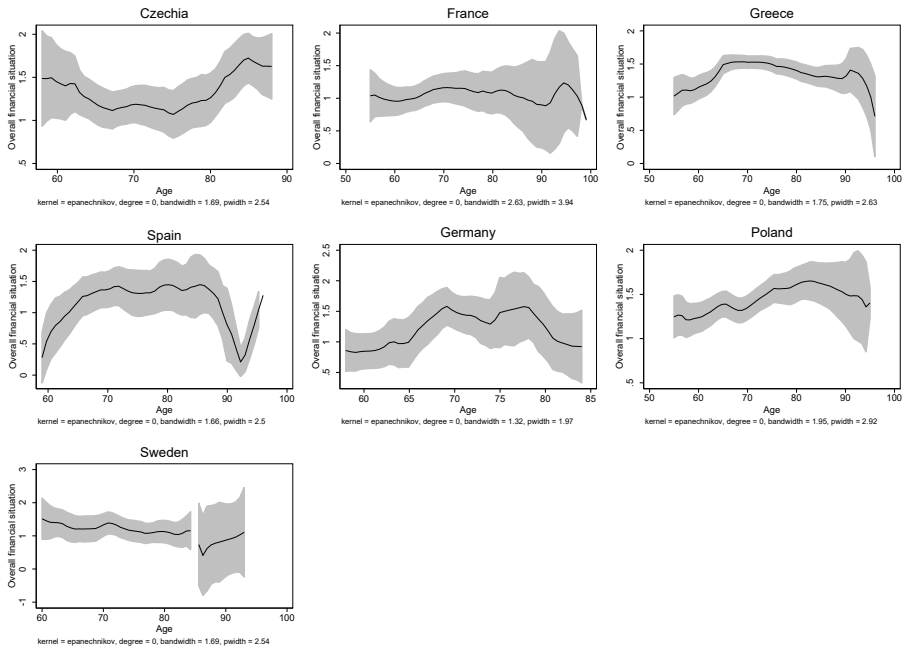


Figure 1: Relationship between the overall financial situation of people aged 50+ in the second phase of the pandemic and age

Source: own analysis based on SHARE.

A similar inversed U-shaped relationship was observed for Germany, Greece, Poland and, to a lesser extent, for France. By contrast, we can observe an U-shaped relationship for Czechia, where households of relatively younger (under 65) and older (over 85) people enjoy the best financial situation.

Figure 2 shows similar kernel regressions illustrating the relationship between the change in the overall financial situation and age. Compared to Figure 1, the diversity of the shape of regression lines is much smaller – in almost all countries, a similar horizontal course of the regression line can be observed, indicating a weak relationship between the change in the overall financial situation and age. In Spain, the average change in the synthetic indicator decreases with age, which indicates a relatively stronger positive change in the overall financial situation of younger households compared to older households. The regression line remains above the line at the level of 0.5 throughout its course, which means that, on average, the financial situation of all households improved in all age groups.

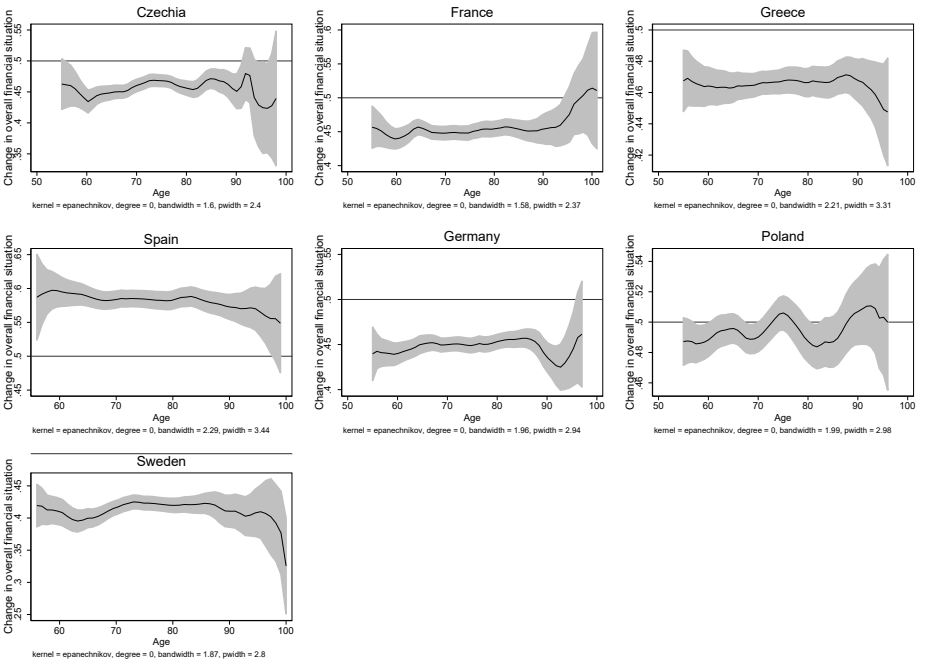


Figure 2: Relationship between the change in the overall financial situation of people aged 50+ during the pandemic and age

Source: own analysis based on SHARE.

In the remaining countries, on average, the change in the overall financial situation was negative for households of people in all age groups, with the exception of Poland, where the 95% confidence bands cover the 0.5 horizontal line, indicating no significant change in financial situation across all age groups.

5. Discussion and conclusion

The results of the analysis provide insight into the changes in the overall financial situation of households with people aged 50+ during the pandemic across several European countries. The analyzed countries experienced different degrees of decline in unemployment rates and GDP during the second phase of the pandemic. However, differences in changes in the Stringency Index and Economic Support Index indicate complex relationships between government policies, support measures, and the overall financial situation of households. In Germany and Greece, the Stringency Index increased while support measures remained unchanged. In Sweden, the Stringency Index decreased while support measures stayed at the same level. In France, both the

Stringency Index and support measures decreased. In Poland, the Stringency Index decreased while support measures increased. In Czechia, the Stringency Index increased while support measures also increased. In each of these cases, the index and support measures differently impacted the changes in the overall financial situation of households with people aged 50+ during the pandemic. Spain, as previously mentioned, was the only country that experienced an improvement in the overall financial situation of households with people aged 50+ during the second phase of the pandemic. The Stringency Index decreased while the Economic Support Index remained unchanged.

The overall financial situation in the surveyed countries during this period was influenced not only by the above-mentioned macroeconomic factors but also by microeconomic factors with varying degrees of impact. For example, Spain had a relatively higher proportion of households consisting of three or more people (28.5%), which may have provided additional support within the household, contributing to financial stability during the pandemic. The relatively lower percentage of single-person households in Spain may also have contributed to this positive outcome. Despite the fact that a significant portion of the population in Spain had education below the secondary level (58.3%), which might typically limit employment opportunities and financial resilience, other factors such as effective government support and intra-household support may have played a more decisive role in mitigating financial difficulties during the pandemic. On the other hand, countries such as Sweden, Czechia, Germany, and France experienced the steepest decline in the overall financial situation of households with people aged 50+ among the surveyed countries during the second phase of the pandemic. These data indicate that households in these countries faced greater financial challenges during the pandemic, possibly due to insufficient economic recovery after the decline in GDP and the rise in unemployment rates during the first phase of the pandemic or inadequate social support measures. In addition to macroeconomic factors, microeconomic factors also contributed to the decline in the overall financial situation of households with people aged 50+ in these countries during the second phase of the pandemic. In Sweden, for example, only 4.7% of households consisted of three or more people, which might suggest a lack of intra-household support during financial hardships. Meanwhile, in France, the high percentage of single-person households (31.2%) could have contributed to the observed financial vulnerability during the pandemic. In Sweden, a higher percentage of respondents had higher education (37.0%), which typically correlates with better employment opportunities and income levels. However, despite this, Sweden experienced a significant decline in overall financial situation during the second phase of the pandemic, indicating other prevailing factors, such as household size and social support structure. Greece and Poland, which had a relatively smaller decline in the financial

situation of older households compared to the aforementioned countries, present two different scenarios. Greece experienced a slight decline in the financial situation of households aged 50+ during the pandemic, while Poland had a more stable financial situation among older households, as the changes during the pandemic were relatively minor.

The analysis of the changes in the financial situation of individuals over 50 years old during the second phase of the pandemic must also consider a range of other factors, not analyzed in the study, that can affect these changes. For example, pension system reforms may have impacted the retirement income of individuals over 50, thereby affecting their financial situation. During the COVID-19 pandemic, some countries postponed their pension reforms, while others decided to introduce policy changes, e.g. to increase pension benefits (Natali, 2020). This underscores the importance of monitoring pension systems and ensuring that reforms do not have negative impacts on the changes in the overall financial situation of older individuals during the second phase of the pandemic. Inflation is another factor that may negatively impact the financial situation of individuals over 50. During the pandemic, inflation rates exceeded 10 percent in most European countries. As the prices of goods and services rose, the real value of household incomes and savings declined, leading to reduced purchasing power. Moreover, Jaravel and O'Connell (2020) found that in 2020 inflation was the highest among older households in comparison to other socio-demographic groups. Age-based discrimination is also a concern for individuals over 50, particularly when it comes to employment opportunities. For example, Lössbroek et al. (2021) conducted a survey among managers, which showed that older workers have smaller chances of being hired, regardless of their skills. Other factors that may contribute to job loss, particularly in sectors dominated by older workers, are automation and digitization of jobs (Georgieff, Milanez, 2021). Finally, the availability of social services and healthcare is another factor that may impact the changes in the overall financial situation of individuals over 50 during the second phase of the pandemic. Budget cuts may limit access to social services and healthcare, which can generate additional costs associated with healthcare or caring for other family members. Several studies have shown that during the first wave of the COVID-19 pandemic, there was a substantial increase in unmet healthcare needs among people aged over 50 (Lourenço et al., 2022; Arnault et al., 2021).

In summary, although the unemployment rate decreased and GDP increased in all the analyzed countries, differences in government policies, economic support, and other factors such as inflation or availability of healthcare affected the changes in the overall financial situation of individuals over 50 during the second phase of the pandemic in different ways. Alongside these macroeconomic factors, microeconomic factors (characteristics of the surveyed individuals and their households) also had

a significant impact on the changes. The complexity of the impact of various factors on the financial situation of the surveyed population underscores the need for further comprehensive analyses and the development of a multi-dimensional social policy strategy aimed at mitigating economic shocks in the event of future economic crises.

References

Abramowska-Kmon, A., Antczak, R., Chełchowska, M., Chłoń-Domińczak, A., Holzer-Żelażewska, D., Łątkowski, W., Magda, I., Oczkowska, M., Panek, T., Perek-Białas, J., Ruzik-Sierdzińska, A., Saczuk, K., Styrc, M., Strzelecki, P., Wróblewska, W. and Zwierzchowski, J., (2023). *The situation of people aged 50+ in Poland and Europe during the second wave of the COVID-19 pandemic: Report from the SHARE Corona 2 telephone survey conducted as part of the 9th wave of the "SHARE: 50+ in Europe" study* (in Polish). SGH Publishing House – Warsaw School of Economics.

Altiparmakis, A., Bojar, A., Brouard, S., Foucault, M., Kriesi, H. and Nadeau, R., (2021). Pandemic politics: Policy evaluations of government responses to COVID-19. *West European Politics, 44*(5–6), pp. 1159–1179

Almeida, V., Barrios, S., Christl, M., De Poli, S., Tumino, A. and van der Wielen, W., (2021). The impact of COVID-19 on households income in the EU. *The Journal of Economic Inequality*, 19(3), pp. 413–431.

Arnault, L., Jusot, F. and Renaud, T., (2021). Economic vulnerability and unmet healthcare needs among the population aged 50+ years during the COVID-19 pandemic in Europe. *European Journal of Ageing*, 19(4), pp. 811–825.

Baptista, I., Marlier, E., Spasova, S., Peña-Casas, R., Fronteddu, B., Ghailani, D., Sabato, S., and Regazzoni, P., (2021). Social protection and inclusion policy responses to the COVID-19 crisis – An analysis of policies in 35 countries. European Commission, Directorate-General for Employment, Social Affairs and Inclusion, Publications Office of the European Union, Luxembourg.

Battistini, N., Stoevsky, G., (2021). The impact of containment measures across sectors and countries during the COVID-19 pandemic. *Economic Bulletin Boxes*, European Central Bank. Bethmann A., Bergmann, M. and Scherpenzeel A., 2019, Sampling guide-wave 8, 9.01, *Working Paper Series*, 33.

Blundell, R., Duncan, A., (1998). Kernel regression in empirical microeconomics. *Journal of Human Resources*, 33(1), pp. 62–87.

Bonfatti, A., Pesaresi, G., Weber, G. and Zambon, N., (2023). The economic impact of the first wave of the pandemic on 50+ Europeans. *Empirical Economics*, 18, pp. 1–53.

Botelho, V., Weißler, M., (2022). COVID-19 and retirement decisions of older workers in the euro area. *Economic Bulletin Boxes*, 6.

Cantillon, B., Seeleib-Kaiser, M. and Van der Veen, R., (2021). The COVID-19 crisis and policy responses by continental European welfare states. *Social Policy & Administration, 55*(2), pp. 326–338.

Castaneda Aguilar, R. A., Dewina, R., Diaz-Bonilla, C., Edochie, I. N., Fujs, T. H. M. J., Jolliffe, D. M., Lain, J., Lakner, C., Lara Ibarra, G. and Gerszon, D., (2022). April 2022 update to the Poverty and Inequality Platform (PIP): What's New. *Global Poverty Monitoring Technical Note Series*, 20*,* The World Bank.

Celbiş, M. G., Wong, P. H., Kourtit, K. and Nijkamp, P., (2023). Impacts of the COVID-19 outbreak on older-age cohorts in European Labor Markets: A machine learning exploration of vulnerable groups. *Regional Science Policy & Practice*, 15(3), pp. 559–584.

Chłoń-Domińczak, A., Holzer-Żelażewska, D., (2022). Economic stress of people 50+ in European countries in the COVID-19 pandemic–do country policies matter?. *European journal of ageing*, 19(4), pp. 883–902.

Clark, A. E., D'Ambrosio, C., Lepinteur, A. and Menta, G., (2022). Pandemic Policy and Individual Income Changes across Europe, 600. ECINEQ, Society for the Study of Economic Inequality.

Di Pietro, G., (2022). Changes in household income during COVID-19: a longitudinal analysis. *SN Business & Economics*, 2(10).

Echebarria Fernández, J., (2021). A critical analysis on the European Union's measures to overcome the economic impact of the COVID-19 pandemic. *European Papers - A Journal on Law and Integration*, 5(3), pp. 1–25.

European Social Policy Network (ESPN), (2021). The impact of the COVID-19 pandemic on social inequalities in the European Union. Publications Office of the European Union, Luxembourg.

Fana, M., Tolan, S., Torrejón, S., Urzi Brancati, C. and Fernández-Macías, E., (2020). The COVID confinement measures and EU labour markets. Publications Office of the European Union, Luxembourg.

Gentilini, U., Almenfi, M., Orton, I. and Dale, P., (2020). Social protection and jobs responses to COVID-19. World Bank Publications – Reports, 33635. The World Bank Group.

Georgieff, A., Milanez, A., (2021). What happened to jobs at high risk of automation?, OECD Social, *Employment and Migration Working Papers*, 255, OECD Publishing, Paris,

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S. and Tatlow, H., (2021). A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nature Human Behaviour*, 5(4), pp. 529–538.

Hale, T., Angrist, N., Kira, B., Petherick, A., Phillips, T. and Webster, S., (2020). *Variation in government responses to COVID-19*. Blavatnik School of Government Working Paper.

Jaravel, X., O'Connell, M., (2020). Real-time price indices: Inflation spike and falling product variety during the Great Lockdown. *Journal of Public Economics*, 191(C).

Lourenco, O., Quintal, C., Moura-Ramos, L. and Antunes, M., (2022). The Impact of the COVID-19 Pandemic on the Unmet Healthcare Needs in People Aged Over 50 in Portugal. *Acta Médica Portuguesa*, 35(6), pp. 416–424.

Lössbroek, J., Lancee, B., van der Lippe, T. and Schippers, J., (2021). Age discrimination in hiring decisions: A factorial survey among managers in nine European countries. *European Sociological Review*, 37(1), pp. 49–66.

Mahler, D. G., Yonzan, N. and Lakner, C., (2022). The impact of COVID-19 on global inequality and poverty. *Policy Research Working Paper Series, 10198*, The World Bank.

Natali, D., (2020). Pensions in the age of COVID-19: recent changes and future challenges. ETUI Research Paper-Policy Brief, 13.

OECD, (2020). Employment outlook 2020: worker security and the Covid-19 crisis. OECD Publishing.

Rajevska, O., Reine, A., Baltmane, D. and Gehtmane-Hofmane, I., (2022). Older Workers in the Baltic States over the Pandemic Year: SHARE Corona Survey Results. In Economic Science for Rural Development Conference Proceedings, 56.

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D. and Roser, M., (2020). Coronavirus pandemic (COVID-19). Published online at OurWorldInData.org.

Scherpenzeel, A., Axt, K., Bergmann, M., Douhou, S., Oepen, A., Sand, G., Schuller, K., Stuck, S., Wagner, M. and Börsch-Supan, A., (2020). Collecting survey data among the 50+ population during the COVID-19 outbreak: The Survey of Health, Ageing and Retirement in Europe (SHARE). In *Survey Research Methods*, 14(2), pp. 217–221).

Theodoropoulos, N., Voucharas, G., (2022). Containment Measures and Job Loss: Evidence from SHARE Corona Surveys. *Economic Policy Papers*. Economic Research Centre, University of Cyprus.

Weber, T., Hurley, J. and Adăscăliței, D., (2021). COVID-19 – Implications for employment and working life. Eurofound, https://data.europa.eu/doi/10.2806/160624.

Appendix

Table A.1. Partial indicators of the financial situation of households in the 2nd wave of the pandemic

Indicators	Definition of indicators
Subjective income assessment	Assessments of how difficult it is for a household to make ends meet with its monthly net income (CCACO107): 1. With difficulty – 1 or 2 2. Easily – 3 or 4
Having savings	Having savings (CAE120): 1. yes – 1 and more 2. no – 0
Dipping into saving	Dipping into saving (CAE112) 1. yes – 1 2. no – 5
Postponing of the bills payment	Postponing of the bills payment: 1. yes – 1 2. no – 5

Source: own work.

Table A.2. Partial indicators of changes in the financial situation of households during the pandemic

Partial indicators	The method of indicators' construction
Income	Difference between lowest net monthly income in the 2nd and the 1st wave of the pandemic: 1. Positive difference 2. Negative difference
Subjective income assessment	Difference between assessments of how difficult it is for a household to make ends meet with its monthly net income in the 2nd and 1st wave of the pandemic: 1. Making ends meet more difficult – deterioration in the situation 2. Unchanged 3. Making ends meet easier – improving the situation

Table A.2. Partial indicators of changes in the financial situation of households during the pandemic (cont.)

Partial indicators	The method of indicators' construction
Dipping into saving	Difference between states of dipping into saving: 1. No dipping into saving in the 1st wave and dipping into saving in the 2nd wave – deterioration in the situation 2. Dipping into saving in the 1st and in the 2nd wave or no dipping into saving in the 1st wave and dipping into savings in the 2nd wave – unchanged 3. No dipping into saving in the 1st wave and in the 2nd wave – improving the situation
Postponing of the bills payment	Difference between states of postponing of the bills payment: 1. No postponing of the bills payment in the 1st wave and postponing of the bills payment in the 2nd wave – deterioration in the situation 2. Postponing of the bills payment in the 1st and in the 2nd wave or no postponing of the bills payment in the 1st and in the 2nd wave – unchanged 3. Postponing of the bills payment in the 1st wave and no postponing of the bills payment in the 2nd wave – improving the situation

Source: own work.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4 pp. 51–77, https://doi.org/10.59139/stattrans-2024-003
Received – 26.03.2024; accepted – 17.07.2024

Modelling Tinnitus Functional Index reduction using supervised machine learning algorithms

Edmund Fosu Agyemang[1]

Abstract

This study aims to model the reduction in the Tinnitus Functional Index (TFI) utilizing supervised machine learning algorithms, focusing primarily on Ordinary Least Squares (OLS), K-Nearest Neighbor (KNN), Ridge, and Lasso regressions. Our analysis highlighted Group, ISI, and SWLS as significant predictors of TFI reduction, identified through the best subset selection and confirmed by both forward and backward selection criteria in the OLS regression. Notably, the shrinkage methods, Ridge and Lasso regressions, demonstrated superior performance compared to OLS and KNN, with the Ridge regression presenting the smallest test mean square error (MSE) of 318.30. This finding establishes the Ridge regression as the best model for analyzing our Tinnitus dataset relative to the other methods, which exhibited test MSEs of 319.28 (Lasso), 330.76 (OLS), and 584.92 (KNN), respectively. This research highlights the potential of supervised machine learning algorithms in advancing personalized Tinnitus treatment, reflecting broader trends in the field as evidenced by studies in the literature. By leveraging these algorithms, we can enhance treatment precision and outcomes, contributing significantly to improved quality of life for individuals with Tinnitus. Future research should explore the integration of multimodal data and longitudinal applications of these algorithms to further refine predictive capabilities and treatment effectiveness.

Key words: Tinnitus, K-Nearest Neighbor regression, Ridge regression, Lasso regression, multiple linear regression.

1. Introduction

Tinnitus, characterized by the perception of noise or ringing in the ears in the absence of any external sound, is a prevalent and often distressing auditory phenomenon (De Ridder et al., 2021). Affecting a significant portion of the population,

[1] Corresponding author. School of Mathematical and Statistical Science, College of Sciences, University of Texas Rio Grande Valley, USA &Department of Statistics and Actuarial Science, College of Basic and Applied Sciences, University of Ghana, Ghana &Department of Computer Science, Ashesi University, No. 1 University Avenue, Berekuso-Ghana. E-mail: edmundfosu6@gmail.com, ORCID: https://orcid.org/0000-0001-8124-4493.

Tinnitus can vary in severity from a mild annoyance to a debilitating condition, impacting quality of life, emotional well-being, and cognitive function (Gasparre et al., 2023). Despite its common occurrence and prevalence, the etiology of Tinnitus remains complex and the precise mechanisms underlying Tinnitus remain poorly understood, and as such, treatment can be challenging, with potential contributions from auditory and non-auditory structures, suggesting a multidimensional pathology (Mohan et al., 2022). The Tinnitus Functional Index (TFI) has emerged as a crucial measure for evaluating the impact of Tinnitus on daily life, encompassing emotional distress, auditory difficulties, and interference with mental concentration among other factors. However, given the multifarious nature of Tinnitus, individual responses to treatment vary significantly, highlighting the need for personalized intervention strategies. Recent advances in supervised machine learning algorithms offer promising new avenues for understanding and managing complex health conditions like Tinnitus (Manta et al., 2023). These algorithms, capable of discerning patterns and associations in large datasets, present an innovative approach to predicting patient outcomes and optimizing treatment protocols. By integrating patient data, including demographic, audiometric, and psychological factors, supervised machine learning models can potentially identify the most influential predictors of TFI reduction, thereby aiding clinicians in tailoring interventions more effectively to individual needs. The integration of machine learning in the medical field is not without challenges. The accuracy of predictive models depends on the quality and quantity of the available data, and in the context of Tinnitus, this includes accurately captured TFI scores, detailed patient histories, and comprehensive treatment records.

Despite these challenges, the potential benefits of applying supervised machine learning to Tinnitus management are considerable. By enabling more precise and personalized treatment approaches, these technologies have the potential to significantly improve quality of life for Tinnitus sufferers. Furthermore, the insights gleaned from machine learning models could contribute to a deeper understanding of Tinnitus pathology and the development of more effective therapeutic interventions. The most common cause of Tinnitus is prolonged exposure to loud noise, which can damage the tiny sensory hair cells in the ear that transmit sound to the brain (Runciman and Johnson, 2023). However, Tinnitus can also result from other factors such as age-related hearing loss, earwax blockage, ear bone changes (otosclerosis), and conditions such as Meniere's disease, which affects the inner ear. Additionally, certain medications, including some antibiotics, cancer medications, and even high doses of aspirin, can induce Tinnitus as a side effect. Other potential causes include head or neck injuries (Biswas et al., 2023), which can affect the auditory nerves or brain function linked to hearing, and various cardiovascular diseases like hypertension, which can interfere with blood flow and cause Tinnitus. Machine learning is rapidly transforming the field of healthcare, offering new ways to enhance diagnostic accuracy and patient

treatment outcomes (Agbota et al., 2024). Understanding the specific cause is crucial for effective treatment and management of the condition.

A systematic review by McCormack et al. (2016) highlighted that Tinnitus prevalence increases with age and exposure to noisy environments. The study also noted that about 10-15% of the adult population experiences Tinnitus, but only a small fraction, around 1-2%, finds it severely debilitating. The exact pathophysiological mechanisms of Tinnitus remain partially understood, but recent research suggests it involves both peripheral and central auditory pathways. Research by Eggermont and Roberts (2004) proposed that Tinnitus results from neural plasticity in response to auditory system damage, leading to altered neural activity perceived as sound. This theory is supported by findings of increased activity in the auditory cortex and changes in the brain's neural network connectivity. Tinnitus is not merely a sensory condition but also has profound psychological impacts. Studies have shown a strong correlation between Tinnitus severity and psychological distress, including anxiety, depression, and reduced quality of life. A study by Langguth et al. (2013) reviewed the impact of Tinnitus on mental health, underscoring the need for holistic approaches in treatment to address both auditory and psychological components. Smith et al. (2021) applied various supervised machine learning techniques to predict the outcomes of cognitive-behavioral therapy in individuals suffering from Tinnitus, as measured by changes in Tinnitus Functional Index (TFI) scores. They found that the random forest algorithm outperformed other models, such as decision trees and support vector machines, in predicting treatment success. The study emphasizes the value of incorporating diverse patient data to enhance the predictive accuracy of treatment outcomes, suggesting a move towards more personalized treatment approaches.

Doborjeh et al. (2023) proposed an Artificial Intelligence algorithm to predict patients' responses to Tinnitus therapies using EEG data. By employing deep learning techniques, the study achieved prediction accuracies ranging from 98% to 100%. The study identified the most informative EEG sensors and demonstrated how EEG frequency and functional connectivity could classify patients into therapy respondents and non-respondent groups, thereby suggesting a potential for real-time monitoring of therapy outcomes. Rodrigo et al. (2021) utilized decision tree models, specifically CART and gradient boosting, to predict the outcomes of Internet-Based Cognitive Behavioral Therapy (ICBT) for Tinnitus. The research highlighted that higher education levels were significantly influential for ICBT outcomes, and the CART decision tree model was able to identify participant groups with an 85% success probability following ICBT. In Cardon et al. (2022), machine learning algorithms, particularly random forest classifiers, were employed to predict outcomes of Tinnitus treatment modalities. The study noted that Tinnitus Functional Index (TFI) scores varied among participants, with some showing no improvement. The study contributes to understanding how machine learning can assist in predicting the response to specific

Tinnitus treatments, indicating a significant advancement toward personalized medicine in audiology. Prominent statistical tools such as cluster analysis (Van den Berge et al., 2017) and factor analysis (Wakabayashi et al., 2020) have been utilized to analyze Tinnitus in the literature.

However, this study aims to explore and identify the significant predictors of Tinnitus reduction, encompassing a range of possible factors including auditory exposure, demographic variables, psychological factors, and underlying health conditions. By employing a comprehensive and integrative approach, this study seeks to unravel the complex web of contributors to TFI reduction. By analyzing data from a diverse cohort of patients, we aim to identify key factors influencing treatment outcomes and to establish a predictive framework that can guide clinical decision-making. This research not only has the potential to enhance individual patient care but also contributes to the broader field of Tinnitus research by incorporating novel data-driven methodologies. The decision to employ four variants of supervised learning algorithms in the analysis of significant predictors of Tinnitus reduction represents a robust approach, leveraging the strengths of both parametric and non-parametric statistical techniques. This diverse approach enhances the ability to accurately model the predictors of Tinnitus reduction, catering for both linear and non-linear dynamics within the data.

2. Data and Methods

The dataset used in this study is of secondary nature and comprises pre and post Tinnitus Functional Index (TFI) scores, along with clinical and demographic information, from a pre-post intervention research involving 142 individuals affected by Tinnitus. The dependent variable for the study is TFI Reduction, which was computed as the difference between the TFI score at the beginning of the study and the TFI score after completion of the study. The predictors for the study include **HHI**: Hearing survey score; **GAD**: Generalized Anxiety Disorder, **PHQ**: Patient Health Questionnaire, **ISI**: Insomnia Severity Index, **SWLS**: Satisfaction with Life Scales, **HYP**: Hyperacusis, **CFQ**: Cognitive Failures, **Duration**: Duration of Tinnitus (in years), **PRE**: TFI score at the beginning of the study, **POST**: TFI score after the completion of the study. The TFI score at the end of the study had 17 missing observations. In this study, the MICE (Multiple Imputation by Chained Equations) algorithm was employed to address missing values in the dataset. Utilizing the Iterative Imputer from Scikit-learn with a Random Forest Regressor, the algorithm iteratively predicted missing numerical data over ten cycles, starting with mean values as place holders. This approach facilitated the estimation of missing data by exploiting inter-variable relationships, ensuring a more accurate and robust imputation compared to simpler methods. 10-fold cross-validation (i.e. setting k = 10) was used in the study. This process is repeated for

10 iterations. In each iteration, a different fold is kept for testing, and the remaining 9 folds are used for training. OLS, KNN, Ridge and Lasso regression were the main statistical tools adopted for the study. The data and python codes used for the study are freely available on Github and can be assessed at the repository at https://github.com/Agyemang1z/Tinnitus-Case-Study-1. The complete description of the Tinnitus dataset is given in Table 1.

Table 1: Description of the Tinnitus Dataset

Attribute	Description	Data Type
Group	Treatment and Control	Binary
Gender	1: Male 2: Female	Binary
HHI _Score	Hearing survey- Overall score- 0-40 (higher score more severe)	Numeric
Generalized Anxiety Disorder (GAD)	Anxiety sum: 0-21 (higher score more severe)	Numeric
Patient Health Questionnaire (PHQ)	Depression sum: 0-28 (higher score more severe)	Numeric
Insomnia Severity Index (ISI)	Insomnia total: 0-28 (higher score more severe)	Numeric
Satisfaction with Life Scales (SWLS)	Overall score, satisfaction with life, like Quality of Life (QOF) Higher scores better QOL (opposite to all other scales)	Numeric
Hyperacusis	0-42 (higher score more severe)	Numeric
Cognitive Failures (CFQ)	0-100 (higher score more severe)	Numeric
Age	In years	Numeric
Duration of tinnitus	In years	Numeric
Pre TFI Score	TFI score at the beginning of the study: Tinnitus score out of 100, higher more severe	Numeric
Post TFI Score	TFI score after the completion of the study: Tinnitus scores out of 100, higher more sever	Numeric

2.1. Ordinary Least Squares and Multiple Regression

Multiple regression allows us to determine the overall fit of the model and the relative contribution of each of the predictors to the total variance. Linearly, we consider the general model given in (1) as:

$$Y_i = \sum_{j=0}^{p} \beta_j X_{i,j} \; ; \quad i = 1,2, \dots, N \tag{1}$$

The least squares estimates, β_j in (1) that minimizes residual sums of squares (RSS) is given in (2) by:

$$RSS(\beta) = \sum_{i=1}^{N}\left[Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right]^2 \quad (2)$$

However, it is possible that not all predictors are significantly associated with the outcome variable. Some predictors might not contribute meaningfully to the model, potentially leading to overfitting. To address this, variable selection techniques such as Best Subset Selection, Forward Selection, and Backward Elimination are employed. These methods aim to find the most relevant subset of predictors that offer the best prediction accuracy, balancing the model's complexity with its predictive power. Best Subset Selection was adopted for the study, where all possible combinations of predictors are evaluated, which can be computationally intensive for a large number of predictors. To evaluate the performance and select the optimal model among different candidate models, various statistical metrics and validation techniques are utilized. Commonly used criteria include the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and adjusted R^2, which help to compare models based on their explanatory power and complexity. Cross-validation techniques, such as k-fold cross-validation, are also widely used to estimate a model's prediction error on new data, ensuring that the selected model generalizes well beyond the observed dataset. Ultimately, the chosen MLR model should strike a balance between complexity (number of predictors) and prediction accuracy, avoiding overfitting while effectively capturing the relationships within the data.

2.2. Ridge and Lasso Regression

The Ordinary Least Squares (OLS) method is commonly used for estimation in regression models, offering adequate predictions under certain conditions. However, its effectiveness diminishes when dealing with nonorthogonal explanatory variables, leading to inaccurate weighting of these predictors (Saleh and Norouzirad, 2018). This issue is particularly pronounced in datasets not derived from controlled experiments, where the assumption of non-orthogonality among variables does not hold, and multicollinearity is present. OLS tends to produce unstable and overfitted models in such scenarios, emphasizing the need for strategies that mitigate model overfitting by reducing variance. To address these limitations, alternative methods like ridge and lasso regression have been developed. These techniques are designed to provide biased estimates of regression coefficients, particularly beneficial in cases of correlated predictors or multicollinearity (Omer, 2022). By introducing a bias into the regression results, they effectively lower the variances of the estimates, which enhances the interpretability of the regression coefficients and the overall model reliability. These strategies aim to utilize all p predictors while reducing their coefficients towards zero,

effectively creating a subset when some coefficients become zero. It is important to note that the intercept is exempt from this reduction process. Using the Lagrangian equation in (3), the estimation of $\hat{\beta}$ is made possible by:

$$\hat{\beta}_{\text{shrink}} = \arg\min_{\beta}\left[\sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{k=1}^{p}\beta_j X_{i,j}\right)^2 + \lambda\sum_{k=1}^{p}\psi(\beta_j)\right] \tag{3}$$

Here, $\lambda\sum_{k=1}^{p}\psi(\beta_j)$ is the shrinkage penalty and $\lambda \geq 0$ is the regularization parameter.

To estimate the ridge (also known as the l_2 regularization), we use the shrinkage penalty $\lambda\sum_{k=1}^{p}\psi(\beta_j)$ so that (3) is modified as in (4) as:

$$\hat{\beta}^{\boldsymbol{Ridge}}_{\text{shrink}} = \arg\min_{\beta}\left[\sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{k=1}^{p}\beta_j X_{i,j}\right)^2 + \lambda\sum_{k=1}^{p}\beta_j^2\right] \tag{4}$$

In similar fashion, to estimate the lasso (also known as the l_1 regularization), we use the shrinkage penalty $\psi(\beta_j) = |\beta_j|$ so that (3) is modified as in (5) as:

$$\hat{\beta}^{\boldsymbol{Lasso}}_{\text{shrink}} = \arg\min_{\beta}\left[\sum_{i=1}^{N}\left(Y_i - \beta_0 - \sum_{k=1}^{p}\beta_j X_{i,j}\right)^2 + \lambda\sum_{k=1}^{p}|\beta_j|\right] \tag{5}$$

Y_i denotes the *i*th observation of the response variable (TFI Reduction), β_0 is a constant term, $X_{i,j}$ represents the *i*th observation of the *j*th explanatory variables (predictors of TFI Reduction), β_j is the associated *j*th coefficient. As λ increases, the absolute value amount of the estimated coefficient shrinks towards zero. Under LASSO, the coefficient of unimportant variables is reduced completely to zero. Ridge does not perform variable selection but minimizes the impact of irrelevant predictors in the model shrinking the estimated coefficient near zero but not completely zero. Whenever $\lambda = 0$ it produces the results of the normal Ordinary Least Squares regression. These techniques effectively address collinearity by design, balancing the coefficients of correlated variables—reducing them to small and negative values when one is significantly large. They also adjust for instances where **X** lacks full rank.

2.3. KNN Regression

Let $X = \{x_1, x_2, \ldots, x_n\}$ be the set of training data and $y = \{y_1, y_2, \ldots, y_n\}$ be the set of corresponding TFI Reduction scores. For a new patient data point x', the KNN algorithm searches the training set to find the k nearest neighbors based on a distance metric (Euclidean distance was adopted in this study) given in (6) by:

$$D(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2} \tag{6}$$

The KNN algorithm was then used to calculate the average severity score of these k neighbors to predict the TFI Reduction score for the new patient:

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

where y_i are the severity scores of the k nearest neighbors to x'.

3. Results and Findings

Table 2: Descriptive statistics of quantitative variables

Specification	HHI	GAD	PHQ	ISI	SWLS	HYP	CFQ	Age	Duration	PRE	POST
count	142.00	142.00	142.00	142.00	142.00	142.00	142.00	142.00	142.00	142.00	125.00
mean	17.79	7.48	8.03	12.96	20.32	19.04	40.59	55.45	11.99	59.37	50.52
std	11.37	5.58	5.67	7.04	7.36	8.50	15.96	12.88	12.50	18.25	21.88
min	0.00	0.00	0.00	0.00	5.00	1.00	7.00	22.00	0.30	24.40	4.00
25%	8.00	3.00	4.00	8.00	14.00	13.00	29.25	46.25	3.00	46.80	32.00
50%	18.00	6.00	7.00	13.00	20.00	18.50	41.00	58.00	10.00	58.60	57.10
75%	26.00	11.00	11.00	18.00	26.00	25.00	50.00	65.00	15.00	73.60	66.00
max	40.00	21.00	27.00	27.00	35.00	42.00	86.00	83.00	55.00	97.20	88.40

Table 2 offers a statistical summary for a selection of variables collected from 142 participants, including TFI score after the completion of the study which has data for 125 participants. The 'HHI' has an average of 17.79, with a standard deviation of 11.37, and ranges from 0 to 40. The 'GAD' and 'PHQ' scores, which assess anxiety and depression, have similar counts and display averages of 7.48 and 8.03, respectively. Both PHQ and ISI have a maximum score of 27. 'ISI' and 'SWLS' scores indicate sleep disturbance and satisfaction with life, averaging 12.96 and 20.32 with a standard deviation of 7.04 and 7.36, respectively. Hyperacusis, or increased sensitivity to sound, has an average score of 19.04, while 'CFQ' averages at 40.59, possibly measuring cognitive failures. Participants' age averages at 55 years with a standard deviation of approximately 13 years. The duration of Tinnitus averages nearly 12 years, and TFI score at the beginning of the study averages at 59.37, with TFI score after the completion of the study having a lower average of 50.52, indicating a potential improvement after an intervention.

3.1. Correlation Analysis of Predictors

The correlogram provides a visual and quantitative depiction of the correlation coefficients between pairs of study variables. Correlation coefficients range from -1 to 1, where values close to 1 indicate a strong positive correlation, values close to -1 indicate a strong negative correlation, and values around 0 indicate no linear relationship. The color scale enhances interpretability: red shades signify positive correlations, blue shades indicate negative correlations, and the intensity of the color corresponds to the strength of the relationship.

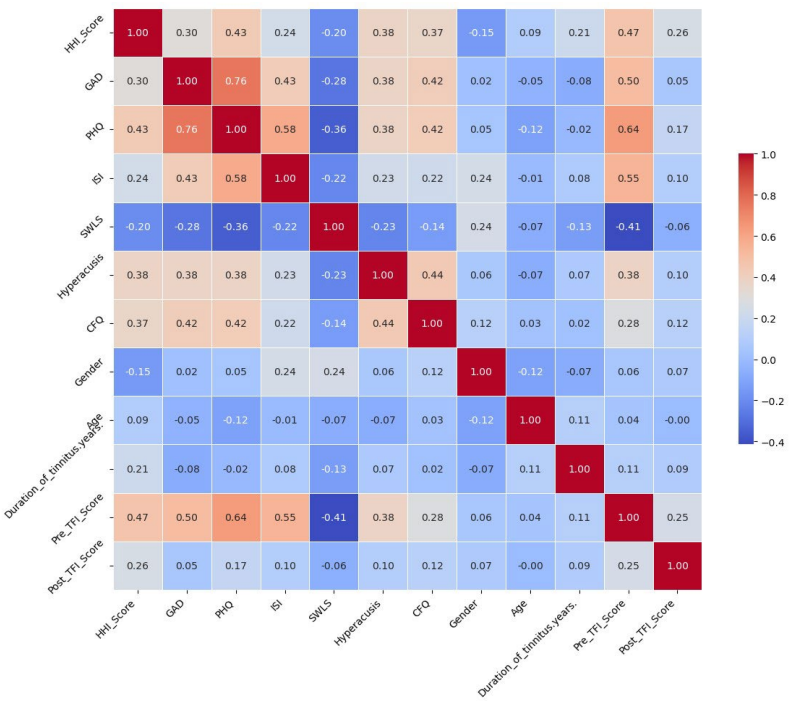


Figure 1: Correlogram of Variables

From Figure 1, notably strong positive correlations are observed between GAD and PHQ scores (0.76), illustrating that higher anxiety levels are associated with higher depression symptomatology. Similarly, PHQ scores and *TFI score at the beginning of the study* (0.64) are strongly correlated, suggesting a significant relationship between depression symptoms and the impact of Tinnitus before treatment. SWLS (Satisfaction with Life Scale) shows negative correlations with GAD, PHQ, and *TFI score at the beginning of the study*, indicating that higher life satisfaction is inversely related to anxiety, depression, and the perceived impact of Tinnitus. This underpins the psychological impact of Tinnitus and its comorbidities on life satisfaction. Moderate positive correlations were observed between GAD and *TFI score at the beginning of the*

study (0.50). Additionally, no linear relationship was observed between Age and *TFI score at the end of the study* (-0.00). The relationship between Hyperacusis and both PHQ and CFQ suggests that sensitivity to sound is linked with hearing challenges and cognitive failures.

3.2. Distribution of quantitative variables

From Figure 2, the 'HHI Score' histogram shows a slightly left-skewed distribution, while 'GAD' (Generalized Anxiety Disorder) and 'PHQ' (Patient Health Questionnaire) show a moderate right skew. The 'ISI' (Insomnia Severity Index), 'SWLS' (Satisfaction with Life Scale), and 'Hyperacusis' measures appear to be somewhat normally distributed, with 'SWLS' showing a slight left skew. The 'CFQ' (Cognitive Failures Questionnaire) histogram suggests a normal distribution with a peak around the 40 score mark. The 'Age' distribution is slightly right-skewed, with a concentration of participants in the middle-age range. The histogram for Duration of Tinnitus in years is highly right-skewed, with most participants having a shorter duration of Tinnitus. Lastly, the 'Pre TFI Score' and 'Post TFI Score' histograms, assessing the impact of Tinnitus before and after treatment, show that most participants' scores are moderately high with a right skew, and the 'Post TFI Score' appears to have a slightly more pronounced peak, indicating a possible concentration of scores post-treatment.

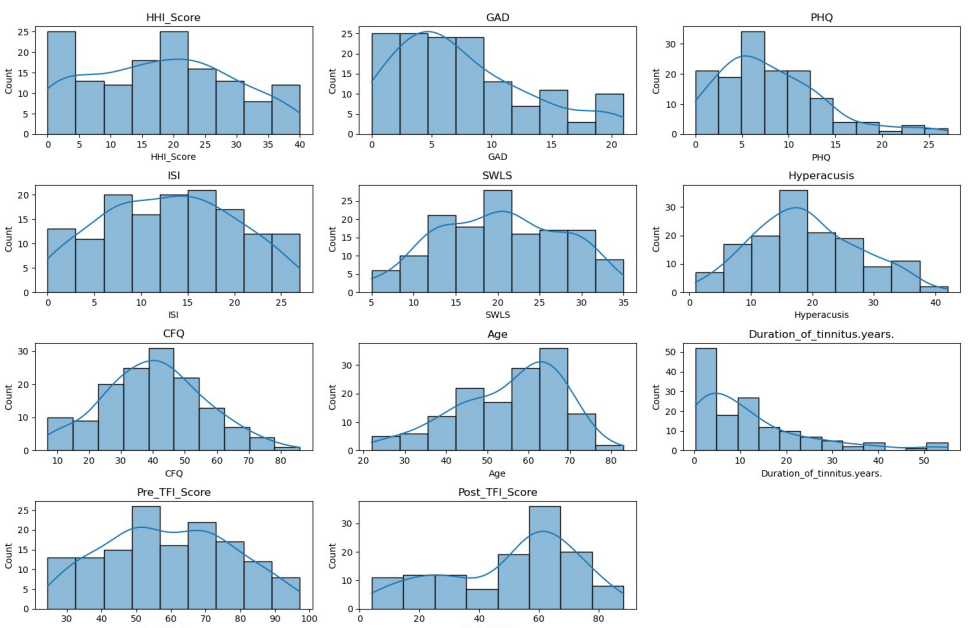


Figure 2: Histogram of quantitative variables

3.3. Descriptive Statistics of qualitative variables

Figure 3 presents two charts displaying the distribution of the study's participants across different groups and genders. The pie chart on the left shows the split between participants in the control group (73 individuals corresponding to 51.4%) and those in the treatment group (69 individuals corresponding to 48.6%). The bar chart on the right indicates the gender distribution, with a count of 80 male participants, making up 56.3% of the total, and 62 female participants, accounting for 43.7%. Both charts include the actual counts along with the corresponding percentages, providing a clear visual representation of the composition of the study's sample.

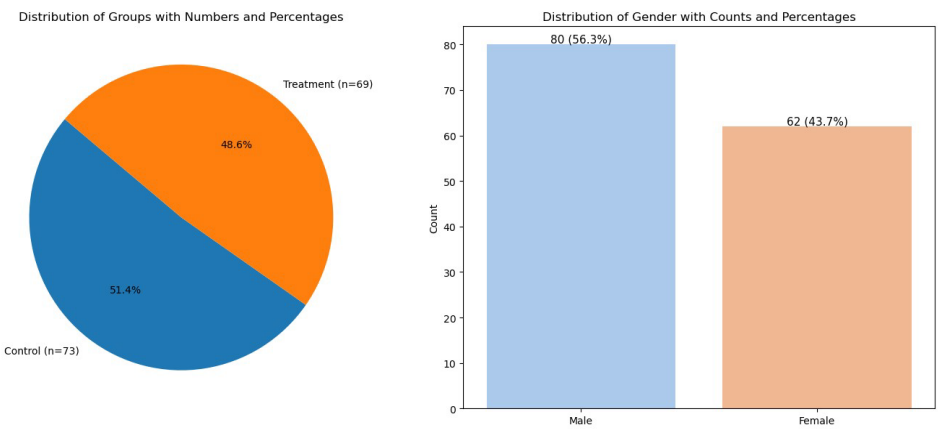


Figure 3: Distribution of Groups and Gender

3.4. Ordinary Least Squares Regression Model

In this section, we utilized the multiple linear regression with best subset selection to predict the TFI Reduction.

Table 3: Full Regression Model Coefficients

Variable	Estimate	Std. Error	$t-value$	$P>\|t\|$	[0.025	0.975]
Constant	-1.0735	13.109	-0.082	0.935	-27.078	24.931
HHI Score	-0.0960	0.196	-0.490	0.625	-0.485	0.293
GAD	0.1591	0.504	0.316	0.753	-0.841	1.159
PHQ	0.2950	0.602	0.490	0.625	-0.899	1.489
ISI	0.8482	0.333	2.551	0.012	0.189	1.508
SWLS	-0.5744	0.284	-2.022	0.046	-1.138	-0.011
Hyperacusis	0.2910	0.244	1.194	0.235	-0.193	0.775
CFQ	-0.0635	0.136	-0.468	0.641	-0.333	0.206

Table 3: Full Regression Model Coefficients (cont.)

Variable	Estimate	Std. Error	*t – value*	*P* > |*t*|	[0.025	0.975]
Gender	-2.1539	4.072	-0.529	0.598	-10.232	5.924
Age	-0.0281	0.146	-0.193	0.848	-0.318	0.262
Duration of Tinnitus (in years)	-0.0886	0.143	-0.620	0.537	-0.372	0.195
Group	25.3661	3.610	7.027	0.000	18.206	32.527

The full regression model in 3 presents the effects of various predictors on TFI reduction. The intercept, while estimated at -1.0735, is not statistically significant (p = 0.935), indicating the threshold TFI Reduction, in the absence of all the predictors. On the other hand, the coefficient of GAD implies that the TFI Reduction would increase by 0.1591 per unit change in SWLS when all other factors are kept constant. The variables with positive coefficients, i.e. PHQ, ISI and Hyperacusis can be interpreted in the same manner as GAD. This means that GAD, PHQ, ISI, Hyperacusis and TFI Reduction move in the same directions, i.e. TFI Reduction increases as they increase and also TFI Reduction decreases as they decrease. Additionally, when all other variables remain constant, the coefficient of HHI Score predicts that TFI Reduction would decrease by 0.0960 per unit change in HHI Score. It also suggests that TFI Reduction and HHI Score have a negative or inverse relationship. SWLS, CFQ, Age and Duration of Tinnitus (in years) all have a negative coefficient and hence can be explained in the same way as HHI _Score. Considering the categorical variables, we observed that the coefficient for gender (male set as the reference category) is -2.1539, which is not statistically significant at the 5% significance level (p = 0.598). A negative coefficient indicates that females, on average, have a lower TFI Reduction compared to males. Likewise, the coefficient for Group (treatment set as the reference category) is 25.3661, which is statistically significant at the 5% significance level (p = 0.000). A positive coefficient indicates that the control group has a higher TFI reduction compared to the treatment group. At the 5% significance level, we observed that ISS (p = 0.012), SWLS (0.046) and group (0.000) were the only significant predictors of TFI Reduction. The other predictors do not statistically contribute significantly to the prediction of TFI Reduction.

Table 4: Summary Statistics for Full Regression Model

*R*2	Adjusted-R^2	F-statistic	Prob (F-statistic)
0.448	0.388	7.451	0.0000

From Table 4, the R^2 value of 0.448 implies that approximately 44.80% of the variations in TFI Reduction was explained by all the predictors (HHI Score, GAD, PHQ, ISI, $\bar{S}\bar{W}$LS, Hyperacusis, CFQ, Gender, Age, Duration of Tinnitus (in years) and Group). Likewise, the Adjusted R^2 value of 0.388 implies that the percentage of variation explained by the independent variables that actually affect TFI Reduction is approximately 38.80%. The p-value of the F-statistic (p=.0000) which is less than 0.05 indicates that the full regression model is statistically significant. This means that at least one of the predictors contribute significantly to predicting TFI Reduction.

3.4.1. Best Subset Selection Criteria

In our pursuit to refine the predictive accuracy of the comprehensive model, the selection of predictors underwent a rigorous statistical evaluation. The criterion for this accurate selection was their statistical significance in predicting the reduction in TFI (TFI Reduction). To achieve this, we employed five distinct measures namely Akaike Information Criterion (AIC), Akaike Information Criterion corrected (AICc), Bayesian Information Criterion (BIC), Mallow's Criterion (C_p) and Final Prediction Error (FPE) as our guiding metrics. These criteria are instrumental in identifying models that strike an optimal balance between complexity and fit, thereby ensuring that only the most relevant predictors are included in the final model. This strategic approach aimed at both minimizing overfitting and enhancing the model's overall predictive capacity.

Now, we let X_1 =Group, X_2 = ISI, X_3 = SWLS, X_4 = Hyperacusis, X_5 =Duration of Tinnitus (in years), X_6 = Gender, X_7 = HHI Score, X_8 =PHQ, X_9 = CFQ, X_{10} =GAD and X_{11} = Age.

Table 5: Best Subsets Regression Summary

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	AIC	AICc	BIC	FPE	Cp
1 ✓											1005.44	1005.47	1008.16	428.28	40.44
2 ✓	✓										1004.55	1004.66	1010.01	424.95	38.62
3 ✓	**✓**	**✓**									**980.64**	**980.86**	**988.82**	**343.91**	**10.37**
4 ✓	✓	✓	✓								981.57	981.95	992.49	346.77	11.27
5 ✓	✓	✓	✓	✓							982.30	982.86	995.94	349.01	11.96
6 ✓	✓	✓	✓	✓	✓						983.85	984.64	1000.21	353.85	13.51
7 ✓	✓	✓	✓	✓	✓	✓					985.59	986.65	1004.67	359.35	15.24
8 ✓	✓	✓	✓	✓	✓	✓	✓				986.83	988.22	1008.65	363.37	16.48
9 ✓	✓	✓	✓	✓	✓	✓	✓	✓			988.57	990.32	1013.12	369.03	18.22
10 ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		990.47	992.63	1017.75	375.34	20.12
11 ✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	992.36	994.97	1022.36	381.72	22.01

NB: A ✓ indicates that a given variable is included in the corresponding model. For instance, the output in the first row in Table 5 above indicates that the model contains only Group as the predictor. Here, the intercept is included in all models.

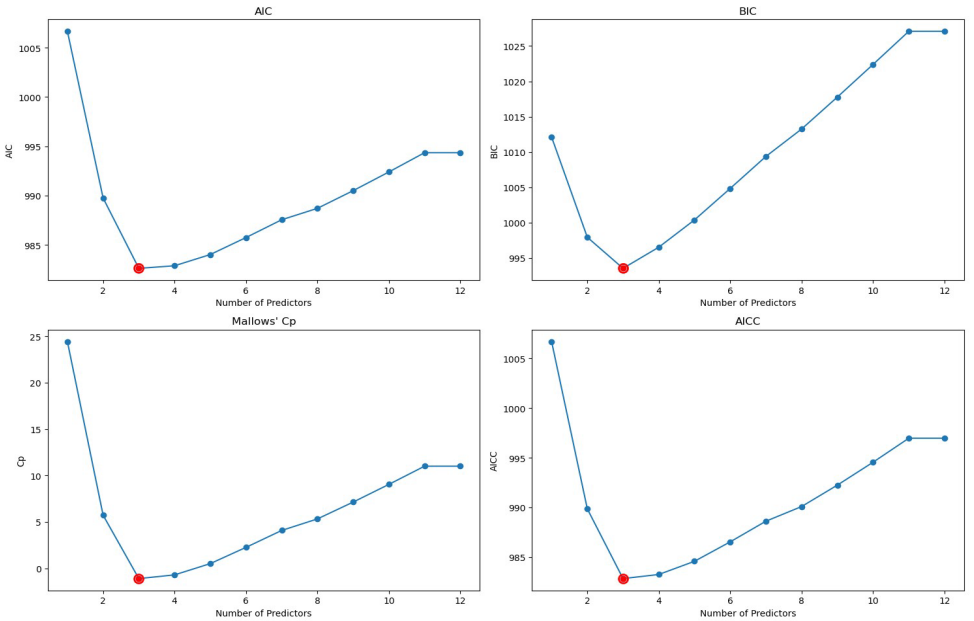


Figure 4: Plots of AIC, AICc, BIC and C(p) for *p* Subsets

From Table 5 and Figure 4, the best subsets of predictors with the least AIC, AICc, BIC and C(p) values include Group, ISI and SWLS. The FPE value in Table 5 also confirms this claim. It is also worth knowing that the results of both the forward and backward selection criteria (not included in this because of brevity) were also in conformity with the one obtained by best subset selection criteria.

The best subset model can be expressed mathematically in (7) as:

$$\textbf{TFI Reduction} = \beta_0 + \beta_1 * \textbf{Group} + \beta_2 * \textbf{ISI} + \beta_3 * \textbf{SWLS}. \quad (7)$$

Table 6 gives the coefficients and summary of the best subsets regression model.

Table 6: Model Coefficients for Best Subset Model from OLS Regression

Variable	Estimate	Std. Error	*t – value*	$P > \|t\|$	[0.025	0.975]
Constant	-0.9332	6.998	-0.132	0.895	-14.793	12.947
Group	25.6116	3.490	7.338	0.000***	18.694	32.529
ISI	0.9787	0.255	3.840	0.000***	0.474	1.484
SWLS	-0.7194	0.238	-3.028	0.003***	-1.190	-0.249

Based on the regression coefficients in Table 6, the linear regression equation of the best subset model is given in (8) by

$$\textbf{TFI Reduction} = -0.9232 - 25.6116 * \textbf{Group} + 0.9787 * \textbf{ISI} - 0.7194 * \textbf{SWLS}. \quad (8)$$

This relationship shows that TFI Reduction decreases with a corresponding change in SWLS when all other variables are kept constant. This means that a one unit increase in SWLS is expected to decrease TFI reduction by 0.7194. Also, TFI Reduction increases with a corresponding increase in ISI when all factors are kept constant. In general, a one unit increase in ISI is expected to increase TFI reduction by 0.9787. Both ISI (p=0.000) and SWLS (p=0.003) are statistically significant at the 5% significance. It was also observed that the coefficient for Group (treatment category as the reference category) is 25.6116, which is statistically significant at the 5% significance level (p = 0.000). A positive coefficient indicates that the control group has a higher TFI reduction compared to the treatment group.

Table 7: Summary Statistics for Best Subset Regression Model

$R2$	Adjusted-R^2	F-statistic	Prob (F-statistic)
0.427	0.411	27.04	0.0000

From Table 7, the R^2 value of 0.427 implies that approximately 42.70% of the variations in TFI Reduction was explained by Group, ISI and SWLS. Likewise, the Adjusted R^2 value of 0.411 implies that the percentage of variation explained by Group, ISI and SWLS that actually affect TFI Reduction is approximately 41.10%. The p-value of the F-statistic (p=.0000) which is less than 0.05 indicates that the best subset regression model is statistically significant. This means that at least one of the predictors contributes significantly to predicting TFI Reduction but in this case all the three predictors (Group, ISI and SWLS) statistically contribute significantly to the prediction of TFI Reduction. The increment in the value of the adjusted R^2 from 0.388 to 0.411 is due to the the removal of the non-significant predictors and thus improving the overall model fit.

3.5. Tests of Model Assumptions

Kutner et al. (2005) defines model validity as the stability and reasonableness of the regression coefficients, the plausibility and usability of the regression function and the ability to generalize inferences drawn from the regression analysis. Validation is a valuable and an essential part of the model building process. Torres-Reyna (2007) established that how good a model is depends on how well it predicts the dependent variable. The tests of the model assumptions are performed on the normality of the

error terms, presence of homoscedasticity, independence of the residuals, multicollinearity and linearity of the regression function.

3.5.1. Test for Normality of Residuals

An informal test of normality was conducted to check the distribution of the data using the Normal Q-Q plot. From Figure 5, it could be seen that most of the data values fall within the confidence bands with some few deviations showing that normality is a suspect. This informal test is further augmented by the Sharpiro Wilk formal test for normality.

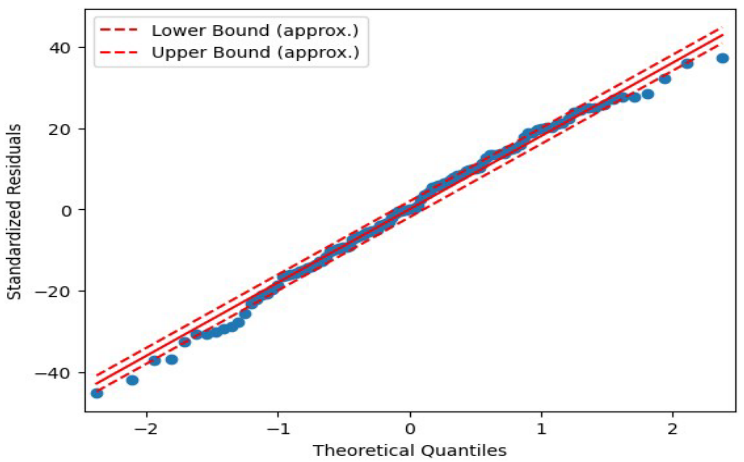


Figure 5: Normal Q-Q with Approximate Bounds

3.5.2. Sharpiro Wilk test for Normality

From Table 8, at the 5% significance level, since the p-value is greater than the alpha level (0.2195 > 0.05) we fail to reject the null hypothesis of normality satisfied and conclude that normality assumption is indeed satisfied. This confirms the initial claim **using the Q-Q plot.**

Table 8: Sharpiro Wilk test statistic and sig-value

Test	Test Statistic	P-value
Shapiro-Wilk	0.9845	0.2195

3.5.3. Test for homoscedasticity of the Residuals

As a prerequisite, the outcome of the test for normality of the error term above satisfies the condition to go on with the test for homoscedasticity. As an informal test, we resort to the scale-location plot.

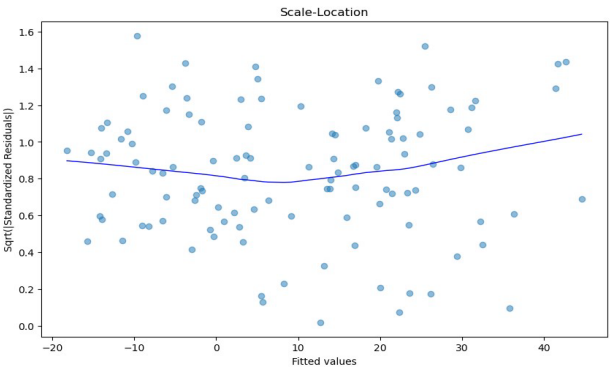


Figure 6: Plot of Scale Location

From Figure 6, we observe that at every fitted value, the spread of the residuals is roughly the same. We thus conclude that homoscedasticity seems plausible. Likewise in Figure 7, the spread of residuals remains constant across different fitted values indicative of homoscedasticity. The Breusch-Pagan test was employed as a formal test to verify our claim.

Table 9: Breusch-Pagan test statistic and sig-value

Test	Test Statistic	P-value
Breusch-Pagan	3.3205	0.3448

Using the Breusch-Pagan test, from Table 9 since the significance probability value = 0.3448 > 0.05, we fail to reject the null hypothesis of homoscedasticity, indicating no strong evidence of heteroscedasticity. This implies that heteroscedasticity is not present thus homoscedasticity assumption is satisfied.

3.5.4. Test for Multicollinearity

The collinearity test was conducted using the Variance Inflation Factor (VIF) approach. The resulting Collinearity statistics are presented in Table 10:

Table 10: Collinearity Statistics

Variable	VIF	VIF> 5
Group	1.0162	False
ISI	1.0654	False
SWLS	1.0551	False

From Table 10, since all the VIFs are all quite moderate ($VIF < 5$), there is no evidence of serious collinearity. Thus, the predictor variables (Group, ISI and SWLS) are not highly correlated with each other.

3.5.5. Test for Independence of Residuals

The Durbin-Watson test of autocorrelation was adopted to test for the independence of residuals of the best subset fitted regression model.

Table 11: Durbin-Watson Test Statistics and sig-value

Test	Lag	Test Statistic
Durbin-Watson	1	2.1254

From Table 11, since the Durbin-Watson Statistic is approximately 2, we do not reject the null hypothesis of no autocorrelation, and therefore, conclude that there is little to no auto or serial correlation in the residuals, indicating that the residuals are random (thus, the residuals are independent of each other).

3.5.6. Testing for linearity of the regression function

Here, we test to see if the regression function is linear using the Residual versus Fitted plot as an informal test.

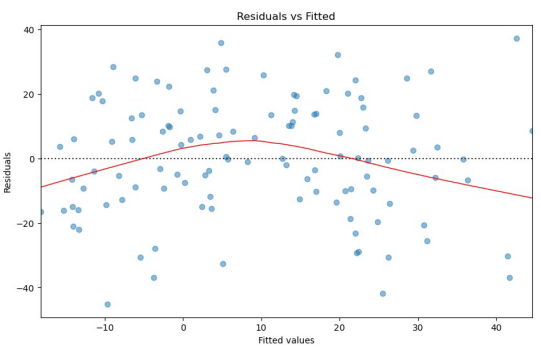


Figure 7: Residuals versus Fitted Plot

From the residuals versus fitted plot in Figure 7, we observe that at every fitted value, the spread of the residuals is roughly the same. We thus conclude that linearity seems plausible. From Figure 7, we also observe that there is no pattern in the residual plot. This suggests that we can assume a linear relationship between the predictors and the outcome variables. Since this is the case, linearity is a suspect. The Harvey-Collier test is then employed as a formal test to validate our suspicion.

Table 12: Harvey-Collier test statistic and sig-value

Test	Test Statistic	P-value
Harvey-Collier	-0.3094	0.7576

From Table 12, we clearly see that since the p-value is greater than the indicated alpha level (0.7576 > 0.05), we do not reject the null hypothesis of linearity, indicating no strong evidence of non-linearity. Hence, the regression function is linear.

In summary, since the best subset regression model satisfies all the assumptions it can be used for predictions based on the test set. The mean square error (MSE) using the test set was found to be **330.76**.

3.6. KNN Regression Results and Data Analysis

In addition to employing the conventional least squares estimation technique, our analysis extends to incorporating KNN regression to assess the impact of the independent variables on the target variable, TFI Reduction in this section. Unlike the traditional regression methods, KNN regression does not presuppose a specific functional form for the outcome variable. This deviation provides enhanced flexibility, allowing the model to adapt more freely to the underlying data structure. The primary objective of integrating this method is to compare and ascertain which algorithm, between least squares and KNN regression, exhibits superior predictive accuracy for the Tinnitus dataset in question. By adopting this dual approach, we aim to gain deeper insights into the dynamics influencing TFI Reduction and identify the most effective predictive framework based on the characteristics of the available data.

In this section, we present 5 different KNN regression models fitted for K = 2, 4, 6, 8, 10 and their test mean squared errors are presented in Figure 8 and Table 13.

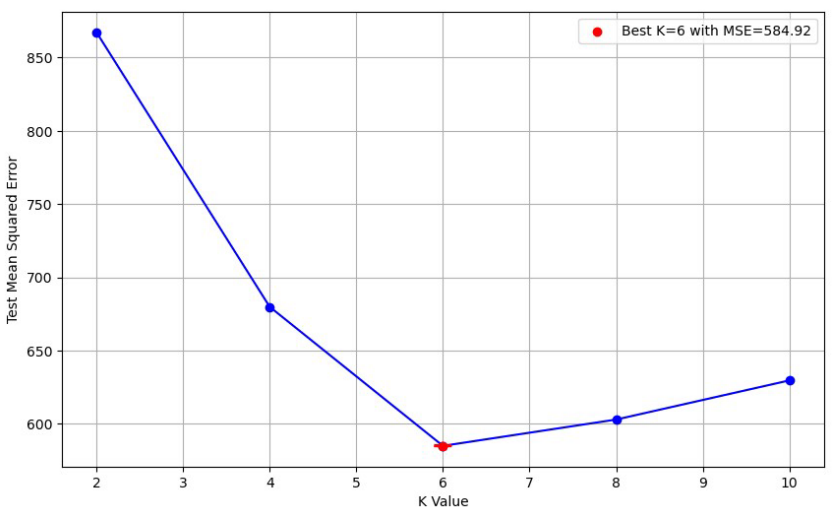


Figure 8: KNN Regression: Plot of Test MSE for Different K values

The outcomes of the cross-validation process from both Figure 8 and Table 13 indicate that the choice of $K = 6$ results in the minimal mean squared error on the test data. Consequently, this particular value was selected for our predictive analysis concerning TFI Reduction using the predictor variables. Given the complexities involved in formulating a direct equation to express TFI Reduction as a function of

these predictors, we opted to employ a K-Nearest Neighbors (KNN) regression model. By applying this model with K set to 6 for our test dataset, we were able to compute the predictions. The performance of these predictions was quantified using the Mean Squared Error (MSE), which amounted to **584.92**, providing a quantitative measure of the prediction accuracy for the TFI Reduction based on the selected predictors.

Table 13: Test MSE for Different K values

K-value	2	4	6	8	10
Test MSE	867.26	679.83	584.92	602.88	629.60

3.7. Lasso and Ridge Regression Analysis

In this section of our analysis, we explore the efficacy of shrinkage methodologies—specifically Ridge and Lasso regression—in predicting the impact of various predictors on the TFI Reduction. The underpinning rationale for employing these techniques lies in their capacity to impose a shrinkage penalty on the coefficients, thereby mitigating the risk of overfitting and enhancing model generalizability compared to the traditional Ordinary Least Squares (OLS) approach. Employing 10-fold cross-validation, we embarked on a rigorous search for the optimal λ (tuning parameter), which minimizes the Mean Squared Error (MSE). This endeavor is not merely a quest for minimal error but a strategic move to discern the most robust model that harmonizes complexity and prediction accuracy. Figure 9 offers a visual exposition of this optimization process, showcasing the relationship between various λ values and their MSE outcomes. This visual analysis is important as it illuminates the trade-offs inherent in model regularization and aids in the selection of a model that judiciously balances bias and variance.

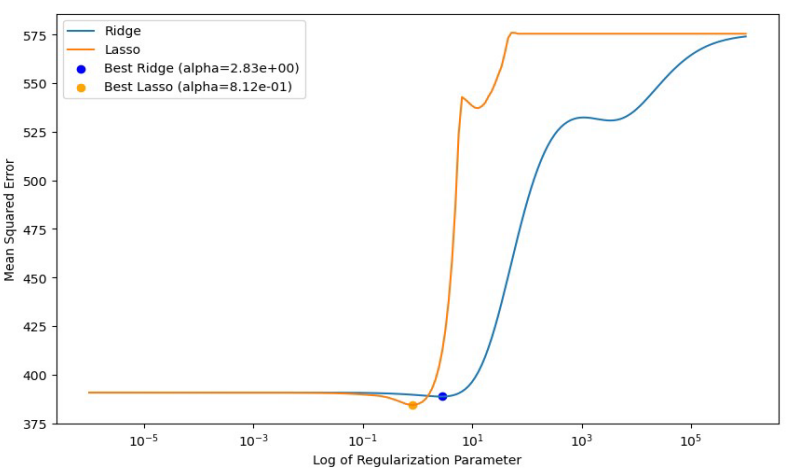


Figure 9: Cross-Validated λ Values for Ridge and Lasso Regression

From the cross-validation plot, the optimal λ-value with the least MSE for each method (Ridge and Lasso) is 2.83 for Ridge and 0.812 for Lasso. We made use of these values for the tuning parameter, λ in each method, to build the model and also to perform the prediction for the testing data.

3.8. Coefficients Estimate and Testing Errors Using the Optimal λ-Values

Table 14 presents the coefficient estimates for each model and their respective test errors of the final model produced by the optimal λ-values.

Table 14: Model Coefficients and MSEs for Ridge and Lasso Shrinkage Methods

Predictors	Ridge (Coefficients)	Lasso (Coefficients)
(Intercept)	0.35578	-1.11778
HHI Score	-0.10206	-0.0633
GAD	0.18036	0.18863
PHQ	0.28242	0.24065
ISI	0.82577	0.77362
SWLS	-0.58996	-0.62518
Hyperacusis	0.28859	0.26886
CFQ	-0.06558	-0.07340
Gender	-2.07104	.
Age	-0.01991	-0.01068
Duration of Tinnitus	-0.09540	-0.09591
Group	22.97202	22.12796
Testing Error	**318.30088**	**319.28195**

The coefficients indicate the relationship between each predictor and the response variable (TFI Reduction). Positive coefficients suggest a direct relationship, while negative coefficients suggest an inverse relationship. Common trends observed in both models include the negative impact of the HHI Score, SWLS, CFQ, Age and Duration of Tinnitus (in years); and the positive impact of measures like PHQ, ISI, and Hyperacusis on TFI Reduction. Differences between the models are evident in the magnitude of the coefficients, with Lasso regression driving some coefficients to zero (e.g., Gender in Lasso has no estimate), indicating it performs variable selection. The MSE for Ridge regression is slightly lower **(318.30)** than for Lasso regression **(319.28)**, suggesting that Ridge regression might be slightly more effective in predicting TFI Reduction with the given dataset and chosen λ values. Both Age and duration of Tinnitus have a slight negative association with TFI Reduction in both models, suggesting that older individuals and those who have had Tinnitus for longer period

might experience less reduction. A positive coefficient for Group in both models suggests that the control group has a higher TFI reduction compared to the treatment group.

4. Discussion and Study Alignment with Health Economics

This study introduces several novel aspects in terms of statistical analysis and methodology that contribute to the existing body of literature on Tinnitus management and treatment effectiveness. It also enriches the statistical analysis of Tinnitus treatment effectiveness through the application of advanced machine learning techniques, rigorous model validation, and comprehensive residual analysis. The study employs a variety of supervised machine learning algorithms, including Ordinary Least Squares (OLS), K-Nearest Neighbor (KNN), Ridge regression, and Lasso regression. Each method is analyzed for its predictive accuracy in modelling Tinnitus Functional Index (TFI) reduction. The inclusion of these diverse algorithms allows for a comprehensive comparison of their performance, highlighting the strengths and weaknesses of each approach in the context of Tinnitus treatment data. Secondly, the application of Ridge and Lasso regressions introduces regularization techniques that address multicollinearity and improve model stability. These methods add a shrinkage penalty to the regression coefficients, which helps in reducing the variance of the estimates. Ridge regression demonstrated superior performance with the smallest test mean square error (MSE) of 318.30, showcasing its effectiveness in handling multicollinear predictors and providing more reliable predictions. Also, the study employs the best subset selection criteria to identify the most significant predictors of TFI reduction. This involves evaluating all possible combinations of predictors and selecting the model that optimally balances complexity and predictive power. The criteria used include Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Mallow's Cp, which are instrumental in ensuring the selected model is both parsimonious and highly predictive.

Moreover, the research emphasizes rigorous model validation techniques, including 10-fold cross-validation, to assess the robustness and generalizability of the models. This approach helps in mitigating overfitting and ensures that the models perform well on unseen data. Various statistical metrics, such as R-squared, adjusted R-squared, and F-statistics, are used to evaluate model fit and significance, providing a detailed understanding of model performance. Additionally, the study conducts extensive residual analysis to check the assumptions underlying the regression models. This includes tests for normality (Shapiro-Wilk test), homoscedasticity (Breusch-Pagan test), independence of residuals (Durbin-Watson test), and multicollinearity (Variance Inflation Factor). Such thorough diagnostic checks ensure the validity and reliability of

the regression models, enhancing the credibility of the findings. To add to the above, the use of correlograms and histograms provides visual insights into the relationships between variables and their distributions. These visual tools help in understanding the data structure and the nature of the predictors. The study also presents visualizations of the cross-validation process for selecting the optimal lambda values in Ridge and Lasso regressions, aiding in the interpretation of the regularization effects. The statistical findings highlight the importance of a holistic approach to Tinnitus management, considering factors beyond just the auditory symptoms. Aside from the statistical implications, the study aligns closely with the field of health economics in several significant ways such as cost-effectiveness analysis, resource allocation, quality of life and economic productivity and long-term economic benefits.

By identifying significant predictors of TFI reduction, the study facilitates the development of more targeted treatment strategies. This can lead to more efficient use of healthcare resources, reducing unnecessary treatments and associated costs. Machine learning models, such as Ridge and Lasso regressions, enable the prediction of treatment outcomes with higher accuracy, allowing healthcare providers to allocate resources more effectively and prioritize interventions that are likely to yield the best results. The findings of this study can inform policy decisions regarding the allocation of resources for Tinnitus treatment. By understanding which factors most significantly influence treatment outcomes, policymakers can prioritize funding and resources towards interventions that address these key areas. Efficient resource allocation based on predictive modeling can help reduce the economic burden of Tinnitus on both patients and the healthcare system as indicated by Tuepker et al. (2018). Tinnitus, particularly when severe, can significantly impair an individual's quality of life and economic productivity. By improving the precision of treatment outcomes, the research contributes to enhancing patients' quality of life, which in turn can have positive economic implications. Better management of Tinnitus can lead to reduced absenteeism and increased productivity among individuals affected by the condition, contributing to broader economic benefits.

5. Conclusion and Recommendations

The main objective of the study was to model Tinnitus Functional Index (TFI) Reduction via supervised machine learning algorithms. Notably, Ordinary Least Squares (OLS), Nearest Neighbor (KNN), Ridge and Lasso regressions were the main statistical tools adopted for the study. The OLS regression revealed that Group, ISI and SWLS were the main significant predictors of TFI Reduction using the best subset selection criteria which was also confirmed by both the forward and backward selection criteria. It was found out that the two shrinkage methods (Ridge and Lasso)

outperformed the OLS and the KNN regression. Based on the test mean square error (MSE), Ridge regression was chosen as the best supervised machine learning algorithm for the analysis of the Tinnitus data under consideration, with the smallest test MSE of **318.30** compared to that of Lasso, OLS, KNN regression's respective test MSE of **319.28**, **330.76** and **584.92**. Hence, by the metric considered, Ridge regression was ranked more accurate in predicting TFI Reduction than the other three methods. The exploration of supervised machine learning algorithms for modeling Tinnitus Functional Index (TFI) reduction presents a promising frontier in the personalized treatment of Tinnitus. Studies such as those by Rodrigo et al. (2021), and investigations into the predictive capabilities of EEG sensors and deep learning algorithms by Doborjeh et al. (2023), demonstrate significant advancements in predicting patient responses to therapies and classifying Tinnitus severity. These machine learning approaches, including decision tree models, random forests, and neural networks, offer new insights into the complex nature of Tinnitus and its management. By effectively predicting treatment outcomes, these tools can facilitate more targeted and effective interventions, ultimately improving the quality of life for individuals suffering from Tinnitus.

As research continues to evolve, the integration of AI and machine learning into Tinnitus treatment protocols holds the potential to revolutionize the field, offering hope for effective management strategies for this challenging condition. The study revealed a significant difference between the treatment and control groups, with the control group showing a higher TFI reduction. This finding may suggest that factors outside the treatment regimen, such as individual coping mechanisms or placebo effects, could influence TFI outcomes. This warrants further investigation to understand the underlying reasons for the observed group differences. Future studies could investigate the integration of multimodal data, such as combining EEG patterns with behavioral and clinical metrics, to develop more comprehensive predictive models. Additionally, exploring the longitudinal application of machine learning algorithms could offer insights into the progression of Tinnitus over time and the long-term effectiveness of different treatment modalities. There is also a significant opportunity to refine machine learning models by incorporating patient feedback loops, enabling the models to learn from each treatment outcome and continuously improve predictions. The study's recommendation to explore the integration of multimodal data and longitudinal applications of machine learning algorithms has long-term economic benefits. Continuous improvement in predictive capabilities can lead to sustained enhancements in treatment strategies, reducing long-term healthcare costs. Investing in such modernistic research aligns with health economics principles of achieving long-term cost savings and health improvements through advanced technological solutions.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability statement

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Additional Information

No additional information is available for this paper.

Acknowledgement

The author acknowledges the enormous support of the University of Texas Rio Grande Valley (UTRGV) Presidential Research Fellowship.

References

Agbota, L., Agyemang, E., Kissi-Appiah, P., Moshood, L., Osei-Nkwantabisa, A., Agbenyeavu, V., Nsiah, A., and Adjei, A. (2024). Enhancing tumor classification through machine learning algorithms for breast cancer diagnosis. *Computer Engineering and Intelligent Systems*, 15(1), p. 71.

Biswas, R., Genitsaridi, E., Trpchevska, N., Lugo, A., Schlee, W., Cederroth, C. R., Gallus, S., and Hall, D. A., (2023). Low evidence for tinnitus risk factors: A systematic review and meta-analysis. *Journal of the Association for Research in Otolaryngology*, 24(1), pp. 81–94.

Cardon, E., Jacquemin, L., Schecklmann, M., Langguth, B., Mertens, G., Vanderveken, O. M., Lammers, M., Van de Heyning, P., Van Rompaey, V., and Gilles, A., (2022). Random forest classification to predict response to high-definition transcranial direct current stimulation for tinnitus relief: A preliminary feasibility study. *Ear and Hearing*, 43(6), p. 1816.

De Ridder, D., Schlee, W., Vanneste, S., Londero, A., Weisz, N., Kleinjung, T., Shekhawat, G. S., Elgoyhen, A. B., Song, J.-J., Andersson, G., et al., (2021). Tinnitus and tinnitus disorder: Theoretical and operational definitions (an international multidisciplinary proposal). *Progress in brain research*, 260, pp. 1–25.

Doborjeh, M., Liu, X., Doborjeh, Z., Shen, Y., Searchfield, G., Sanders, P., Wang, G. Y., Sumich, A., and Yan, W. Q., (2023). Prediction of tinnitus treatment outcomes based on eeg sensors and tfi score using deep learning. *Sensors*, 23(2), p. 902.

Eggermont, J. J., Roberts, L. E., (2004). The neuroscience of tinnitus. *Trends in neurosciences*, 27(11), pp. 676–682.

Gasparre, D., Pepe, I., Laera, D., Abbatantuono, C., De Caro, M. F., Taurino, A., D'Erasmo, D., Fanizzi, P., Antonucci, L. A., Pantaleo, A., et al., (2023). Cognitive functioning and psychosomatic syndromes in a subjective tinnitus sample. *Frontiers in Psychology*, 14.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W., (2005). *Applied linear statistical models*. McGraw-hill.

Langguth, B., Kreuzer, P. M., Kleinjung, T., and De Ridder, D., (2013). Tinnitus: causes and clinical management. *The Lancet Neurology*, 12(9), pp. 920–930.

Manta, O., Sarafidis, M., Schlee, W., Mazurek, B., Matsopoulos, G. K., and Koutsouris, D. D., (2023). Development of machine-learning models for tinnitus-related distress classification using wavelet-transformed auditory evoked potential signals and clinical data. *Journal of Clinical Medicine*, 12(11), p. 3843.

McCormack, A., Edmondson-Jones, M., Somerset, S., and Hall, D., (2016). A systematic review of the reporting of tinnitus prevalence and severity. *Hearing research*, 337, pp. 70–79.

Mohan, A., Leong, S. L., De Ridder, D., and Vanneste, S., (2022). Symptom dimensions to address heterogeneity in tinnitus. *Neuroscience & Biobehavioral Reviews*, 134, p 104542.

Omer, P. A., (2022). Improving prediction accuracy of lasso and ridge regression as an alternative to ls regression to identify variable selection problems. *Zanco Journal of Pure and Applied Sciences*, 34(s6), pp. 33–45.

Rodrigo, H., Beukes, E. W., Andersson, G., and Manchaiah, V., (2021). Exploratory data mining techniques (decision tree models) for examining the impact of internet-based cognitive behavioral therapy for tinnitus: Machine learning approach. *Journal of Medical Internet Research*, 23(11), e28999.

Runciman, L., Johnson, C., (2023). Young adults' knowledge and perceptions of permanent noise-induced tinnitus and its influence on behavioural intentions. *Noise and Health*, 25(119), pp. 236–246.

Saleh, A. M. E., Norouzirad, M., (2018). On shrinkage estimation: Non-orthogonal case. *Statistics, Optimization & Information Computing*, 6(3), pp. 427–451.

Smith, S. S., Kitterick, P. T., Scutt, P., Baguley, D. M., and Pierzycki, R. H., (2021). An exploration of psychological symptom-based phenotyping of adult cochlear implant users with and without tinnitus using a machine learning approach. *Progress in Brain Research*, 260, pp. 283–300.

Torres-Reyna, O., (2007). Linear regression using stata. *Princeton University. New Jersey: Princeton University*.

Tuepker, A., Elnitsky, C., Newell, S., Zaugg, T., and Henry, J. A., (2018). A qualitative study of implementation and adaptations to progressive tinnitus management (ptm) delivery. *PLoS One*, 13(5), e0196105.

Van den Berge, M. J., Free, R. H., Arnold, R., De Kleine, E., Hofman, R., Van Dijk, J. M. C., and Van Dijk, P., (2017). Cluster analysis to identify possible subgroups in tinnitus patients. *Frontiers in neurology*, 8, p. 115.

Wakabayashi, S., Oishi, N., Shinden, S., and Ogawa, K., (2020). Factor analysis and evaluation of each item of the tinnitus handicap inventory. *Head & face medicine*, 16, pp. 1–9.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4 pp. 79–97, https://doi.org/10.59139/stattrans-2024-004
Received – 13.06.2022; accepted – 18.01.2024

Efficient use of auxiliary information in estimating finite population variance in sample surveys

Housila P. Singh[1], **Rajesh Tailor**[2], **Priyanka Malviya**[3]

Abstract

This paper addresses the problem of estimating the finite population variance of the study variable y using information on the known population variance of the auxiliary variable x in sample surveys. We have suggested a class of estimators for population variance using information on population variance of x. The bias and mean squared error of the suggested class of estimators up to first order of approximation was obtained. Preference regions were derived under which the suggested class of estimators is more efficient than the usual unbiased estimator, Das and Tripathi (1980) estimators, Isaki (1983) ratio estimator, Singh et al (1973, 1988) estimator and Gupta and Shabbir (2007) estimator. An empirical study as well as simulation study were carried out in support of the present study.

Key words: study variable, auxiliary variable, class of estimators, bias, mean squared error.

1. Introduction

It is tradition to use the auxiliary information at the estimation stage in improving the precision of the estimates of population parameters such as mean and variance. A large amount of work has been carried out towards the estimation of population mean $\bar{Y}$ of the study variable y in the presence of auxiliary information by various authors including Cochran (1940), Robson (1957), Singh, M.P. (1965, 1967), Srivastava (1971, 1980), Srivastava and Jhajj (1980), Sahai and Sahai (1985), Ray and Singh (1981), Gupta (1978), Adhvaryu and Gupta (1983), Singh and Upadhyaya (1986), Singh, H. P. (1986, 1987), Singh and Singh (1984), Tracy et al (1996), Bahl and Tuteja (1991), Singh, S. (2003), Reddy (1978), Walsh (1970), Vos (1980), Singh and Ruiz Espejo (2003), Singh

[1] School of Studies in Statistics, Vikram University, Ujjain, (M. P.), India. E-mail: hpsujn@gmail.com. ORCID: https://orcid.org/0000-0002-7816-9936.

[2] School of Studies in Statistics, Vikram University, Ujjain, (M. P.), India. E-mail: tailorraj@gmail.com. ORCID: https://orcid.org/0000-0003-2097-7313.

[3] School of Studies in Data Science and Forecasting, Devi Ahilya Vishwavidyalaya Indora, (M. P.), India. E-mail: sarsodiapriyanka@gmail.com. ORCID: E-mail: https://orcid.org/0000-0001-5241-8300.

and Tailor (2005), Singh et al (2012), Singh and Yadav (2020) and Singh and Nigam (2020) and the references cited therein. However in many situations of practical importance, the problem of estimation of population variance S_y^2 of the study variable y deserves special attention. Singh, Pandey and Hirano (1973) and Searls and Intrapanich (1990) forwarded an improved estimator that utilizes the kurtosis $\left(\beta_2(y)\right)$ of the study variable y. Later various authors including Das and Tripathi (1978), Isaki (1983), Srivastava and Jhajj (1980), Singh, Upadhyaya and Namjoshi (1988), Gupta and Shabbir (2007), Solanki and Singh (2013), Singh and Solanki (2013a, 2013b), Yadav et al (2013), Pal and Singh (2018), Singh et al. (2003) among others, have paid their attention towards the estimation of population variance S_y^2 of the study variable y using information on population variance S_x^2 of the auxiliary variable x and suggested different estimators for population variance S_y^2. The goal of this paper is to suggest a new class of estimators for population variance S_y^2 utilizing the knowledge on population variance S_x^2 of the auxiliary variable x. The properties of the envisaged class of estimators up to the first order of approximation are studied. The present study is supported through numerical illustration.

2. Notations and Expected Values

Let $U = \{U_1, U_2, \ldots, U_N\}$ be a finite population of N units. Let y and x be the study and auxiliary variables respectively. The aim is to estimate the population variance S_y^2 of y using information on population variance S_x^2 of x. A simple random sample (SRS) of size n ($<N$) is drawn from U without replacement (WOR) to estimate S_y^2 of y when S_x^2 of x is known. Let us denote:

$S_y^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - \bar{Y})^2$: The population variance/mean square of the study variable y;

$S_x^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(x_i - \bar{X})^2$: The population variance/mean square of x,

$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$: The population mean of y,

$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$: The population mean of x,

$\beta_2(y) = \frac{\sum_{i=1}^{N}(y_i-\bar{Y})^4}{\left(\sum_{i=1}^{N}(y_i-\bar{Y})^2\right)^2}$: The coefficient of kurtosis of y,

$\beta_2(x) = \frac{\sum_{i=1}^{N}(x_i-\bar{X})^4}{\left(\sum_{i=1}^{N}(x_i-\bar{X})^2\right)^2}$: The coefficient of kurtosis of x,

$$\gamma = \frac{(N-1)\sum_{i=1}^{N}(y_i - \bar{Y})^2(x_i - \bar{X})^2}{(\sum_{i=1}^{N}(y_i - \bar{Y})^2)(\sum_{i=1}^{N}(x_i - \bar{X})^2)}$$

Population size N is large enough so that the finite population correction (fpc) term $\left(1-\frac{n}{N}\right)=(1-f)\cong 1$ is ignored and

$S_y^2=\mu_2(y),\ S_x^2=\mu_2(x), \beta_2(y)=\frac{\mu_4(y)}{\mu_2^2(y)}=\frac{\mu_4(y)}{S_y^4},$

$\beta_2(x)=\frac{\mu_4(x)}{\mu_2^2(x)}=\frac{\mu_4(x)}{S_x^4},$

$\gamma=\frac{\mu_{22}(y,x)}{\mu_2(y)\mu_2(x)}=\frac{\mu_{22}(y,x)}{S_y^2S_x^2}, \mu_2(y)=\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y})^2,$

$\mu_2(x)=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{X})^2,$

$\mu_{22}(y,x)=\frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{Y})^2\,(x_i-\bar{X})^2.$

For a SRS of size n, we have

$s_y^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2$: sample variance/mean square of *y*,

$s_x^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2$: sample variance/mean square of *x*,

$s_y^2=S_y^2(1+e_0), s_x^2=S_x^2(1+e_1)$

such that

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

and to the first degree of approximation and ignoring fpc we have

$E(e_0^2)=\frac{1}{n}(\beta_2(y)-1), E(e_1^2)=\frac{1}{n}(\beta_2(x)-1)$ and $E(e_0e_1)=\frac{1}{n}(\gamma-1).$

3. Reviewing Some Existing Estimators

The usual unbiased estimator of S_y^2 is given by

$$t_0=s_y^2=\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i-\bar{y})^2. \tag{3.1}$$

The variance /mean square of s_y^2 under SRSWOR scheme (ignoring fpc) is given by

$$MSE(t_0)=\frac{S_y^4}{n}(\beta_2(y)-1). \tag{3.2}$$

Utilizing knowledge on the kurtosis $\beta_2(y)$ of y Singh, Pandey and Hirano (1973) and Searls and Intrapanich (1991) envisaged the following class of estimators for S_y^2 as

$$t_{SPH}=ws_y^2, \tag{3.3}$$

where w is a constant such that the mean squared error (MSE) of t_{SPH} is minimum.

The mean squared error of t_{SPH} ignoring fpc is given by

$$MSE(t_{SPH})=S_y^4\left[1+w^2\left\{1+\frac{\beta_2(y)-1}{n}\right\}-2w\right] \tag{3.4}$$

which is minimum when

$$w=\frac{n}{(n+\beta_2(y)-1)}=w_{(opt)}\ \text{(say)}. \tag{3.5}$$

This leads to the resulting estimator

$$t_{SPH} = \frac{n}{(n+\beta_2(y)-1)} s_y^2. \tag{3.6}$$

Substitution of $w_{(opt)}$ at (3.5) in (3.4) yields the MSE of t_{SPH}as

$$MSE(t_{SPH}) = S_y^4 \frac{(\beta_2(y)-1)}{(n+\beta_2(y)-1)}. \tag{3.7}$$

When population variance S_x^2 of x is known, Isaki (1983) suggested a ratio estimator for population variance S_y^2 of y as

$$t_R = s_y^2 \frac{S_x^2}{s_x^2}. \tag{3.8}$$

To the first degree of approximation, the MSE of the ratio estimator t_R ignoring fpc term is given by

$$MSE(t_R) = \frac{S_y^4}{n}[\beta_2(y) + \beta_2(x) - 2\gamma]. \tag{3.9}$$

When S_x^2 is known, Das and Tripathi (1978) suggested the following classes of estimators of S_y^2 as

$$t_{DT1} = s_y^2 \left(\frac{S_x^2}{s_x^2}\right)^{\alpha} \tag{3.10}$$

and

$$t_{DT2} = s_y^2 \frac{S_x^2}{\{S_x^2+\alpha(s_x^2-S_x^2)\}}, \tag{3.11}$$

where α being suitably chosen constant.

The common minimum MSE of t_{DTi} (i=1,2) to the first degree of approximation, is given by

$$MSE_{\min}(t_{DTi}) = \frac{S_y^4}{n}\left[(\beta_2(y) - 1) - \frac{(\gamma-1)^2}{(\beta_2(x)-1)}\right] \tag{3.12}$$

which equals to the minimum MSE of the difference estimator

$$t_D = s_y^2 + d(S_x^2 - s_x^2),$$

where 'd' is a suitable chosen constant to be determined such that MSE of t_D is the least.

Singh, Upadhyaya and Namjoshi (1988) proposed a class of difference type estimators for S_y^2 as

$$t_{SUN} = w_1 s_y^2 + w_2(S_x^2 - s_x^2), \tag{3.14}$$

where (w_1, w_2) are suitable chosen constants.

The mean squared error of the estimator t_{SUN} ignoring fpc term is given by

$$MSE(t_{SUN}) = S_y^4[1 + w_1^2 a_1 + w_2^2 a_2 - 2w_1 w_2 a_3 - 2w_1], \tag{3.15}$$

where

$$a_1 = \left[1 + \frac{1}{n}(\beta_2(y) - 1)\right], a_2 = \frac{1}{nR^2}(\beta_2(x) - 1), a_3 = \frac{1}{nR}(\gamma - 1), \quad R = \frac{S_y^2}{S_x^2}.$$

The MSE(t_{SUN}) at (3.15) is minimum when

$$\left.\begin{aligned} w_1 &= \frac{a_2}{(a_1a_2-a_3^2)} = w_{10}(say) \\ w_2 &= -\frac{a_3}{(a_1a_2-a_3^2)} = w_{20}(say) \end{aligned}\right\}. \tag{3.16}$$

Thus, the resulting minimum MSE of t_{SUN} is given by

$$MSE_{min}(t_{SUN}) = S_y^4\left[1 - \frac{a_2}{(a_1a_2-a_3^2)}\right]. \tag{3.17}$$

Gupta and Shabbir (2007) envisaged the following class of estimators for S_y^2 as

$$t_{GS} = \left[w_1s_y^2 + w_2(S_x^2 - s_x^2)\right] exp\left(\frac{S_x^2-s_x^2}{S_x^2+s_x^2}\right), \tag{3.18}$$

where (w_1, w_2) are suitable chosen constants.

The MSE of t_{GS} to the first degree of approximation (ignoring fpc term) is given by

$$MSE(t_{GS}) = S_y^4[1 + w_1^2b_1 + w_2^2b_2 + 2w_1w_2b_3 - 2w_1b_4 - 2w_2b_5], \tag{3.19}$$

where

$$b_1 = \left[1 + \frac{1}{n}(\beta_2(y) + \beta_2(x) - 2\gamma)\right], b_2 = \frac{1}{nR^2}(\beta_2(x) - 1),$$
$$b_3 = \frac{1}{nR}(\beta_2(x) - \gamma), b_4 = \left[1 + \frac{1}{2n}\left\{\frac{3}{4}(\beta_2(x) - 1) - \gamma + 1\right\}\right],$$
$$b_5 = \frac{1}{2nR}(\beta_2(x) - 1).$$

The $MSE(t_{GS})$ at (3.19) is minimum when

$$\left.\begin{aligned} w_1 &= \frac{(b_2b_4-b_3b_5)}{(b_1b_2-b_3^2)} = w_{10(1)}(say) \\ w_2 &= -\frac{(b_1b_5-b_3b_4)}{(b_1b_2-b_3^2)} = w_{20(2)}(say) \end{aligned}\right\}. \tag{3.20}$$

Thus, the least MSE of t_{GS} is given by

$$MSE_{min}(t_{GS}) = S_y^4\left[1 - \frac{(b_2b_4^2-2b_3b_4b_5+b_1b_5^2)}{(b_1b_2-b_3^2)}\right]. \tag{3.21}$$

One may also consider a class of estimators for S_y^2 as

$$t_{SS} = \left[w_1s_y^2 + w_2(S_x^2 - s_x^2)\right]\left(\frac{S_x^2}{s_x^2}\right). \tag{3.22}$$

The bias and MSE of the estimator t_{SS} to the first degree of approximation (ignoring fpc term) are respectively given by

$$B(t_{SS}) = S_y^2[w_1c_4 + w_2c_5 - 1], \tag{3.23}$$

$$MSE(t_{SS}) = S_y^4[1 + w_1^2c_1 + w_2^2c_2 + 2w_1w_2c_3 - 2w_1c_4 - 2w_2c_5], \tag{3.24}$$

where

$$c_1 = \left[1 + \frac{1}{n}(\beta_2(y) + 3\beta_2(x) - 4\gamma)\right], c_2 = \frac{1}{nR^2}(\beta_2(x) - 1),$$
$$c_3 = \frac{1}{nR}(2\beta_2(x) - \gamma + 1), c_4 = \left[1 + \frac{1}{n}(\beta_2(x) - \gamma)\right], c_5 = \frac{1}{nR}(\beta_2(x) - 1),$$

The $MSE(t_{SS})$ at (3.24) is minimum for

$$\left.\begin{aligned} w_1 &= \frac{(c_2c_4-c_3c_5)}{(c_1c_2-c_3^2)} = w_{10(2)}(say) \\ w_2 &= -\frac{(c_1c_5-c_3c_4)}{(c_1c_2-c_3^2)} = w_{20(2)}(say) \end{aligned}\right\}. \tag{3.25}$$

Thus, the resulting minimum MSE of t_{SS} is given by

$$MSE_{min}(t_{SS}) = S_y^4\left[1-\frac{(c_2c_4^2-2c_3c_4c_5+c_1c_5^2)}{(c_1c_2-c_3^2)}\right]. \tag{3.26}$$

In the following section we have made an effort to develop a new class of estimators for population variance S_y^2 using the knowledge of S_x^2 and its properties are studied. The proposed study is well supported through numerical illustration.

4. The Proposed Class of Estimators

We suggest the following class of estimators for S_y^2 of the study variable y as

$$T = w_1s_y^2\frac{S_x^2}{[S_x^2+\eta(s_x^2-S_x^2)]} + w_2s_y^2\, exp\left[\frac{\eta(S_x^2-s_x^2)}{2S_x^2+\eta(s_x^2-S_x^2)}\right]. \tag{4.1}$$

Expressing T in terms e's we have

$$T = S_y^2\left[w_1(1+e_0)(1+\eta e_1)^{-1} + w_2(1+e_0)\, exp\left[\frac{-\eta e_1}{2+\eta e_1}\right]\right]. \tag{4.2}$$

Expanding the right-hand side of (4.2), multiplying out and neglecting terms of e's having power greater than two, we have

$$\begin{aligned}(T-S_y^2) = S_y^2\Bigg[&w_1\{1+e_0-\eta e_1-\eta e_0e_1+\eta^2e_1^2\} \\ &+ w_2\left\{1+e_0-\frac{\eta e_1}{2}-\frac{\eta e_0e_1}{2}+\frac{3\eta^2e_1^2}{8}\right\}\Bigg]\end{aligned}$$

or

$$\begin{aligned}(T-S_y^2) = S_y^2\Big[&w_1(1+e_0-\eta e_1-\eta e_0e_1+\eta^2e_1^2) + w_2\Big(1+e_0-\tfrac{\eta e_1}{2}-\tfrac{\eta e_0e_1}{2}+ \\ &\tfrac{3\eta^2e_1^2}{8}\Big)-1\Big].\end{aligned} \tag{4.3}$$

Taking expectation of both sides of (4.3) we get the bias of T to the first degree of approximation as

$$B(T) = S_y^2[w_1\Sigma_4 + w_2\Sigma_5 - 1], \tag{4.4}$$

where

$$\Sigma_4 = \left[1+\frac{1}{n}\{\eta^2(\beta_2(x)-1)-\eta(\gamma-1)\}\right],$$
$$\Sigma_5 = \left[1+\frac{1}{n}\left\{\frac{3}{8}\eta^2(\beta_2(x)-1)-\frac{\eta}{2}(\gamma-1)\right\}\right].$$

Squaring both sides of (4.3) and neglecting terms of e's having power greater than two, we have

$$\left(T - S_y^2\right)^2 = S_y^4 \begin{bmatrix} 1 + w_1^2(1 + 2e_0 - 2\eta e_1 + e_0^2 - 4\eta e_0 e_1 + 3\eta^2 e_1^2) + w_2^2(1 + 2e_0 - \eta e_1 + e_0^2 - 2\eta e_0 e_1 + \eta^2 e_1^2) \\ +2w_1 w_2\left(1 + 2e_0 - \frac{3\eta e_1}{2} + e_0^2 - 3\eta e_0 e_1 + \frac{15\eta^2 e_1^2}{8}\right) - 2w_1(1 + e_0 - \eta e_1 - \eta e_0 e_1 + \eta^2 e_1^2) \\ -2w_2\left(1 + e_0 - \eta e_1 - \eta e_0 e_1 + \frac{\eta^2 e_1^2}{2}\right) \end{bmatrix}.$$

Taking the expectation of both sides of the above expressions we get the MSE of T to the first degree of approximation as

$$MSE(T) = S_y^4[1 + w_1^2\Sigma_1 + w_2^2\Sigma_2 + 2w_1 w_2\Sigma_3 - 2w_1\Sigma_4 - 2w_2\Sigma_5], \tag{4.5}$$

where

$$\Sigma_1 = \left[1 + \frac{1}{n}\{(\beta_2(y) - 1) - 4\eta(\gamma - 1) + 3\eta^2(\beta_2(x) - 1)\}\right],$$
$$\Sigma_2 = \left[1 + \frac{1}{n}\{(\beta_2(y) - 1) - 2\eta(\gamma - 1) + \eta^2(\beta_2(x) - 1)\}\right],$$
$$\Sigma_3 = \left[1 + \frac{1}{n}\left\{(\beta_2(y) - 1) - 3\eta(\gamma - 1) + \frac{15}{8}\eta^2(\beta_2(x) - 1)\right\}\right].$$

Differentiating (4.5) with respect w_1 and w_2 and equating them to zero, we have

$$\begin{bmatrix} \Sigma_1 \Sigma_3 \\ \Sigma_3 \Sigma_2 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} \Sigma_4 \\ \Sigma_5 \end{bmatrix}. \tag{4.6}$$

After simplification of (4.6) we get the optimum values of w_1 and w_2 as

$$\left.\begin{aligned} w_1 &= \frac{(\Sigma_2\Sigma_4 - \Sigma_3\Sigma_5)}{(\Sigma_1\Sigma_2 - \Sigma_3^2)} = w_{1(opt)} \\ w_2 &= \frac{(\Sigma_1\Sigma_5 - \Sigma_3\Sigma_4)}{(\Sigma_1\Sigma_2 - \Sigma_3^2)} = w_{2(opt)} \end{aligned}\right\}. \tag{4.7}$$

Substitution of (4.7) in (4.5) yields the minimum MSE of T as

$$MSE_{min}(T) = S_y^4\left[1 - \frac{(\Sigma_2\Sigma_4^2 - 2\Sigma_3\Sigma_4\Sigma_5 + \Sigma_1\Sigma_5^2)}{(\Sigma_1\Sigma_2 - \Sigma_3^2)}\right]. \tag{4.8}$$

which holds true if

$$0 < \frac{(\Sigma_2\Sigma_4^2 - 2\Sigma_3\Sigma_4\Sigma_5 + \Sigma_1\Sigma_5^2)}{(\Sigma_1\Sigma_2 - \Sigma_3^2)} < 1 \quad \text{and} \quad (\Sigma_1\Sigma_2 - \Sigma_3^2) > 0$$

5. Efficiency Comparison

From (3.2) and (3.7) we have

$$MSE\left(s_y^2\right) - MSE(t_{SPH}) = \frac{S_y^4\left(\beta_2(y - 1)\right)^2}{n(n + \beta_2(y) - 1)} \geq 0$$

which gives the inequality

$$MSE(t_{SPH}) \leq MSE\left(s_y^2\right). \tag{5.1}$$

This shows that the Singh et al (1973) estimator t_{SUN} is more efficient than usual unbiased estimator s_y^2.

From (3.2), (3.9) and (3.12) we have

$$MSE\left(s_y^2\right) - MSE_{min}\left(t_j\right) = \frac{S_y^4(\gamma-1)^2}{n(\beta_2(x)-1)} \geq 0, \tag{5.2}$$

$$MSE(t_R) - MSE_{min}\left(t_j\right) = \frac{S_y^4(\beta_2^*(x)-\gamma^*)^2}{n(\beta_2(x)-1)} \geq 0, \tag{5.3}$$

where

j= DT1, DT2, D; $\beta_2^*(x) = (\beta_2(x) - 1)$ and $\gamma^* = (\gamma - 1)$.

It follows from (5.2) and (5.3) that the estimators (t_{DT1}, t_{DT2}, t_D) are better than usual unbiased estimator s_y^2 and ratio estimator t_R due to Isaki (1983).

From (3.12) and (3.17) we have

$$MSE_{min}\left(t_j\right) - MSE_{min}(t_{SUN}) = \frac{S_y^4\left(a_1a_2-a_2-a_3^2\right)^2}{a_2(a_1a_2-a_3^2)} \geq 0 \tag{5.4}$$

j= DT1, DT2, D;

It follows from (5.4) that the estimator t_{SUN} is more efficient than $s_y^2, t_R, t_{DT1}, t_{DT2}$ and t_D.

From (3.17) and (4.8) we have that

$$MSE_{min}(T) < MSE_{min}(t_{SUN}) \text{ if}$$

$$\frac{\left(\Sigma_2\Sigma_4^2-2\Sigma_3\Sigma_4\Sigma_5+\Sigma_1\Sigma_5^2\right)}{\left(\Sigma_1\Sigma_2-\Sigma_3^2\right)} > \frac{a_2}{\left(a_1a_2-a_3^2\right)}. \tag{5.5}$$

From (3.21) and (4.8) it is observed that

$$MSE_{min}(T) < MSE_{min}(t_{GS}) \text{ if}$$

$$\frac{\left(\Sigma_2\Sigma_4^2-2\Sigma_3\Sigma_4\Sigma_5+\Sigma_1\Sigma_5^2\right)}{\left(\Sigma_1\Sigma_2-\Sigma_3^2\right)} > \frac{\left(b_2b_4^2-2b_3b_4b_5+b_1b_5^2\right)}{\left(b_1b_2-b_3^2\right)}. \tag{5.6}$$

From (3.26) and (4.8) we have that

$$MSE_{min}(T) < MSE_{min}(t_{SS}) \text{ if}$$

$$\frac{\left(\Sigma_2\Sigma_4^2-2\Sigma_3\Sigma_4\Sigma_5+\Sigma_1\Sigma_5^2\right)}{\left(\Sigma_1\Sigma_2-\Sigma_3^2\right)} > \frac{\left(c_2c_4^2-2c_3c_4c_5+c_1c_5^2\right)}{\left(c_1c_2-c_3^2\right)}. \tag{5.7}$$

Thus, the proposed class of estimators T is more efficient than the estimators t_{SUN},t_{GS} and t_{SS} as long as the conditions (5.5), (5.6) and (5.7) are satisfied respectively. Hence under the condition (5.5) the proposed class of estimators T is also more efficient than the estimators s_y^2, $t_R, t_{DT1}, t_{DT2}, t_D$ and t_{SPH}.

6. Empirical Study

To illustrate the performance of the suggested class of estimators T over the estimators $s_y^2, t_{SPH}, t_R, t_{DT1}, t_{DT2}, t_D, t_{SUN}, t_{GS}$ and t_{SS}, we consider two natural data sets earlier considered by Das and Tripathi (1982), Kadilar and Cingi (2007) and Singh and Solanki (2013).

Population-I The population consists of 353 villages /towns/ward under Panskura Police Station, (Source: Census 1961, West Bengal, District Census Hand Book, Mindnapore.) The characters y and x are number of persons and area of villages/towns/ ward in acres respectively.

For this population, the required parameters were obtained as follows:

$S_y^2 = 412624.88$, $S_x^2 = 40533.195$, $\gamma = 12.3063$,

$\beta_2(x) = 16.3895$, $\beta_2(y) = 15.05$, $N = 353$, $n = 30$.

Population-II The data sets earlier used by Kadilar and Cingi (2007) and Singh and Solanki (2013).

In this population data set the level of apple production amount (in 100 tones) is a study variable y and number of apple trees is an auxiliary variable x in 104 villages of the East Anatolia Region of Turkey in 1999. The required values of the parameter are:

$S_y^2 = 136.189$, $S_x^2 = 530202800.90$, $\gamma = 14.398$,

$\beta_2(y) = 16.523$, $\beta_2(x) = 17.516$, $N = 104$, $n = 20$.

We have computed the percent relative efficiencies (PRE's) of the estimators $s_y^2, t_{SPH}, t_R, t_{DT1}, t_{DT2}, t_D, t_{SUN}, t_{GS}$ and t_{SS} for both population data sets (I and II) and the resulting values are compiled in Tables 6.1 and 6.2 respectively.

Table 6.1: PRE's of different Estimators of Population Variance with respect to usual unbiased estimator s_y^2 for Population-I and Population-II

Estimator	Population-I	Population-II
	PRE $(.\,, s_y^2)$	PRE $(.\,, s_y^2)$
s_y^2 Usual unbiased estimator	100.00	100.00
t_R Isaki (1983) ratio estimator	205.80	296.07
t_{SPH} Singh, Pandey and Hirano (1973) estimator	146.83	177.62
(t_{DT1}, t_{DT2}, t_D) Das and Tripathi (1978) estimator	244.62	333.52
t_{SUN} Singh, Upaadhyaya and Namjoshi (1988) estimator	291.45	411.13
t_{GS} Gupta and Shabbir (2007) estimator	**342.77**	549.81
$\boldsymbol{t_{SS}}$	340.78	**779.07**

Table 6.2: PRE's of proposed class of estimators T with respect to usual unbiased estimator s_y^2 for populations I and II

Population I		Population II	
Values of constant η	PRE (T, s_y^2)	Values of constant η	PRE (T, s_y^2)
-5.25	1220.64	-5.25	817.99
-5.00	1289.65	-5.00	864.91
-4.75	1364.36	-4.75	915.85
-4.50	1445.38	-4.50	971.29
-4.25	1533.36	-4.25	1031.73
-4.00	1629.09	-4.00	1097.79
-3.75	1733.39	-3.75	1170.15
-3.50	1847.25	-3.50	1249.60
-3.25	1971.70	-3.25	1337.07
-3.00	2107.95	-3.00	1433.61
-2.75	2257.28	-2.75	1540.44
-2.50	2421.11	-2.50	1659.00
-2.25	2600.93	-2.25	1790.94
-2.00	2798.19	-2.00	1938.17
-1.75	3014.12	-1.75	2102.89
-1.50	3249.25	-1.50	2287.53
-1.25	3502.13	-1.25	2494.66
-1.00	3765.87	-1.00	2726.41
-0.75	4015.81	-0.75	2982.79
-0.50	4147.05	-0.50	3255.02
-0.25	3161.99	-0.25	3492.46
0.25	7368.44	0.25	8020.37
0.50	7891.91	0.50	6778.99
0.75	8839.35	0.75	7851.08
1.00	10358.80	1.00	10307.08
1.25	13373.12	1.25	19118.06
1.50	25878.68	**1.42**	**895292.40**
1.63	**401889.70**	1.90	860.29
2.00	1045.41	2.00	1942.39
2.25	4238.41	2.25	3689.14
2.50	5703.97	2.50	4797.45
2.75	6537.20	2.75	5576.82
3.00	7023.40	3.00	6123.91
3.25	7271.59	3.25	6473.92
3.50	7338.32	3.50	6645.46
3.75	7262.57	3.75	6657.09
4.00	7076.72	4.00	6532.59
4.25	6809.62	4.25	6300.44
4.50	6486.87	4.50	5990.92
4.75	6130.28	4.75	5632.61

Table 6.2: PRE's of proposed class of estimators T with respect to usual unbiased estimator s_y^2 for populations I and II (cont.)

Population I		Population II	
Values of constant η	PRE (T, s_y^2)	Values of constant η	PRE (T, s_y^2)
5.00	5757.64	5.00	5249.97
5.25	5382.63	5.25	4862.17
5.50	5015.27	5.50	4482.99
5.75	4662.38	5.75	4121.53
6.00	4328.29	6.00	3783.14
6.25	4015.37	6.25	3470.41
6.50	3724.59	6.50	3183.99
6.75	3455.92	6.75	2923.29
7.00	3208.69	7.00	2686.98
7.25	2981.82	7.25	2473.30
7.50	2774.01	7.50	2280.33
7.75	2583.84	7.75	2106.14
8.00	2409.89	8.00	1948.86
8.25	2250.75	8.25	1806.73
8.50	2105.11	8.50	1678.15
8.75	1971.72	8.75	1561.66
9.00	1849.44	9.00	1455.97
9.25	1737.22	9.25	1359.90

7. Simulation Study

To access the performance of the proposed class of estimators a simulation study is performed using R-software to verify the theoretical results. We have generated artificial population of two variables (y, x) based on regression model as x = rnorm $(N, 0, 1)$ and $y = x$ + rnorm $(N, 0, 1)$ of size N. We have generated two populations:

Population-I: $N = 5000$, $n = 2000$; Population-II: $N = 10000$, $n = 4000$.

Table 7.1: PRE's of different Estimators of Population Variance with respect to usual unbiased estimator s_y^2 for simulated Populations I and II

Estimator	Population-I	Population-II
	PRE $(. , s_y^2)$	PRE $(. , s_y^2)$
s_y^2 Usual unbiased estimator	100.00	100.00
t_R Isaki (1983) estimator	25.56	25.92
t_{SPH} Singh, Pandey and Hirano (1973) estimator	99.95	99.98
(t_{DT1}, t_{DT2}, t_D) Das and Tripathi (1978) estimator	1173.71	729.83
t_{SUN} Singh, Upaadhyaya and Namjoshi (1988) estimator	1173.65	729.81
t_{GS}Gupta and Shabbir (2007) estimator	1173.45	729.74
$\boldsymbol{t_{SS}}$	1173.66	729.81

Table 7.2: PRE's of proposed class of estimators T with respect to usual unbiased estimator s_y^2 for simulated populations I and II

Population I		Population II	
Values of constant η	PRE (T, s_y^2)	Values of constant η	PRE (T, s_y^2)
-13.00	480401.38	-13.00	1257339.40
-12.75	516601.08	-12.75	1325095.50
-12.50	555811.16	-12.50	1395274.40
-12.25	598235.74	-12.25	1467685.90
-12.00	644072.75	-12.00	1542089.00
-11.75	693506.78	-11.75	1618190.70
-11.50	746700.10	-11.50	1695645.50
-11.25	803781.66	-11.25	1774057.60
-9.75	1227383.64	-9.75	2237550.70
-9.50	1309460.73	-9.50	2308716.00
-9.25	1393511.34	-9.25	2376775.80
-9.00	1478746.04	-9.00	2441295.80
-8.75	1564256.06	-8.75	2501903.10
-8.50	1649039.04	-8.50	2558296.40
-8.25	1732035.48	-8.25	2610252.30
-8.00	1812174.46	-8.00	2657630.10
-7.75	1888425.13	-7.75	2700371.20
-7.50	1959849.53	-7.50	2738497.30
-7.25	2025651.43	-7.25	2772104.40
-7.00	2085216.23	-7.00	2801355.40
-6.75	2138137.75	-6.75	2826470.40
-6.50	2184229.77	-6.50	2847716.30
-6.25	2223521.84	-6.25	2865396.10
-6.00	2256241.20	-6.00	2879837.80
-5.75	2282783.69	-5.75	2891384.40
-5.50	2303677.91	-5.50	2900384.40
-5.25	2319546.34	-5.25	2907183.70
-5.00	2331067.19	-5.00	2912118.40
-4.75	2338939.59	-4.75	2915509.00
-4.50	2343853.86	-4.50	2917656.00
-4.25	2346467.62	-4.25	2918835.70
-4.00	2347387.74	-4.00	2919298.30
-3.75	2347157.61	-3.75	2919265.50
-3.50	2346248.83	-3.50	2918929.40
-3.25	2345056.27	-3.25	2918451.90
-3.00	2343895.73	-3.00	2917963.90
-2.75	2343002.95	-2.75	2917565.40
-2.50	2342533.61	-2.50	2917325.30
-2.25	2342563.39	-2.25	2917281.50
-2.00	2343088.09	-2.00	2917440.70

Table 7.2: PRE's of proposed class of estimators T with respect to usual unbiased estimator s_y^2 for simulated populations I and II (cont.)

Population I		Population II	
Values of constant η	PRE (T, s_y^2)	Values of constant η	PRE (T, s_y^2)
-1.75	2344023.33	-1.75	2917778.50
-1.50	2345203.99	-1.50	2918239.30
-1.25	2346383.36	-1.25	2918736.10
-1.00	2347233.10	-1.00	2919150.50
-0.75	2347344.01	-0.75	2919333.60
-0.50	2346226.93	-0.50	2919102.60
-0.25	2343320.54	-0.25	2918247.60
0.25	2329569.50	0.25	2913680.20
0.50	2317332.74	0.50	2909398.80
0.75	2300564.35	0.75	2903373.30
1.00	2278568.07	1.00	2895266.80
1.25	2250709.14	1.25	2884729.40
1.50	2216455.49	1.50	2871407.00
1.75	2175418.12	1.75	2854945.30
2.00	2127388.91	2.00	2835001.80
2.25	2072370.42	2.25	2811254.50
2.50	2010593.42	2.50	2783413.30
2.75	1942519.02	2.75	2751231.80
3.00	1868823.93	3.00	2714517.50
3.25	1790369.45	3.25	2673143.00
3.50	1708156.83	3.50	2627054.60
3.75	1623273.93	3.75	2576279.00
4.00	1536838.68	4.00	2520927.40
4.25	1449945.03	4.25	2461196.60
4.50	1363616.11	4.50	2397366.20
4.75	1278767.89	4.75	2329792.60
5.00	1196184.82	5.00	2258900.10
5.25	1116507.36	5.25	2185168.50
5.50	1040230.08	5.50	2109119.50
5.75	967708.11	5.75	2031301.30
6.00	899169.65	6.00	1952273.20
6.25	834732.14	6.25	1872590.10
6.50	774419.98	6.50	1792789.50
6.75	718182.47	6.75	1713378.80
7.00	665910.63	7.00	1634826.00
7.25	617452.29	7.25	1557552.20
7.50	572625.18	7.50	1481927.10
7.75	531227.79	7.75	1408266.40
8.00	493048.19	8.00	1336831.50
8.25	457870.94	8.25	1267830.80

Table 7.2: PRE's of proposed class of estimators T with respect to usual unbiased estimator s_y^2 for simulated populations I and II (cont.)

Population I		Population II	
Values of constant η	PRE (T, s_y^2)	Values of constant η	PRE (T, s_y^2)
8.50	425482.40	8.50	1201422.50
8.75	395674.52	8.75	1137718.70
9.00	368247.63	9.00	1076789.40
9.25	343012.21	9.25	1018667.60
9.50	319789.97	9.50	963354.50
9.75	298414.38	9.75	910823.80
10.00	278730.81	10.00	861027.00
10.25	260596.31	10.25	813897.40
10.50	243879.25	10.50	769353.80
10.75	228458.78	10.75	727304.40
11.00	214224.20	11.00	687649.70
11.25	201074.31	11.25	650284.90
11.50	188916.74	11.50	615102.20
11.75	177667.26	11.75	581992.80
12.00	167249.13	12.00	550848.10
12.25	157592.45	12.25	521561.10
12.50	148633.62	12.50	494027.30
12.75	140314.73	12.75	468145.30
13.00	132583.09	13.00	443817.30
13.25	125390.71	13.25	420949.70
13.50	118693.92	13.50	399453.10
13.75	112452.92	13.75	379242.50
14.00	106631.46	14.00	360237.50
14.25	101196.47	14.25	342362.00
14.50	96117.81	14.50	325544.20
14.75	91367.94	14.75	309716.70
15.00	86921.71	15.00	294816.20
15.25	82756.16	15.25	280783.20
15.50	78850.24	15.50	267562.00
15.75	75184.72	15.75	255100.70
16.00	71741.98	16.00	243350.40
16.25	68505.84	16.25	232265.90
16.50	65461.49	16.50	221804.50
16.75	62595.29	16.75	211926.70
17.00	59894.75	17.00	202595.50
17.25	57348.36	17.25	193776.40
17.50	54945.51	17.50	185437.20
17.75	52676.46	17.75	177548.00
18.00	50532.20	18.00	170080.90

8. Discussion

It is observed from Table 6.1 that in population-I, the estimator t_{GS} due to Gupta and Shabbir (2007) appears to be the best (in the sense of having least MSE) followed by the estimator t_{SS}while in population-II, the estimator $t_{SS}$is the best followed by the estimator t_{GS} due to Gupta and Shabbir (2007).

Comparing the results of Tables 6.1 and 6.2 it is observed that the PRE $\left(T, s_y^2\right)$ = **401889.70%** is the largest at η = **1.63**, which is very high as compared to the Gupta and Shabbir (2007) estimator t_{GS} $[\boldsymbol{PRE}\left(\boldsymbol{t_{GS}, s_y^2}\right) = \mathbf{342.77\%}]$ in population-I. The PRE $\left(T, s_y^2\right)$ is very high as compared to all estimators including t_{GS} in population-I for other values of constant η also. It is further observed from Table 6.2 that in population-II, the maximum PRE $\left(T, s_y^2\right)$ = **895292.40%** at η = **1.42**, which is very large as compared to the estimator t_{SS} $[PRE\left(t_{SS}, s_y^2\right) = 779.07\%]$. However, for other values of η in population-II, the PRE $\left(T, s_y^2\right)$ gives the larger values than the estimator t_{SS}. Thus, from Tables 6.1 and 6.2 it is observed that there is enough scope of selecting the values of η in obtaining estimators better than the estimators $s_y^2, t_{SPH}, t_R, t_{DT1}, t_{DT2}, t_D, t_{SUN}, t_{GS}$ and t_{SS}closed in Table 6.1. Finally, we conclude that the proposed class of estimators perform well as compared to the existing estimators discussed here. So, we recommend the proposed estimator T for its use in practice.

The results of simulation experiments which reveal the ascendance of PRE of the estimators

$s_y^2, t_{SPH}, t_R, t_{DT1}, t_{DT2}, t_D, t_{SUN}, t_{GS}$ and t_{SS} and the proposed class of estimators T with respect to conventional unbiased estimator s_y^2 are displayed in Tables 7.1 and 7.2 for various values of scalar 'η'.

Table 7.1 exhibits that the common PRE due to Das and Tripathi's (1978) estimators t_{DT1}, t_{DT2} and t_D is the largest among the estimators $\left(s_y^2, t_{SPH}, t_R, t_{SUN}, t_{GS,} t_{SS,} t_{DT1}, t_{DT2}, t_D\right)$ followed by the estimators t_{SS}. The PREs of the estimators $t_{DT1}, t_{DT2}, t_D, t_{SUN}, t_{GS}$ and t_{SS} with respect to s_y^2 are almost same in both the population I and II. It follows that the performance of the estimators $t_{DT1}, t_{DT2}, t_D, t_{SUN}, t_{GS}$ and t_{SS} are almost same. It is further observed that for both the artificial populations I and II, the performance of Isaki (1983) ratio-type estimator t_R and Singh, Pandey and Hirano (1973) estimator t_{SPH} even worse than the usual unbiased estimator s_y^2 (which does not utilize auxiliary information).

From the perusal of the simulated results summarized in Tables 7.1 and 7.2 for artificial populations I and II, it can be seen that the performance of the suggested class of estimators T is better than the usual unbiased estimator s_y^2, Isaki's (1983) ratio-type estimator t_R, Singh, Pandey and Hirano (1973) estimator t_{SPH}, Das and Tripathi's (1978) estimators (t_{DT1}, t_{DT2}, t_D)Singh, Upadhyaya and Namjoshi (1988) estimator t_{SUN}, Gupta and Shabbir's (2007) estimator t_{GS} and the estimator t_{SS} for various values of the scalar 'η'. Thus, the suggested class of estimators T is recommended for its use in practice based on the simulation study results too.

9. Conclusion

This article addresses the problem of estimating the population variance S_y^2 of a study variable y when information on population variance S_x^2 of the auxiliary variable x is available under simple random sampling without replacement (SRSWOR). We have suggested a class of estimators for population variance S_y^2 of the study variable y using information on population variance S_x^2 of the auxiliary variable x. We have obtained the bias and mean squared error of the suggested class of estimators up to first order of approximation. The optimum conditions are obtained under which the proposed class of estimators has least MSE. The merits of the suggested class of estimators are judged through two natural population data sets. It has been shown empirically that the suggested class of estimators is more efficient than the existing estimators considered here with substantial gain in efficiency. This fact can be seen from Tables 6.1 and 6.2. We have also carried out simulation study based on two artificial populations I and II. We have computed PRE's of different estimators of population variance S_y^2 relative to s_y^2 and the results are presented in Tables 7.1 and 7.2. Larger gain efficiency is observed by using the suggested class of estimators T over other existing estimators for a wide range of scalar "η". Finally, the results theoretically and empirically are very encouraging and useful to the researcher engaged in this area of interest. So, we recommend the proposed estimator for its use in practice.

Acknowledgement

Authors are thankful to the learned referees for their valuable suggestions regarding improvement of the paper.

References

Adhvaryu, D., Gupta, P. C., (1983). On some alternative sampling strategies using auxiliary information. *Metrika*, 30, pp. 217–226.

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

Cochran, W. G., (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *The Journal of Agricultural Science*, 30, pp. 262–275.

Das, A. K., (1982). *Contributions to the theory of sampling strategies based on auxiliary information*. Ph.D. Thesis, B. C. K. V., Mohanpur, Nadia, West Bengal, India.

Das, A. K., Tripathi, T. P., (1978). Use of auxiliary information in estimating the finite population variance. *Sankhya*, C(40), pp. 139–148.

Das, A. K., Tripathi, T. P., (1980). Sampling strategies for population mean when the coefficient of variation of an auxiliary character is known, *Sankhya*, C(42), pp. 76–86.

Gupta, P. C., (1978). On some quadratic and higher degree ratio and product estimator. *Journal of Indian Society of Agriculture Statistics*, 30, pp. 71–80.

Gupta, S., Shabbir, J., (2007). On the use of transformed auxiliary variables in estimating population mean. *Journal of Statistical Planning and Inference*, 137(5), pp. 1606–1611.

Isaki, C. T., (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78, pp. 117–123.

Kadilar, C., Cingi, H., (2007). Improvement in variance estimation in simple random sampling. *Communications in Statistics Theory Methods*, 36(11), pp. 2075–2081.

Pal S. K., Singh, H. P., (2018). Improved estimators of the finite universe variance and mean using auxiliary variable in sample surveys. *International Journal of Mathematics and Computation*, 29(2), pp. 58–70.

Ray, S. K., Singh, R. K., (1981). Difference cum ratio type estimators. *Jour. Ind. Sta. Assoc.*, 19, pp. 147–151.

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

Robson, D. S., (1957). Application of multivariate polykays to the theory of unbiased ratio-type estimation. *Journal of American Statistical Association*, 52, pp. 511–522.

Sahai, A., Sahai, A., (1985). On efficient use of auxiliary information. *Journal of Statistical Planning and Inference*, 12, pp. 203–212.

Searls, D. T., Intrapanich, R., (1990). A note on an estimator for variance that utilizes the kurtosis. *The American Statistician*, 44(4), pp. 295–296.

Singh J., Pandey, B. N. and Hirano, K., (1973). On the utilization of a known coefficient of kurtosis in estimation procedure of variance. *Annals of the Institute of Statistical Mathematics*, 25, pp. 51–55.

Singh, H. P., Ruiz Espejo, M., (2003). On Linear regression and ratio-product estimation of a finite population mean. *The Statistician*, 52, 1, pp. 59–67.

Singh, H. P., Solanki, R. S., (2013a). Improved estimation of finite population variance using auxiliary information. *Communications in Statistics Theory Methods*, 42, pp. 2718–2730.

Singh, H. P., Solanki, R. S., (2013b). A new procedure for variance estimation in simple random sampling using auxiliary information. *Statistical Papers*, 54, 2, pp. 479–497.

Singh, H. P., Tailor, R., (2005). Estimation of finite population mean using known correlation coefficient between auxiliary characters. *Statistica*, 65, 4, pp. 407–418.

Singh, H. P., (1987). Class of almost unbiased ratio and product type estimators for finite population mean applying quenocillles method. *Journal of Indian Society of Agriculture Statistics*, 39, pp. 280–288.

Singh, H. P., Upadhyaya, L. N. and Lachan, R., (1988). Estimation of finite population variance. *Current Science*, 57, 24, pp. 1331–1334.

Singh, H. P., (1986). A generalized class of estimators of ratio, product and mean using supplementary information on an auxiliary character in PPSWR sampling scheme. *Gujarat Statistical Review*, 13(2), pp. 1–30.

Singh, H. P., Nigam, P., (2020). A general class of dual to ratio estimators. *Pakistan Journal of Statisticas and Operation Research*,16, 3, pp. 421–431.

Singh, H. P., Upadhyaya, L. N., (1986). A dual to modified ratio estimators using coefficient of variation of auxiliary variable. *Proceedings of the National Academy of Sciences*, 56, A, pp. 336–340.

Singh, H. P., Yadav, A., (2020). A new exponential approach for reducing the mean squared errors of the estimators of population mean using conventional and non-conventional location parameters. *Journal of Modern Applied Statistical Methods*, 18(1), pp. 1–47.

Singh, H. P., Tailor, R. and Tailor, R., (2012). Estimation of finite population mean in two-phase sampling with known coefficient of variation of an auxiliary character. *Statistica*, 72(1), pp. 111–126.

Singh, J., Pandey, B. N. and Hirano, K., (1973). On the utilization of a known coefficient of kurtosis in the estimation procedure of variance. *Annals of the Institute of Statistical Mathematics*, 25, pp. 51–55.

Singh, M. P., (1965). On the estimation of ratio and product of the population parameters. *Sankhya*, B (27), pp. 321–328.

Singh, M. P., (1967). Ratio-cum-product method of estimation. *Metrika*, 12, 1, pp. 34–43.

Singh, R. K., Singh, G., (1984). A class of estimators for population variance using information on two auxiliary variates. *Aligarh Journal of Statistics*, 3&4, pp. 43–49.

Singh, S., (2003). Advanced sampling theory with application: How Micheal "selected" A my, *Kluwer Academic Publishers*.

Solanki, R. S., Singh, H. P., (2013). An improved class of estimators for population variance. *Model Assisted Statistics and Application*, 8, 3, pp. 229–238.

Srivastava, S. K., (1971). A generalized estimator for the mean of a finite population using multi-auxiliary information. *Journal of American Statistical Association*, 66, pp. 404–407.

Srivastava, S. K., (1980). A class of estimators using auxiliary information in sample surveys. *The Canadian Journal of Statistics*, 8, 2, pp. 253–254.

Srivastava, S. K., Jhajj, H. S., (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhya*, C(42), pp. 87–96.

Tracy, D. S., Singh, H. P. and Singh, R., (1996). An alternative to the ratio-cum-product estimator in sample surveys. *Journal of Statistical Planning and Inference*, 53, 3, pp. 375–387.

Vos, J. W. E., (1980). Mixing of direct, ratio and product methods estimators. *Statist. Netherlands Society for Statistics and Operations Research*, 34, 4, pp. 209–218.

Walsh, J. E., (1970). Generalisation of ratio estimate for population total. *Sankhya*, A(33), pp. 99–106.

Yadav, R., Upadhyaya, L. N., Singh, H. P. and Chatterjee, S., (2013). A generalized family of transformed ratio product estimator for variance in sample surveys. *Communications in Statistics Theory and Methods*, 42(10), pp. 1839–1850.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 99–117, https://doi.org/10.59139/stattrans-2024-005
Received – 28.02.2024; accepted – 02.08.2024

AMUSE: Analysis of mobility using simultaneous equations. Present population of refugees in Poland

Sebastian Wójcik[1]

Abstract

Due to the conflict in Ukraine, which escalated on 24^{th} February 2022, and caused a large inflow of Ukrainian citizens to Poland, a need to investigate this phenomenon by official statistics has arisen. When it comes to tracking the movement of refugees, statistical and administrative data sources fail due to the lack of timeliness or spatial granularity. Therefore, official statistics is reaching for big data sources which seem to be complementary to statistical and administrative data sources. In this paper, we deal with the synthetic Mobile Network Operator (MNO) daily data obtained from SIM cards issued to Ukrainian refugees by one of MNOs operating in Poland. We propose AMUSE, a workflow for data analysis, a model for the data deduplication and mobility estimation as well as a simple estimator of the present population. All these functions of AMUSE are based on the aggregated signaling data on time and territory.

Key words: mobility, Mobile Network Operator data, refugees, simultaneous equations, experimental statistics.

1. Introduction

The full-scale war taking place in Ukraine caused widespread damage of residential buildings, industrial facilities and critical infrastructure. Intense military operations led to mass migration of Ukrainian people, both inside the country and abroad. According to the UN Refugee Agency estimates, approximately 3.7 million people have been internally displaced and over 6.3 million emigrated abroad. Most people leaving Ukraine decided to cross the border into Poland. Admitting such a large number of refugees in a short time was a great challenge for the Polish authorities, non-governmental organizations and ordinary citizens. Some Ukrainians, after a short stay in Poland, went to other countries. However, many of them decided to stay in Poland for longer. This required creating a support system in the form of social benefits, health care and educational services (some recent reports concerning situation of Ukrainian can be found in References).

Large-scale inflow of refugees is a challenge for labor market, real estate market, education system, and health care system. For the process of planning humanitarian assistance and public services for refugees, information about the number of refugees being hosted becomes substantial. There are several approaches to estimating the number of refugees

[1]Institute of Mathematics, University of Rzeszów, Rzeszów, Poland. E-mail: swojcik@ur.edu.pl, & Statistical Office in Rzeszów, Poland. E-mail: s.wojcik@stat.gov.pl. ORCID: https://orcid.org/0000-0003-2425-9626.

including a sample survey, an administrative data source, Mobile Network Operator data, Payment Card Operator data, Social Media data, etc. Applied approach is heavily conditioned to the available data sources. The moment of crossing the border as well as applying for government services by a particular refugee should left a footprint in one or more administrative data sources. Thus, the administrative data sources, if available, seem to be a primary data source for estimating the number of refugees. It is not the case for many emergencies and crises across the globe, e.g. Afghanistan, Syria, and Sudan. It turns out that even in the case of countries with a well-developed system of registers, some other obstacles may occur in utilizing administrative data sources in estimating the population of refugees solely. Recently, Statistics Poland in close collaboration with the World Health Organization (WHO) carried out a sample survey of Ukrainian refugees in Poland (details can be found in the report *Health of refugees from Ukraine in Poland 2022. Household survey and behavioural insights research*). The survey was conducted by offices in Rzeszów and Lublin. The quantitative component was used to collect health information about the refugees while qualitative component was used to collect the experiences of refugees in accessing health services. To compute the estimates, the following administrative data sources were utilized: Border Guard data on evacuees of Ukrainian nationality at the Polish-Ukrainian border and population register (PESEL). It is worth noting that in the process of computing sample weights there was a substantial role of a pilot study conducted just after the war breakout which revealed that 54% of refugees decided to leave Poland and travel further to other countries. Since Poland is in the Schengen Zone, there is no regular border control within and so the movement of refugees from Poland to other countries in the Schengen Zone would be unregistered. Hence, without the pilot study, the number of refugees in Poland would be harshly overestimated.

The administrative data sources are often used to estimate the *usually resident population*, which is based on actual stay in the given area over a twelve months. Recently, some new concepts of population have been developed. One of them is the so-called *present population* also known as *de facto population* (Lanzieri (2013)). In opposition to the usually resident population, the present population is a snapshot, that is, it consists of all individuals present in a given area in the given moment of time. Some researchers (Lanzieri (2019)) advocate these alternative concepts of population as the complementary statistics providing information on population movements. The administrative data sources seem to be insufficient for the task of estimating the present population. Therefore, there is a need for an alternative data source. Letouzé and Jütting, representatives of the Data-Pop Alliance (a think-tank on Big Data and development) and Paris21 (The Partnership in Statistics for Development in the 21st Century), in their inaugural paper from 2015, argue that Big Data may provide faster and cheaper data with better granularity. Still, it shall be a complementary data source instead of being a replacement for standard surveys carried out by official statistics, including population statistics.

Mobile Network Operator data was utilized by official statistics, among others, in estimating the present population (Ahas et al. (2015), Deville et al. (2014)), mobility (Alexander et al. (2015), Diao et al. (2016)), and migration (Lai et al. (2019)). MNO data were also used to support policy against COVID-19 pandemic (Badr et al. (2019)). Nevertheless, there are research papers giving a warning of possible biases of population and migration

estimates based on MNO data (Wesołowski et al. (2013)) and e-mail data (Zagheni et al. (2012)).

Recently, Statistics Poland obtained Mobile Network Operator daily data pertaining to SIM cards issued to Ukrainian refugees by one of MNOs operating in Poland aggregated to LAU level. In this paper we deal with a synthetic version of data. The synthetic dataset preserves trends, seasonality and spatial dependency of the original dataset. The original dataset values were scaled and the noise was added. Hence, we propose a workflow for analyzing such type of a dataset in terms of seasonality and spatial dependency. In the main part of this paper, we propose AMUSE: a mobility model for the data deduplication and mobility estimation as well as a simple estimator of the present population.

2. Problem Statement

Statistical models of the present population are strongly embedded in the datasets obtained from MNO. These datasets include a cell plan and event data. The cell plan contains information about the geographical location of the Base Transceiver Stations (BTS) or alternatively *the cell towers* and their properties such as range, propagation direction, networks serviced, etc. The next figure presents the geographical location of the cell towers in the capital city (on the left) and in the small town, namely Ustrzyki Dolne (on the right).

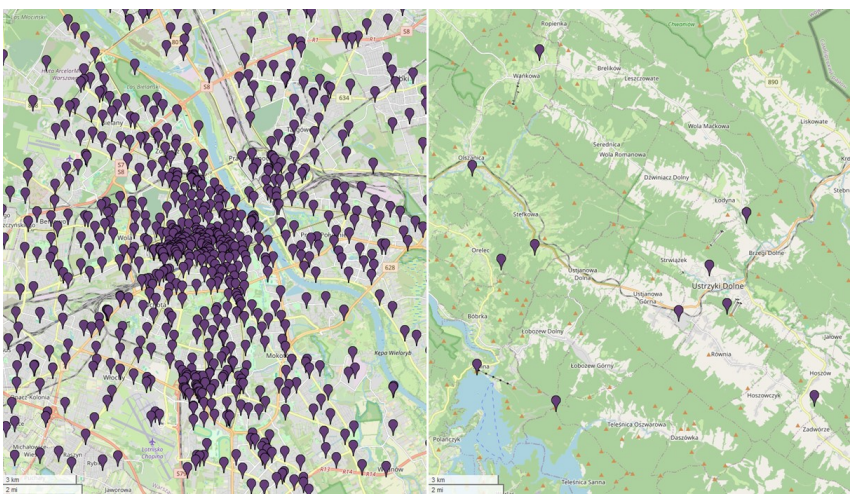


Figure 1: Cell plan

Density of BTS is really high in the capital city which has around 2 million population. On there other, in the small town with roughly 18 thousand inhabitants, there are only few BTS.

The event data contain information about the activity of particular cell phones. *Call Details Records* (CDR) consist of the records about calls (initiating and receiving) and SMSes (sending and receiving). Moreover, if CDR provide details on mobile data usage, then it is

often called *Data Details Records* (DDR) (Tennekes and Gootzen (2021)). Thus, CDR and DDR cover information about activity of the mobile phone users. MNO collects also *signaling data*, which are passive data consisting of logs to the cell towers. It is worth noting that CDR do not provide an information about the location of a particular mobile phone. The event record just states that in a given moment of time a particular mobile phone logged to a given BTS. Moreover, it is mostly a BTS with the strongest signal in a range of a particular mobile phone.

When dealing with MNO data, researchers and data analyst must keep in mind several issues concerning data quality aspects (Saidani, Bohnensteffen, and Hadam (2022)):

- Accuracy differences between MNOs - Variation in the events generated as a result of different practices between MNOs.
- Spatial accuracy - Short distances are underrepresented since location changes are only identified when a SIM card moves into a new mobile network cell. Moreover, smallest spatial unit varies greatly at the regional level due to different degrees of cell coverage, and accuracy of location estimation differs considerably.
- Asymmetric data losses - Units with little activity are disproportionately affected by anonymization losses.
- Validity - The assumption that a SIM card always communicates with the nearest antenna is not supported empirically.
- Biased features - Socio-demographic characteristics, e.g. age and gender are biased.
- Undercoverage - Not all sections of the population use a mobile device.

The event data may be provided by MNO in the form of micro data and aggregated data. In the case of micro data, one of the first steps includes modelling the spatial coverage patterns of BTS, that is mapping each cell tower to a geographical territory. This function is called *cell geolocation* (Salgado et al. (2020)). Ricciato et al. (2020) grouped existing geolocation methods into two families:

- Tessellations: after mapping geographical territories to cell towers, these areas remain disjoint. The simplest approach generates the simple fixed grid. A more data-driven approach includes Voronoi partitioning (Baccelli and Błaszczyszyn (2006)).
- Overlapping cells: mapping allows overlapping of geographical territories (Ricciato and Coluccia (2021)). Thus, it is a more general approach than tessellation.

The later step involves building a probabilistic model and estimating. In the recent literature several solutions can be found, namely:

- Maximum Likelihood Estimator based on multinomial distribution (Riccatio (2016)),
- Maximum Likelihood Estimator based on Poisson distribution (Shepp and Vardi (1982), Van der Laan, de Jongey (2019)),

- Simple Bayes-rule estimator (Tennekes and Gootzen (2021)).

In the case of aggregated data, the data can be obtained in various forms. In this paper we focus on the form of MNO data strictly connected with the use case of estimating the present population of Ukrainian refugees in Poland.

Several Mobile Network Operators issued SIM cards to Ukrainian citizens crossing the Polish-Ukrainian border after the war in Ukraine began. In order to register a SIM card in Poland for Ukrainian citizens, an identity document (ID card, passport or permanent residence card) is needed in the case of registration at registration points, or name, surname and PESEL (Universal Electronic System for Registration of the Population) in the case of registration through a bank. Due to the need to have an identity document or a PESEL number, the estimated results on the basis of Mobile Network Operator data should be comparable to the register of Ukrainian citizens who were assigned a PESEL number based on the Act published on March 12, 2022. Therefore, data obtained from a Mobile Network Operator generally does not include people who:

- came to Poland before the outbreak of the war,
- came to Poland after the outbreak of the war, but did not use a dedicated SIM card,
- registered a SIM card and then stopped using it (inactive card).

The obtained data are the daily counts of active SIM cards on Local Administrative Units level (LAU). We shall use also a term 'area' interchangeability with LAU. The MNO's system counts a given SIM card as active in a specific area provided the telephone with this card was active in that area for at least 3 hours a day. This means that if someone travelled between LAUs, SIM card could be counted multiple times in one day. Taking into account travel time between areas such a person could be assigned to a maximum of 7 neighboring or non-neighboring LAUs. Hence, we face a problem of multiple counts of particular SIM cards.

Let us note that the dataset does not contain any information on the directions of movements. Therefore, it is not possible to distinguish on its basis whether a given SIM card belonged to a resident of the given area or somebody who commuted to work, to school or to university as well as somebody who had a business or holiday trip. This distinction is essential to build a register of Ukrainian residents in Poland. Moreover, in the case of possibly longer business or holiday trip (at least three LAUs visited) we cannot distinguish which consecutive LAUs were visited within a trip. Deriving information how refugees commute and travel in Poland is also a subject of concern. Such information can be presented in the form of mobility matrix or series of mobility matrices. The idea of mobility matrix is incorporated in the proposed AMUSE model presented in details in the fourth section and it is a crucial tool in estimating the number of unique SIM cards.

After estimating the number of unique SIM cards from MNO, we need to estimate a total number of unique SIM cards from all MNOs operating in Poland. Finally, we shall derive a population of unique SIM card holders and then the whole population, that is, including persons who are not SIM card holders, especially children or some elderly persons. To this end, we start with data analysis for some insights.

3. Data analysis

The dataset contains information on the total daily number of active SIM cards issued by one of MNOs to Ukrainian citizens in the period from July to December, 2023 for 2,477 LAUs. No additional variables and no metadata are available. Since the data are already aggregated, there is no possibility to control quality issues on micro level. Further, we deal with issues of biased features and undercoverage.

The next plot presents the data aggregated to the national level. In the given period, the total number of active SIM cards increased almost by 10% (comparing the end points of the time interval). Strong deviation from the trend is observed, e.g. in September when the total number of active SIM cards decreased by 17% between 14 and 17 of September and then increased by 18% in a one day.

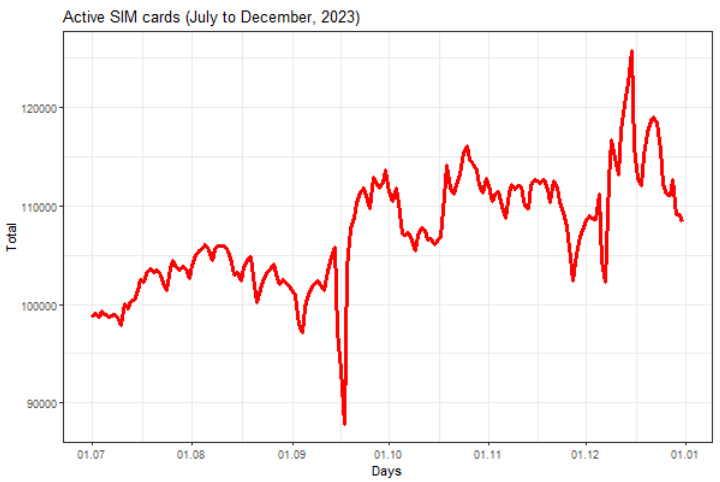


Figure 2: Active SIM cards time-series

The time-series analysis is ended with the seasonal decomposition. Seasonal decomposition by LOESS method proposed by Cleveland et al. (1990) was used. Figure 3 presents the results.

In the analyzed period, the rising trend is observed but with some turbulences. The seasonal part ranged from -1.6 thousand to 0.9 thousand while the remainder component ranged from -10.3 thousand to 6.3 thousand Weekly seasonality is not clearly visible. To investigate it, we used three seasonality tests, that is: Test for Seasonal Unit Roots proposed by Osborn et al. (1988), Test for Seasonal Stability by Canova, Hansen (1995) and Test for Seasonal Unit Roots presented in Hylleberg et al. (1990). Trigonometric version of Canova and Hansen test for seasonal stability indicated no seasonal cycles of seasonal frequencies $\frac{2\pi}{7}$, $\frac{4\pi}{7}$ and $\frac{6\pi}{7}$. Canova and Hansen test with dummy variables for seasonal stability revealed that on Friday there is significant and above-average traffic while on weekends the traffic is below the average. Osborn, Chui, Smith, Birchenhall test as well as Hylleberg, Engle, Granger, Yoo test indicated no significant seasonality.

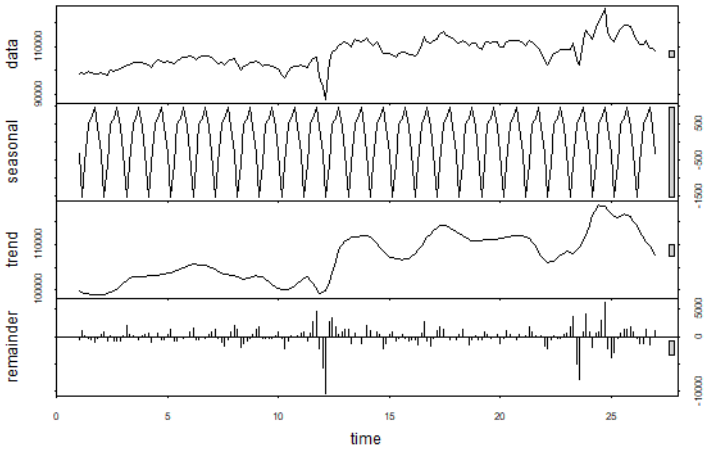


Figure 3: Seasonal decomposition of active SIM cards time-series

Let us investigate a spatial distribution of active SIM cards. The highest number of active SIM cards, for the majority of the analyzed period, was observed in the capital city of Warsaw. At the same time, there were 99 LAUs without a single active SIM card. The next figure presents the number of active SIM cards aggregated from LAU level to poviats (using former names, that is before 2017, aggregated from LAU level 2 to LAU level 1).

SIM cards mostly occur in urban areas, especially in big cities, or in tourist areas. One may ask if it follows a spatial distribution of population of the Polish citizens or the urban areas are in more favor of the refugees than rural areas. It may stem from the better access to accommodation, labor market, education and health services etc. To verify if this phenomenon occurs, let us recall first some studies on urban population. Auerbach (1913) observed a statistical regularity in the urban population of Germany. He discovered that the population of the city P is proportional to reciprocal of the population rank r in a decreasing order. Later, American linguist George Zipf (Zipf 1949) discovered that the similar regularity holds for rankings of words in text corpus for many natural languages. Zipf's law - applied to the population of the cities in a given country - states that the population of a given city follows the power law

$$P \sim \frac{1}{r^v} \tag{1}$$

where v is a fixed parameter ($v = 1$ for Auerbach discovery). Zipf also proposed a rank plot, also called Zipf plot, which is used to investigate if Zipf's law holds. The next figure presents the Zipf plots for MNO data and LAUs' population in Poland.

The solid lines represent fitted values from the model given by (1). In both models, R-squared reached 0.99. Note that the coefficient v amounted to 1.085 in the case of MNO data while it attained a value of 0.785 for Polish LAUs. Hence, SIM cards revealed that the Ukrainian refugees more often reside in highly urbanized areas, especially large cities. It could result from improved access to housing, labor market, education, health services, and so forth. Moreover, large cities are well connected in terms of rail and road transportation.

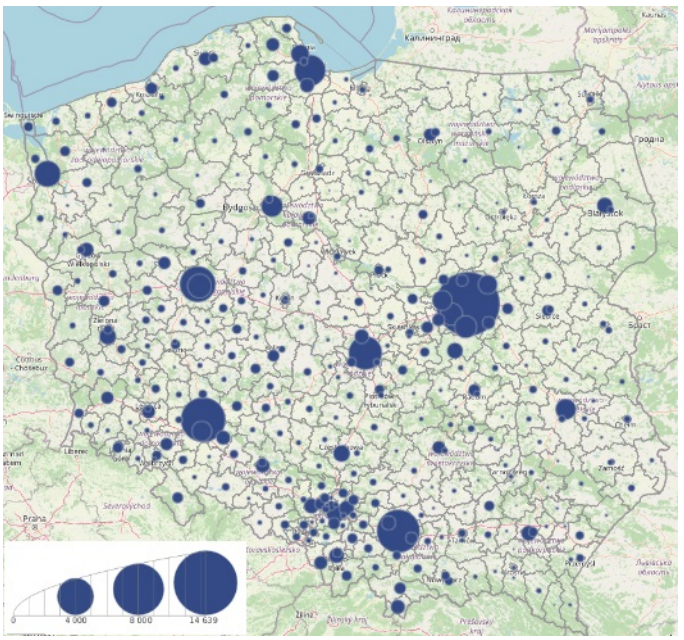


Figure 4: Spatial distribution of active SIM cards

4. AMUSE model

Due to the fact that a single SIM card could be counted several times, there was a need to build a model to estimate the number of unique SIM cards. Another important issue was to determine the mobility patterns of SIM card holders within the country. The following notations were used for modelling:

- $i \in \{1,...,n\}$ denotes index of spatial unit (area);
- $y = (y_1,\ldots,y_i,\ldots,y_n)$ denotes the vector of active SIM cards, $y_i > 0$ for $i \in \{1,...,n\}$;
- $x = (x_1,\ldots,x_i,\ldots,x_n)$ denotes the vector of unique SIM cards, $x_i \geq 0$ for $i \in \{1,...,n\}$;
- T - number of unique SIM cards in total.

In the model, y and T are known, while x is unknown. Determining x is the goal of the first stage of the estimation process. SIM card holders may exhibit many different behaviors in terms of inter-area movements in terms of the number of areas visited, their neighborhood, etc. Intuitively, the most cases will be:

- single counting of SIM card holders who live, work, or study within the same area;
- double counting of SIM card holders who live in one area and work or study in another area. At the same time, we can expect more cases of double counting with neighboring than with non-neighboring areas.

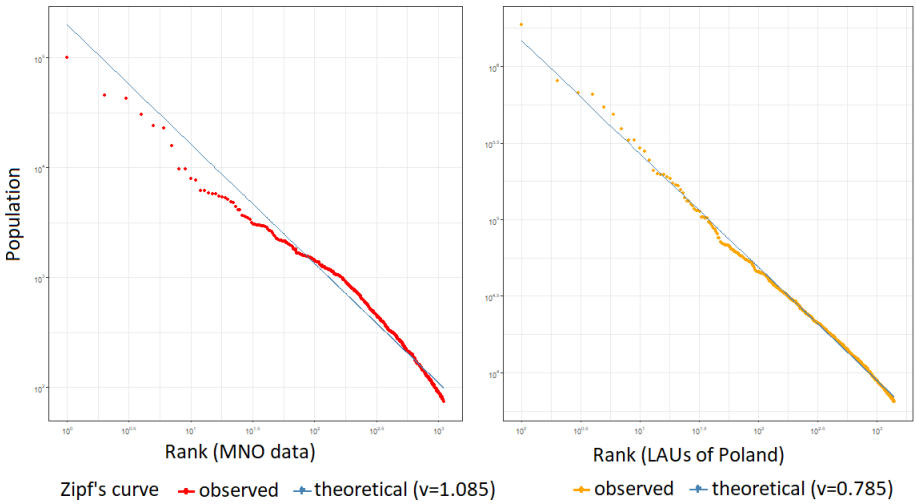


Figure 5: Zipf plots for MNO data and LAUs' population in Poland

In addition to commuting to the place of work or study, there are also movements related to trips for tourist and business purposes. These trips may involve more than two visited areas within a single trip. However, the number of such trips should be small when compared to all movement patterns.

4.1. Model setup

In order to build the model we start with the simplest case of a universe of two areas. Let us denote by p_{12} a share of SIM card holders residing in the area number 1 who were counted also in the area number 2 and by p_{21} a share of SIM card holders residing in the area number 2 who were counted also in the area number 1. Thus, by definition, $p_{12}, p_{21} \in [0,1]$, hence $x_i \leq y_i$ for $i \in \{1,2\}$. The next figure presents inter-area flows in the two-area universe.

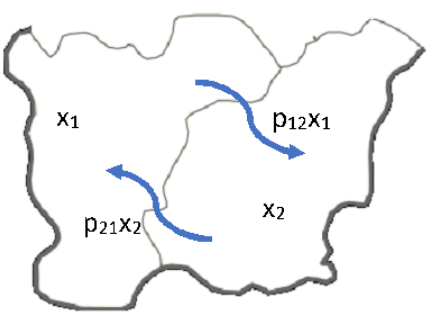


Figure 6: Flows in two-area universe

The number of active SIM cards y_1 in the area 1 is a sum of active SIM cards of residents x_1 and the number of active SIM cards of residents of the area 2, who visited the area 1 for

at least three hours, that is long enough to be counted by system of MNO as an active card. Therefore, the universe of two areas must meet the following equations:

$$\begin{cases} y_1 = x_1 + p_{21}x_2, \\ y_2 = x_2 + p_{12}x_1, \\ T = x_1 + x_2. \end{cases} \tag{2}$$

The system (2) of three equations contains four unknown variables. Hence, the system is undetermined.

In the case of a universe of, e.g. three areas, a particular SIM card can be counted even three times. For instance, resident of the area 2 could visit the area 3, and then could visit the area 1 or could visit the area 1 first, and then could visit the area 3. Hence, a more general model can be build two-fold:

- with dynamic approach,
- with static approach.

In dynamic approach we need to take into consideration a route followed by a particular SIM card holder. By route we understand a sequence of areas in which a SIM card holder were counted in a given day. Formally, let $s \in \{1,...,S\}$. Then, the S-step route is a finite sequence $(g_1,...,g_S) \in \{1,...,n\}^S$, that is a sequence of the area indices visited consecutively by a SIM card holder. When a resident of the area g_1 stayed within in a given day, then the route reduces to (g_1).

Further, we can define the flow frequencies between areas taking into account a stage of a route. Let $s \geq 2$ and $(g_1,...,g_s,...,g_S)$ be a route. By $p^{(s)}_{ij|g_1,\ldots,g_{(s-1)}}$ we denote a share of SIM card holders who travelled from the area number i to the area number j in a s-th step of the route residing in the area g_1 after visiting the areas $g_2,\ldots,g_{(s-1)}$ (note that the visit must take at least three hours in our setting). Formally, $p^{(s)}_{ij|g_1,\ldots,g_{(s-1)}} \in [0,1]$ and $p^{(s)}_{ij|g_1,\ldots,g_{(s-1)}} = 0$ for $i = j$.

Let us consider a case of three areas. The next figure presents possible routes of travelling to the area 1 (on the left, routes from the area 3; on the right, route from the area 2).

Therefore, the universe of three areas must meet the following equations:

$$\begin{cases} y_1 = x_1 + p_{21}x_2 + p_{31}x_3 + p_{23}p^{(2)}_{31|2}x_2 + p_{32}p^{(2)}_{21|3}x_3, \\ y_2 = x_2 + p_{12}x_1 + p_{32}x_3 + p_{13}p^{(2)}_{32|1}x_1 + p_{31}p^{(2)}_{12|3}x_3, \\ y_3 = x_3 + p_{13}x_1 + p_{23}x_2 + p_{12}p^{(2)}_{23|1}x_1 + p_{21}p^{(2)}_{13|2}x_2, \\ T = x_1 + x_2 + x_3. \end{cases} \tag{3}$$

Compared to (2), the system (3) of four equations contains 15 unknown variables. In a general case, the number of unknown variables increases proportionally to n^2.

The dynamic approach would be preferred to static approach whenever detailed information about mobility is of interest. When it comes to the estimation of the present population,

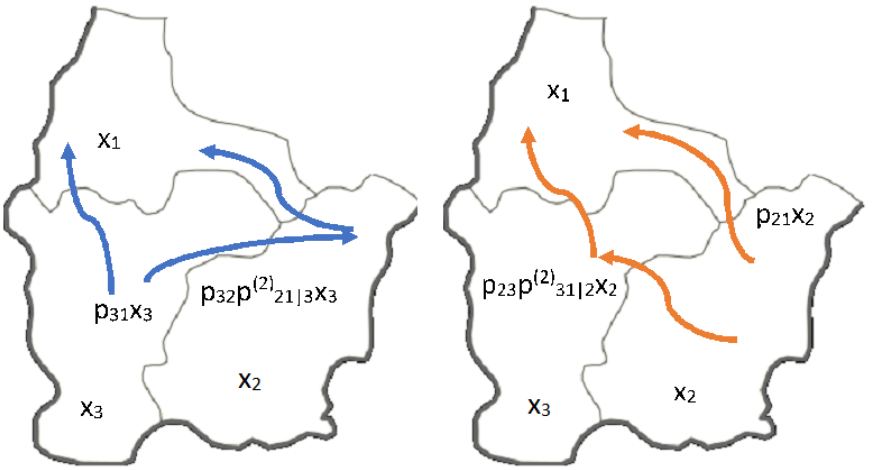


Figure 7: Flows in three-area universe

the static approach would be more straightforward and explainable. In the static approach we are only interested in the fact if a given SIM card holder visited a particular area. For instance, we are indifferent if a resident of i-th area visited j-th area directly or through the k-th area. That is, the routes (ij) and (ikj) are indistinguishable. In such a setting, the static model can be easily derived from the dynamic model. First note that we can rewrite (3) in the following way:

$$\begin{cases} y_1 = x_1 + \left(p_{21} + p_{23}p^{(2)}_{31|2}\right)x_2 + \left(p_{31} + p_{32}p^{(2)}_{21|3}\right)x_3, \\ y_2 = x_2 + \left(p_{12} + p_{13}p^{(2)}_{32|1}\right)x_1 + \left(p_{32} + p_{31}p^{(2)}_{12|3}\right)x_3, \\ y_3 = x_3 + \left(p_{13} + p_{12}p^{(2)}_{23|1}\right)x_1 + \left(p_{23} + p_{21}p^{(2)}_{13|2}\right)x_2, \\ T = x_1 + x_2 + x_3. \end{cases} \tag{4}$$

Putting $q_{ij} := p_{ji} + p_{jk}p^{(2)}_{ki|j}$ for $i, j, k = 1, ..., 3$, $i \neq j \neq k$, (4) takes the following form

$$\begin{cases} y_1 = x_1 + q_{21}x_2 + p_{31}x_3, \\ y_2 = x_2 + q_{12}x_1 + q_{32}x_3, \\ y_3 = x_3 + q_{13}x_1 + q_{23}x_2, \\ T = x_1 + x_2 + x_3. \end{cases} \tag{5}$$

Let Q' denote transposition of the matrix Q and assume that 1_n is a column vector of length n. In general, that is for an arbitrary $n \in \mathbb{N}$, the system of equations (5) can be presented in a matrix form

$$\begin{cases} y = Q'x, \\ T = 1'_n x. \end{cases} \tag{6}$$

The static approach has definitely less unknown parameters to estimate then the dynamic

approach. And so, the results are more data-driven than in the dynamic approach. On the other hand, it cannot be used to provide complex results on mobility. Nevertheless, it still captures a major part of movement, that is, commuting to the nearest area for the purpose of working or studying.

4.2. Priors

Since we are dealing with undetermined system of equations, the unknown x and Q can be estimated in an iterative way starting from a given prior x_0 and Q_0. The simplest form of the initial values (x_0, Q_0) may be based on a single point in time, that is, without concerning changes of y_i over time and any additional variables characterizing areas in terms of labor market, housing market, etc. Such a naïve prior may take a form

$$\begin{cases} \lambda = \frac{\sum_i^n y_i}{T}, & \\ x_i = \frac{y_i}{\lambda} & \text{for } i = 1,..,n, \\ q_{ij} = \frac{\lambda - 1}{n-1} & \text{for } i,j = 1,..,n,\ i \neq j. \end{cases} \tag{7}$$

Simple calculations show that the prior (7) satisfies $\sum_i^n x_i = T$. Moreover, $x_i \geq 0$ and $q_{ij} \in [0,1]$. Indeed, since $\sum_i^n y_i \geq T$, we have

$$0 = \frac{\frac{T}{T} - 1}{n-1} \leq \frac{\frac{\sum_i^n y_i}{T} - 1}{n-1} = q_{ij} \leq \frac{\frac{nT}{T} - 1}{n-1} = 1.$$

Further, (7) can be modified to take into account that q_{ij} is positive, e.g. only for neighbouring areas. Then, denoting by n_i the number neighboring areas of i-th area, we have

$$\begin{cases} \lambda = \frac{\sum_i^n y_i}{T}, & \\ x_i = \frac{y_i}{\lambda} & \text{for } i = 1,..,n, \\ q_{ij} = \frac{\lambda - 1}{n_i} & \text{for } i,j = 1,..,n,\ i \neq j. \end{cases} \tag{8}$$

The prior (8) is well-defined whenever $\sum_i^n x_i \leq 2T$ since that inequality ensures it holds $q_{ij} \in [0,1]$.

4.3. Estimation of the parameters

Now, we shall propose two methods for estimating the parameters of (6). The first one is based on the fixed-point iterations method. Among numerical methods for solving nonlinear equations or optimization problems, fixed-point iterations are widely used (cf. Shams et. al. (2022), Zhu et al. (2023)). These methods iteratively update an initial guess until a fixed point is reached, where the updated value equals the value from previous iteration. Their underlying idea share similarities with Banach's fixed-point theorem. Both involve finding points where certain conditions are satisfied, leading to estimation or convergence. The general idea is to define an operator $W : X \to X$ on a given space X that transforms and updates a prior value $z_0 \in X$, that is $W(z_0) = z_1$. Repeating that transformation consecutively

$W(z_t) = z_{t+1}$ should lead to the solution z^* satisfying

$$\lim_{t\to\infty} W(z_t) = z^*.$$

The solution is a fixed-point of the operator W, that is, it holds that $W(z^*) = z^*$. Existence of such fixed-point as well as convergence of the operator are conditioned to the properties of the operator W as well as the properties of the space X itself. Keeping that in mind we propose a fixed-point iterations method allowing updates of the values from the previous iteration only if the new values are in the domain of interest. For the learning rate $\alpha \in (0,1]$, the algorithm goes as follows:

(1) for $i = 1,...,n$ calculate

$$x_i^* = (1-\alpha)x_i + \alpha\left(y_i - \sum_{j,j\neq i}^{n} q_{ji}x_j\right)$$

and replace x_i by x_i^* in the $(i+1)$-th equation, provided $0 \leq x_i^* \leq y_i$. Elsewhere, skip the iteration.

(2) for $i, j = 1,...,n$, $i \neq j$ calculate

$$q_{ij}^* = (1-\alpha)q_{ij} + \alpha\left(\frac{y_j - \sum_{j,j\neq i}^{n} q_{ji}x_j}{x_i}\right)$$

and replace q_{ij} by q_{ij}^*, provided $0 \leq q_{ij}^* \leq 1$. Elsewhere, skip the iteration.

In this procedure, x_i^* is a convex combination of the current estimate x_i and the estimate $y_i - \sum_{j,j\neq i}^{n} q_{ji}x_j$ satisfying the i-th equation. The same idea holds for q_{ij}^*. The learning rate controls the speed of convergence to the set of values satisfying (6). For the prior which is not a data-driven, e.g. of the form (7), it is advised to use a small learning rate which shall produce estimates within a range of valid values more likely. On the other hand, keep in mind that the small learning rate will increase a computational burden.

The second method is based on the optimization approach. The optimization criterion can combine the relative squared change or relative squared error for x_i^* and Kullback–Leibler divergence for q_{ij}^*. Thus, the problem is to minimize

$$L(x^*, Q^*) = \sum_{i=1}^{n}\left(\frac{x_i^*}{x_i} - 1\right)^2 + B\sum_{i=1}^{n}\sum_{j,j\neq i}^{n} q_{ji}\log\left(\frac{q_{ji}}{q_{ji}^*}\right) \tag{9}$$

subject to condition (6) for x^*, Q^*. The parameter B in (9) serves to set a trade-off between emphasis on relative squared change and emphasis on Kullback–Leibler divergence. Moreover, for Kullback–Leibler divergence we adopt a standard convention that $0\log 0 = 0$. The optimization approach is harder to implement but it has sound statistical background.

5. Estimation of the size of the refugees' population

In this chapter, we present a simple approach to estimating the size of the refugees' population. Let us recall that after determining the AMUSE model developed in the previous step, the number of unique SIM cards for one analyzed MNO was derived. Hence, the next step is to compute the number of unique SIM cards for all MNOs operating in Poland. Next, having information about the total number of unique active SIM cards, we shall proceed to determine the size of the refugees' population.

The basis for computing the estimates were reports published by the Office of Electronic Communications. The reports contain information on, among other things, the number of users, mobile traffic, revenues, market shares, and types of services provided. Some data are presented in additional breakdowns, e.g. divided into SIM cards and M2M, pre-paid, post-paid, by operators, etc. The reports include data on several services including mobile telephony service, Internet access service, VoIP telephony service, landline telephony service, bundled services, and paid TV services. Various services have different level of market penetration. According to the state as of the end of 2022 in the telecommunication services market (based on the aforementioned report), there were 52.6 million SIM cards, 6.7 million M2M SIM cards, 17.91 million Internet users, 13.92 million subscribers to bundled services, and 10.83 million subscribers to paid TV services. Among the services provided by operators, mobile telephony services have by far the widest reach. For this reason, operator shares in mobile telephony services were taken into account in further calculations. Table 1 presents market shares in 2022.

Table 1: Market shares of MNOs in 2022

MNO	*market share*
P4	30.2
Orange	26.6
Polkomtel	20.4
T-Mobile	19.2
Others	3.6

Due to the very limited scope of available data on the mobile network market concerning citizens of Ukraine, in particular the absence of such data at the Office of Electronic Communications, it was necessary to make a series of assumptions:

(i) The structure of operators shares for all SIM cards in the market is similar to the structure of SIM cards issued to Ukrainian refugees.

(ii) The spatial distribution of SIM cards issued to citizens of Ukraine is similar for each MNO.

(iii) Movement patterns are similar for each MNO.

The first assumption could be partially verified based on online sources referring to the number of SIM cards issued to citizens of Ukraine. The information pertained (depending on the source) to only two or three operators. From the most recent data, which only covered

two operators, it was found that by March 16, 2022, they had issued 275,000 (57.9% of the total number of SIM cards issued by both operators combined) and 200,000 (42.1%) SIM cards, respectively. It turns out that the market shares of mobile telephony services of these two operators are very similar and remain in a proportion of 58.1% to 41.9%.

Let us denote the total number of SIM cards issued by k-th MNO to Ukrainian refugees and to Polish citizens by S_k^{UA} and S_k^{PL}, respectively. Taking into account the assumption (i), the estimator of the number of SIM cards issued to Ukrainian refugees by all Polish MNOs denoted by S^{UA} can be given by

$$S^{UA} = S_k^{UA} \cdot \frac{\sum_k S_k^{PL}}{S_k^{PL}}. \tag{10}$$

Observe that (10) can be interpreted as a direct estimator (Horvitz-Thompson estimator) of the total population provided a simple random sampling without replacement with inclusion probabilities equal to $p = \frac{S_k^{PL}}{\sum_k S_k^{PL}}$. Then, under the assumption (ii), the estimator of the number of SIM cards in i-th area S_i^{UA} can be computed in the following way:

$$S_i^{UA} = S^{UA} \cdot \frac{x_i}{\sum_i x_i}. \tag{11}$$

It should be borne in mind that, according to the law, each SIM card must be registered and assigned to a person or company. In particular, multiple SIM cards can be registered to one person. In recent years, their number has been around 50 million in Poland, which gives an average of over 1.32 SIM card per person. On the other hand, individuals aged 13 or older are eligible to register a card. Consequently, due to legal and other circumstances (e.g. the level of digital literacy among different age groups), not every person owns a mobile phone and a SIM card. According to a survey conducted by the Office of Electronic Communications, 78.0% of Polish people (at age 15 or more) own a smartphone.

By N, N_{SIM} and $N_{holders}$ let us denote the size of population, the total number of SIM cards and the total number of SIM cards holders, respectively. Then, note that we have the following equality:

$$N = \frac{N}{N_{holders}} \cdot \frac{N_{holders}}{N_{SIM}} \cdot N_{SIM}. \tag{12}$$

Moreover, $P := \frac{N_{holders}}{N}$ can be interpreted as a percentage of persons with SIM cards while $M := \frac{N_{SIM}}{N_{holders}}$ gives an average number of SIM cards per person. In a result, we obtain

$$N = \frac{N_{SIM}}{P \cdot M} \tag{13}$$

The equality (12) holds when all indicators pertain to the same market. While estimating the number of Ukrainian refugees on the basis of SIM cards, two statistics, that is *percentage of persons with SIM cards* and *average number of SIM cards per person*, are not known for this population and are replaced by the corresponding statistics from the Polish market.

In the case when age-gender characteristics of refugees differs from age-gender characteristics of the host country, the indicators *percentage of persons with SIM cards* and

average number of SIM cards per person can be harshly biased. If the age-gender cohort structure of refugees is available from, e.g. sample survey, then it can be used to weight the aforementioned indicators with respect to cohorts. It should be kept in view that age-gender characteristics of Ukrainian refugees in Poland can be investigated through, e.g. the register of Ukrainian residents under temporary protection, which was developed due to the act *Council Directive 2001/55/EC of 20 July 2001 on minimum standards for giving temporary protection in the event of a mass influx of displaced persons and on measures promoting a balance of efforts between Member States in receiving such persons and bearing the consequences thereof* and Polish act *Ustawa z dnia 12 marca 2022 r. o pomocy obywatelom Ukrainy w związku z konfliktem zbrojnym na terytorium tego państwa (Dz.U. z 2023r. poz. 103 z późn.zm.)*. Enormous disparities between age-gender cohorts of Polish citizens and Ukrainian refugees are presented in the next figure.

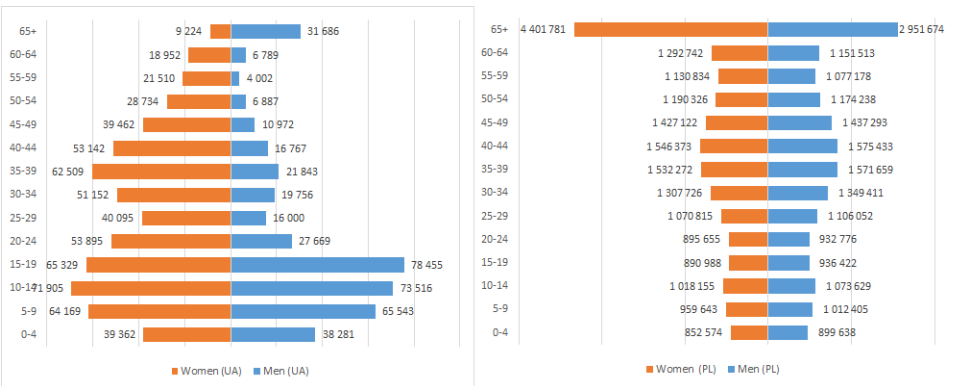


Figure 8: Age-gender cohorts of Polish citizens and Ukrainian refugees (as of May 31, 2023)

Size of cohorts was derived from the register of Ukrainian residents under temporary protection and Polish population register. Consider that the population of refugees mostly consists of children and young females. Adult (but not retired) males or elderly females are in the minority. Note that most of the children do not own smartphones formally. Since there are a lot of children in the population of Ukrainian refugees, the percentage of persons with SIM cards is lower than for the Ukrainian refugees in general.

Applying (10), (13), taking into account data from the Office of Electronic Communications and age-cohorts statistics from the register of Ukrainian residents under temporary protection, we obtained that for, e.g. MNO P4, to estimate the total number of Ukrainian refugees in Poland, the total number of unique SIM cards should be multiplied by 5.087362.

6. Conclusions

The article discusses the significant influx of Ukrainian refugees into Poland following the escalation of the conflict in Ukraine in February 2022. It highlights the challenges in tracking refugee movements using traditional statistical and administrative data sources

due to issues such as timeliness and spatial granularity. As a result, official statistics are turning to big data sources, such as mobile network operator (MNO) data, to supplement existing data. The paper focuses on utilizing synthetic MNO daily data from SIM cards issued to Ukrainian refugees by a Polish MNO. It proposes AMUSE model: mobility model for data deduplication and a simple estimator for estimating the present refugee population based on aggregated signalling data over time and areas. Further research shall be focused on including data variability over time into modelling.

References

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M. and Zook, M., (2015). Everyday space–time geographies: using mobile phone based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science*, 29(11), pp. 2017–2039.

Alexander, L., Jiang, S., Murga, M. and González, M. C., (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58, pp. 240–250.

Auerbach, F., (1913). Das gesetz der be volkerungskonzentration, Petermanns Geographische Mitteilungen, 59.

Baccelli, F., Błaszczyszyn, B., (2006). Tessellation in Wireless Communication Networks: Voronoi and Beyond it. Lorenz Center, Leiden University, 6-10 March 2006.

Badr, H., Du, H., Marshall, M., Dong, E., Squire, M. and Gardner, L., (2020). Association between mobility patterns and covid-19 transmission in the USA: a mathematical modelling study. The Lancet Infectious Diseases, 20(11).

Cleveland, R. B., Cleveland, W. S., McRae and J. E., Terpenning, I., (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6, pp. 3–73.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D. and Tatem, A. J., (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45), pp. 15888–15893.

Diao, M., Zhu, Y., Joseph Ferreira, J. and Ratti, C., (2016). Inferring individual daily activities from mobile phone traces: A Boston example. *Environment and Planning B: Planning and Design*, 43(5), pp. 920–940.

GUS, (2022). Health of refugees from Ukraine in Poland 2022, Household survey and behavioural insights research.

Lai, S., Erbach-Schoenberg, E., Pezzulo, C., Ruktanonchai, N., Sorichetta, A., Steele, J., Li, T., Dooley, C. and Tatem, A., (2019). Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications 5*, 34.

Lanzieri, G., (2013). Population definitions at the 2010 censuses round in the countries of the UNECE region, in: 15th Meeting of the UNECE Group of Experts on Population and Housing Censuses, Geneva, Switzerland.

Lanzieri, G., (2019). Towards a single population concept for international purposes: definitions and statistical architecture, in: 16th Meeting of the Task Force on the Future EU Censuses of Population and Housing, Luxembourg.

Osborn, D., Chui, A., Smith, J. and Birchenhall, C., (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics*, 50(4), pp. 361–377.

Ricciato, F., Coluccia, A., (2021). On the estimation of spatial density from mobile network operator data. arXiv:2009.05410v3 [eess.SP].

Ricciato, F., Lanzieri and G., Wirthmann, A., (2020). Towards a methodological framework for estimating present population density from mobile network operator data. Pervasive and Mobile Computing, 68.

Ricciato, F., Widhalm, P., Craglia and M., Pantisano, F., (2016). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. Pervasive and Mobile Computing.

Särndal, C-E., Swensson and B., Wretman, J., (1992). *Model Assisted Survey Sampling*, New York: Springer.

Saidani, Y., Bohnensteffen and S., Hadam, S., (2022). Quality Of Mobile Network Data – Project Experience And Use Cases In Official Statistics.

Salgado, D., Sanguiao, L., Onacea, B., et al. (2021). An end-to-end statistical process with mobile network data for official statistics. EPJ Data Sci. 10, 20(2021).

Shams, M., Kausar, N., Agarwal, P.and Oros, G. I., (2022). Efficient iterative scheme for solving non-linear equations with engineering applications. *Applied Mathematics in Science and Engineering*, 30:1, pp. 708–735, doi: 10.1080/27690911.2022.2130914.

Shepp, L. Vardi, Y., (1982). Maximum likelihood reconstruction for emission tomography. IEEE Transactions on Medical Imaging.

Tennekes, M., Gootzen, Y., (2021). A Bayesian approach to location estimation of mobile devices from mobile network operator data. *Journal of Spatial Information Science*.

UNHCR, (2023). Displacement Patterns, Protection Risks and Needs of Refugees from Ukraine, Regional Protection Analysis #3, Trends analysis: Moldova, Poland, Romania, and Slovakia, November 2023.

UNHCR, (2024). Ukraine Situation: Regional Refugee Response Plan, January-December 2024.

Urząd Komunikacji Elektronicznej, (2022). Raport o stanie rynku telekomunikacyjnego w Polsce w 2022 r.

Van der Laan, J., de Jongey, E., (2019). Maximum likelihood reconstruction of population densities from mobile signalling data. In NetMob'19.

Wesołowski, A. Eagle, N., Noor, A. M., Snow, R. W. and Buckee, C. O., (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society*, Interface, Vol. 10(81).

Zagheni, E., Ingmar, W., (2012). You are where you E-mail: Using E-mail Data to Estimate International Migration Rates, WebSci 2012.

Zhu, Z., Klein, A. B., Li, G.and Pang, S., (2023). Fixed-point iterative linear inverse solver with extended precision. Sci Rep, 13(1):5198. https://doi.org/10.1038/s41598-023-32338-5.

Zipf, G. K., (1949). Human behavior and the principle of least effort: Cambridge, Massachusetts: Addison-Wesley.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 119–136, https://doi.org/10.59139/stattrans-2024-006
Received – 28.08.2023; accepted – 09.10.2024

Forecasting under-five child mortality in Bangladesh: progress towards the SDGs target by 2030

Sacchidanand Majumder[1], **Soma Chowdhury Biswas**[2]

Abstract

In order to facilitate progress towards achieving the SDGs target regarding under-five child mortality in Bangladesh, the study explores the relevant mortality trends and makes a projection of the situation by 2030. The yearly dataset regarding mortality among children aged five and under (per 1,000 live births) in Bangladesh employed in this study was collected from the World Bank Databank (https://data.worldbank.org/indicator) for the 1972-2022 period. The selection of the best-fitted model for the purpose of forecasting was between the ARIMA model and the Double Exponential Smoothing Holt's Method. Compared with the ARIMA (1,2,1) model, the Double Exponential Smoothing Holt's method proved the best-fitted model for forecasting the under-five child mortality in the future. The results show that under-five child mortality in Bangladesh is an annually declining trend. The average under-five child mortality is forecasted to drop by one during the 2023-2035 period. Thus, the predicted value of under-five child mortality would be 26 in 2025 and 22 in 2030, which contributes to the achievement of the national target of 27 (per 1,000 live births) in 2025 and the SDGs target (25 deaths per 1,000 live births) regarding the under-five mortality rate. Bangladesh will then achieve the SDGs target regarding the under-five child mortality by 2026 if the existing strategy and plan of reducing under-five child mortality is successful.

Key words: under-five child mortality, forecasting, SDGs, Bangladesh.

1. Introduction

Under-five child mortality is a significant metric that serves as a measurement of child health and the overall progress of a nation. It provides insights into the socio-economic conditions in which children live, encompassing their access to healthcare (McGuire, 2006). The reduction of child mortality is a crucial target

[1] Department of Statistics, University of Chittagong, Bangladesh. E-mail: mithu.m2011@gmail.com. ORCID: https://orcid.org/0000-0002-9197-0381

[2] Department of Statistics, University of Chittagong, Bangladesh. E-mail: soma.c.biswas@gmail.com. ORCID: https://orcid.org/0009-0005-1034-1735.

within the framework of the Sustainable Development Goals (SDGs). The SDG 3 mentioned ensuring healthy lives and promoting well-being for all ages. The achievement of SDG 3 is contingent upon advancements made in many other SDGs, including No Poverty, Zero Hunger, Quality Education, Gender Equality, Clean Water and Sanitation, Affordable and Clean Energy, and Sustainable Cities and Communities. In order to attain these health goals, a series of targets have been established. Several of the targets outlined in the SDGs pertain to the domains of maternity, neonatal, and child health (Nino, 2015; WHO, 2016). The SDGs frameworks are referenced in target 3.2 within SDG 3, which aims to drop new-born and under-five mortality rates to 12 and 25 deaths per 1,000 live births correspondingly by 2030 (UN, 2015; Chao F., et al., 2018).

The advancements made in dropping child mortality globally have been noteworthy. Globally, the child mortality under five experienced a significant drop to 38 deaths per 1,000 live births in 2019, compared to 93 in 1990 and 76 in 2000, representing a decline of 59% and 50%, respectively, and the neonatal mortality fell to 17 deaths in 2019 from 37 deaths in 1990 and 30 in 2000, representing a decline of 52% and 42%, respectively (UNIGME, 2020). Globally, there was a substantial drop in the figure of deaths every day among under-five children, with a decline to 14,000 in 2019 from 27,000 in 2000 and 34,000 in 1990 (UNIGME, 2020).

Children persistently encounter significant regional inequalities in their prospects for survival. In the region of Sub-Saharan Africa, the under-five child mortality continues to be the highest (average 76 deaths per 1,000 live births in 2019) among all regions globally. This statistic indicates that the under-five child mortality is 1 in 13, 20 times higher than the Australia and New Zealand region (1 in 264) and equal to the world average in 1999 (UNIGME, 2020).

However, it is worth noting that by 2019, 122 nations had successfully attained an under-five mortality rate that fell below the specified target set by the SDGs, which aims to achieve 25 or less deaths per 1,000 live births. Those nations must strive towards sustaining advancements and actively mitigating inequalities within their people. To meet the SDGs target by 2030, 53 of the 73 remaining countries would need to step up their efforts. More nations are expected to fail to achieve the new-born mortality objective by 2030. Specifically, over 60 countries will be required to expedite their efforts to meet SDGs neonatal mortality target within the designated timeframe (UNIGME, 2020).

Bangladesh, situated in Asia, is categorized as a developing nation with a notable prevalence of under-five child mortality. However, it has substantially reduced this mortality rate, witnessing a decline from 133 deaths per 1000 live births in 1990 to 30.2 in 2018 (UN, 2015; Khan and Awan, 2017). This achievement can be attributed to the effective execution of the MDGs. In Bangladesh, the mortality rate among children under five decreased significantly to 28 deaths per 1,000 live births in 2019 from 125 deaths in 1995, 84 deaths in 2000, and 46 deaths in 2014 (GoB, 2020a;

MICS, 2019; BDHS 2014). Nevertheless, the prevalence remains elevated within South Asian nations, with Bangladesh ranking third after Pakistan (69.3) and India (36.6) (UNICEF, 2018; WHO, 2018). It is indisputable that a substantial amount of work and ongoing endeavors are crucial in order to secure additional declines in neonatal mortality and mortality rates among children under five, with the ultimate aim of attaining the corresponding SDG targets. Bangladesh has made notable progress in effectively decreasing child mortality over the past few decades (Ahmed et al., 2012). This achievement has played a crucial role in attaining the MDG 4 target. To make strides towards attaining the SDGs target, it is imperative to decrease the death rate among children under five. It is imperative to examine the trend of under-five child mortality to mitigate the susceptibility of child survival. This study aims to examine the patterns of mortality among under-five children and estimate projections for the year 2030 to advance the attainment of the SDGs target pertaining to under-five child mortality in Bangladesh.

2. Methodologies

2.1. Source of Data

The present study mainly employs secondary data sources. The annual dataset about the mortality rate of children under five (per 1,000 live births) in Bangladesh, utilized in this study, is collected from the World Bank Databank. The dataset covers the period from 1972 to 2022 and is categorized under the indicator "Mortality rate, under-5 (per 1,000 live births)." The dataset can be found at the following URL: https://data.worldbank.org/indicator. The dataset utilized in this study is displayed in Table 1 below:

Table 1: The yearly under-five child mortality rate (per 1,000 live births) in Bangladesh

Year	Number	Year	Number	Year	Number	Year	Number
1972	228	1985	178	1998	97	2011	47
1973	227	1986	171	1999	91	2012	45
1974	226	1987	165	2000	86	2013	42
1975	224	1988	159	2001	81	2014	40
1976	222	1989	152	2002	76	2015	39
1977	218	1990	146	2003	72	2016	37
1978	215	1991	140	2004	68	2017	35
1979	210	1992	133	2005	65	2018	34
1980	206	1993	127	2006	61	2019	32
1981	201	1994	121	2007	58	2020	31
1982	195	1995	115	2008	55	2021	30
1983	190	1996	109	2009	52	2022	29
1984	184	1997	103	2010	49		

Source: World Bank (Access on 2nd July 2024).

2.2. Methods of Analysis

Various statistical approaches are available for analyzing trends, ranging from basic linear regression to advanced parametric and non-parametric approaches (Hesel and Hirsch, 1992; Chen et al., 2007). Nevertheless, it is imperative to select a suitable methodology that guarantees the precise identification of all events pertaining to variability and change, as well as their subsequent influence on future predictions (Machiwal and Jha, 2008). The Mann-Kendall test, first proposed by Mann in 1945 (Mann, 1945) and further developed by Kendall in 1975 (Kendall 1975), was employed in this study to conduct trend analysis and ascertain the slope of the trend line. The time series modeling and forecasting task involved the utilization of the Non-seasonal Auto-Regressive Integrated Moving Average (ARIMA) model, as suggested by Box and Jenkins in 1976 (Box and Jenkins, 1976).

Double exponential smoothing Holt's method, often called Holt's method, was developed by Holt in 1957 to make predictions closer to actual values while displaying a trend in the data series (Holt, 1957). The selection of the most suitable forecasting model was based on evaluating error measures such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The chosen model would be utilized to forecast the under-five child mortality (per 1,000 live births) in Bangladesh for 2023–2035.

2.2.1. Mann-Kendall Test

The Mann-Kendall test is a statistical procedure commonly applied to analyses trends in time series data. It is a non-parametric method, meaning that it does not rely on any specific assumptions about the underlying distribution of the data (Kendall, 1975). One significant benefit of the Mann-Kendall test is its independence from the statistical distributions necessary for parametric methods. The null hypothesis (H_0) in the context of the Mann-Kendall test posits an absence of trend or serial correlation within the population under analysis. Conversely, the alternative hypothesis (H_1) implies the presence of a monotonic trend, either growing or decreasing.

2.2.2. Auto-Regressive Integrated Moving Average (ARIMA) Model

The Box-Jenkins procedure is often regarded as the most effective approach for model selection, as it leverages real-world datasets to determine the most suitable model. This methodology has several advantages, including calculating a reduced number of parameters and assessing the presence of seasonality in the data. The model construction process primarily comprises of four key steps: identification, estimation, diagnostic verification, and forecasting (Box and Pierce, 1970). The ARIMA model is a widely used time series. The parameters (p, d, q) are commonly used in time series

analysis. In this context, "p" represents the autoregressive coefficient, "d" denotes the order of differencing required to achieve stationarity in the time series, and "q" signifies the number of moving average components involved in the analysis. Conduct an initial stationarity test on the time series, and if it fails to meet the stationarity assumption, employ techniques such as differencing and logarithmic transformations to induce stationarity. The autocorrelation function (ACF) exhibits a decay for several lags, indicating the absence of significant correlation. Conversely, spikes in the plot suggest the presence of q parameters. On the other hand, the partial autocorrelation function (PACF) diminishes for numerous lags, indicating the lack of substantial correlation. However, spikes in the plot suggest the presence of p parameters.

In the event that multiple models are constructed, the Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC) are employed to select the most suitable model, as determined by the model with the lowest value of AIC and BIC. In addition to diagnostic checks, a test is employed to assess the model's appropriateness in relation to random errors by utilizing the Ljung-Box Q Statistics. The Q statistic is employed to assess the degree of randomness exhibited by the error term and to determine the statistical significance of the estimated parameters. In the event that the parameters lack significance or exhibit non-random errors, it is advisable to explore other p and q ordering until statistically significant p and q values are obtained. In this study, many models were employed to analyze the train dataset and determine the optimal model based on multiple criteria for predicting the yearly under-five child mortality rates in Bangladesh. The non-seasonal ARIMA model is

$$y_t = \mu + \emptyset_1 y_{t-1} + \emptyset_2 y_{t-2} + \cdots .. + \emptyset_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \theta_q \varepsilon_{t-q} - \epsilon_t;\ \epsilon_t \sim iid$$

where $\emptyset$, θ are unknown parameters.

2.2.3. Double Exponential Smoothing Holt's Method

Double exponential smoothing, often called Holt's method, was developed by Holt in 1957 to make predictions closer to actual values while displaying a trend in the data series (Holt, 1957). Holt's method is used when predicting data with a linear trend. Due to the data trend, the simple smoothing model will frequently make significant errors that swing from positive to negative or vice versa. The model includes a forecast equation as well as two smoothing equations (Wilson and Keating 2007). Holt's method smooths the slope and trend of the time series using two separate smoothing constants (α for the level and γ for the trend).

The equation for forecast, $y_{t+h} = l_t + hb_t$

The equation for the level, $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$

The equation for a trend, $b_t = \gamma(l_t - l_{t-1}) + (1-\gamma)b_{t-1}$

where y_{t+h} is the prediction for h periods into the future, l_t is an estimate of the level at the time t, b_t is an estimate of trend (slope) at the time t, h are future forecasting periods, periods to be forecast into the future, α is the level's smoothing constant ($0 \le \alpha \le 1$), and γ is the trend's smoothing constant ($0 \le \gamma \le 1$).

2.2.4. Measures of Forecasting Accuracy

When faced with the task of choosing among several types of forecasting methodologies, it is of utmost importance to consider the accuracy of the forecasts. In this context, the term "accuracy" pertains to the differentiation between the observed and projected value of a specific timeframe, also known as the forecasting error. The three forecasting error factors applied in this study are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The measure commonly used to quantify the average discrepancy between the observed and predicted values for a given time frame is denoted as MAE. At the same time, the RMSE represents the root of the average of the squared differences between the observed and predicted values. On the other hand, the MAPE is used to represent the average of absolute percentage mistakes. The equations utilized for calculating the aforementioned errors are as follows:

$$\mathrm{MAE} = \frac{\sum|\mathrm{D_t} - \mathrm{F_t}|}{\mathrm{n}}$$

$$MAPE = \frac{\sum|e_t/D_t|}{n} \times 100$$

$$RMSE = \sqrt{\frac{\sum(D_t - F_t)^2}{n-1}}$$

where D_t refers to the observed demand during period t, while F_t represents the estimated demand during the same period. The variable n represents the designated number of periods, and e_t denotes the discrepancy between the observed demand (D_t) and the estimated demand (F_t), which is calculated as the difference between the two (D_t -F_t).

3. Results and Discussion

All the analysis is done by SPSS (V23.0). The analysis of continuous data on the under-five child mortality rate (per 1,000 live births) is of significant importance in examining the patterns of variation in under-five child mortality. The Mann-Kendall test is utilized to determine the trends in time series data of under-five child mortality

rate. The utilization of ARIMA model and Double Exponential Smoothing Holt's Method for time series data analysis yields results that are subsequently employed for selecting the best-fitted model for the purpose of forecasting. The chosen model would then be utilized to forecast the under-five child mortality rate (per 1,000 live births) in Bangladesh for the period 2023–2035.

3.1. Mann-Kendall Test

In order to identify the seasonal trend in time series data pertaining to the under-five child mortality rate, the Mann-Kendall statistics (Zc statistics) is performed. The hypotheses to be examined under Kendall's Tau test for trend were as follows:
H_0: There is no trend in the under-five child mortality rate (per 1,000 live births).
H_1: There is a trend in the under-five mortality rate (per 1,000 live births).

Table 2: The Results of Mann-Kendall trend test (two-tailed test).

Statistic	Estimate
Kendall's Tau Coefficient	-.1.000**
p-value (two-tailed)	.000
N	51

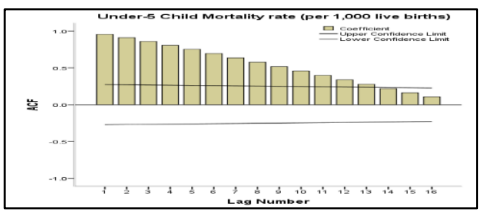
The calculated p-value is found to be below the predetermined significance level of α=0.05. As a result, the null hypothesis H_0, which states that there is no observed trend in the mortality rate of children under five in Bangladesh, is rejected. Conversely, the alternative hypothesis H_1, which suggests the presence of a trend in the aforementioned mortality rate, is accepted.

3.2. Model Selection and Forecasting of Under-Five Child Mortality Rate

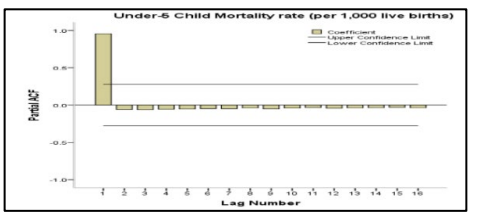
3.2.1. ARIMA Model

Figure 1 displays the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots regarding the under-five mortality rate in Bangladesh. The plots visually represent the confidence limits by the continuous line positioned above and below the X-axis. Figure 1(a) presents the ACF plot without applying differencing. It is observed that first thirteen spikes at lag 1-13 are above the confidence limit and the last three spikes at lag 14-16 are within the confidence limit. Figure 1(a) demonstrates that the ACF steadily decreases to zero as time delays t rise, suggesting a geometric decay representative of an AR process. Once more, Figure 1(b) illustrates the PACF plot without differencing. It indicates a single spike over the confidence limit at lag 1, and all other lags are within the significant level. Additionally, Figure 1(c) illustrates the

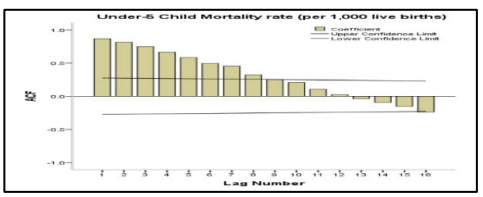
autocorrelation function (ACF) plot after applying differencing. It is observed that first eight spikes at lag 1–8 are above the confidence limit and the last one spike at lag 16 which falls below the confidence level. Moreover, Figure 1(d) illustrates the plot of the partial autocorrelation function (PACF) after applying differencing. The plot indicates two spikes, with the first one above the confidence limit at lag 1 and the second one below the confidence limit at lag 8. All other lags are within significant level.



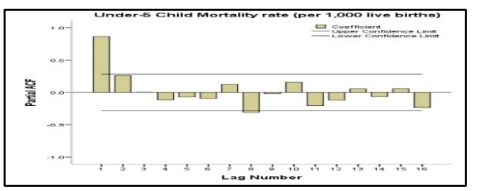
(a). Without differencing



(b). Without differencing



(c). With differencing

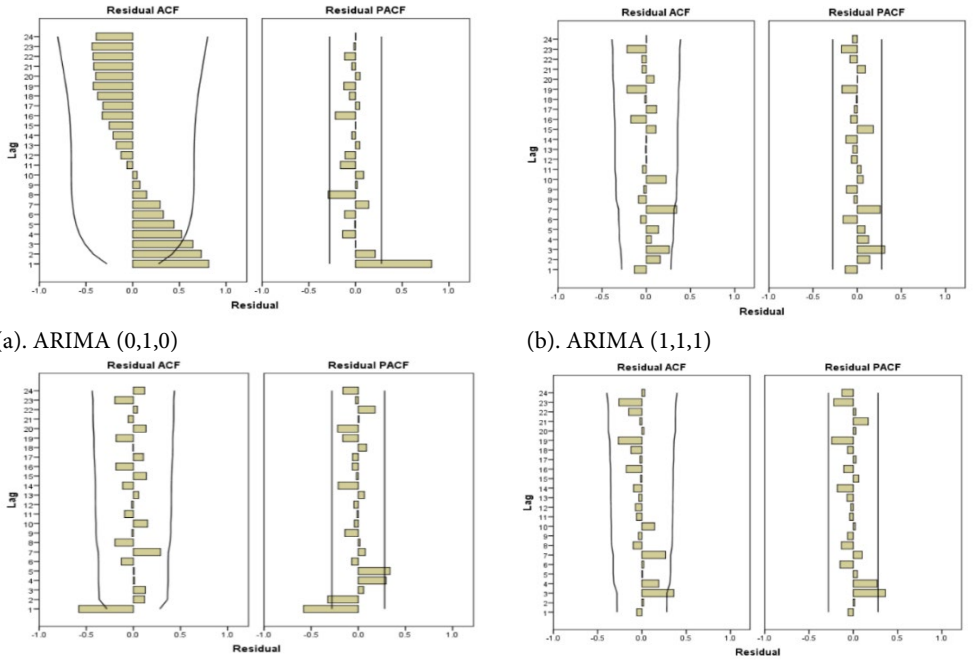


(d). With differencing

Figure 1: Autocorrelation and partial autocorrelation of under-five child mortality rate in Bangladesh

After detecting the theoretical attributes of the ACF and PACF as an AR process, it is evident that the ACF of the differenced series exhibits a progressive decline towards zero due to the AR component's influence. This decline in the ACF effectively mitigates the impact of the moving average (MA) component, hence reducing the presence of any underlying trend. The ARIMA model can be specified as ARIMA (0, 1, 0), ARIMA (1, 1, 1), ARIMA (0, 2, 0), or ARIMA (1, 2, 1), depending on whether the decision is made to incorporate a lagged error term.

In order to investigate whether the four selected models present any apparent systematic patterns that may be eliminated to enhance the predictability of these models, we conducted a further analysis of the ACF and partial ACF of the residuals. Figure 2 displays the ACF and Partial ACF of the residuals obtained from the models used to analyze the under-five mortality rate. Figure 2 demonstrates that ACF and partial ACF of the residuals fall within the confidence interval and do not exhibit substantial deviation from zero. This finding suggests that the models have been chosen properly.



(a). ARIMA (0,1,0)

(b). ARIMA (1,1,1)

(c). ARIMA (0,2,0)

(d). ARIMA (1,2,1)

Figure 2: Autocorrelation (ACF) and partial autocorrelation (PACF) of residuals for four selected models for the under-five child mortality rate

Table 3 displays the proposed models determined by applying the principle of parsimony and the recognizable pattern of the ACF and PACF. The optimal estimate to the data within the ARIMA family is determined by evaluating the Akaike's Information Criterion (AIC), Schwarz's Bayesian Criterion (BIC) values and Ljung-Box Q Statistics.

Table 3: ARIMA models selection for diagnostic analysis for the under-five child mortality rate in Bangladesh

Model	AIC	BIC	Ljung-Box Q		
			Statistics	DF	Sig.
ARIMA (0,1,0)	210.911	212.823	166.651	18	.000
ARIMA (1,1,1)	121.732	127.468	23.006	16	.114
ARIMA (0,2,0)	123.159	125.051	35.454	18	.008
ARIMA (1,2,1)	115.141	120.816	20.632	16	.193

Compared to the values of AIC, the BIC and Ljung-Box Q Statistics for the possible models displays that ARIMA (1,2,1) has the lowest values of AIC, BIC and Ljung-Box Q Statistics. Hence, ARIMA (1,2,1) is selected as the best-fitted ARIMA model for the data. Figure 3 displays the result of the fitted model of the under-five child mortality rate in Bangladesh using the ARIMA (1,2,1) model.

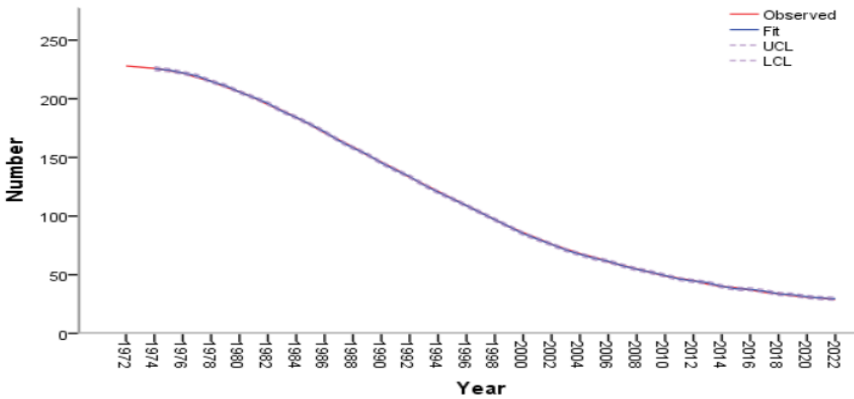


Figure 3: The Fitted Model of the Under-five Child Mortality Rate in Bangladesh using ARIMA (1,2,1) Model

3.2.2. Double Exponential Smoothing Holt's method

The yearly data for the period 1972-2022 are displayed in Table 1 and it is observed that this data pattern is a linearly decreasing trend. So, the Double Exponential Smoothing method is applied in this study. The step in using Double Exponential Smoothing Holt's method for forecasting is to plot the data to determine the data pattern. The goal of this step is to find out whether there is a trend, cycle, seasonal, random, or stationary component. Figure 4 shows the plotted results of the data series pattern from the actual data.

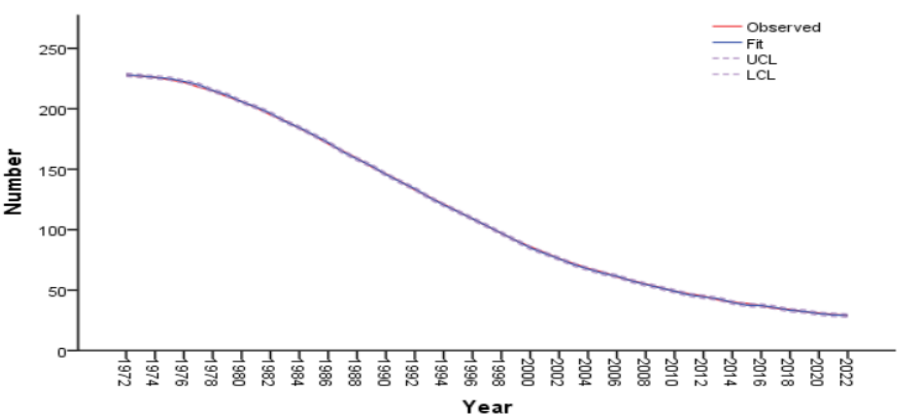


Figure 4: Time Series Data Plot of Number of under-five child deaths (per 1,000 live births) using Double Exponential Smoothing Holt's method

Figure 4 illustrates that the data pattern of under-five child mortality (per 1,000 live births) is a linear trend, thus Double Exponential Smoothing Holt's approach is the best and appropriate way to apply. In smoothing out linear trend data, this method employs two smoothing parameters: alpha (α) and gamma (γ).

Determination of Smoothing Parameters

The alpha parameter (α) is used to smooth the actual data while smoothing the trend on a regular basis. Parameter values ranging from 0 to 1 (Mega, 2003). The value of these factors can be determined through trial and error. This parameter determines the difference between the forecasted value and the actual data. When the alpha value approaches one, the weight given to the most recent data increases, resulting in the minor smoothing effect. When the alpha value approaches 0, it will have a minimal response to the most recent data, resulting in the substantial smoothing effect.

Although theoretically, alpha (α) and gamma (γ) can be assumed to be worth 0 and 1, in practice, alpha (α) and gamma (γ) parameters can only be determined using a limited range of values. This was narrowed due to the choice of alpha (α) and gamma (γ); the Double Exponential Smoothing method is typically seen as a more simply implemented method (Nurkse, 1953). The gamma parameter (γ) is used to remove some flexibility from the forecasted data. The result of the estimation of exponential smoothing model parameters is displayed in Table 4.

Table 4: Exponential Smoothing Model Parameters

Specification	Estimate
Alpha (Level)	.700
Gamma (Trend)	1.000

Table 4 displays the parameter estimates for the double exponential smoothing Holt's model. The result mentions that the model gave the estimated values of alpha and gamma of 0.700 and 1.000 respectively. The value of alpha (0.700) is relatively high, indicating that the estimate of the level at the current time point is based upon more recent observations. Again, the value of gamma (1.000) is high; indicating that the estimate of the trend at the current time point is based upon more recent observations.

3.2.3. Forecast Evaluation

The complete data are from 1972 to 2022, but to calibrate the model the study used data from 1972 to 2014. Then, the study used the data from 2015 to 2022 with the assumption that they have not been observed to conduct an out-of-sample forecast.

The values of the out-of-sample forecast as compared to the actual (observed) values are presented in Table 5.

Table 5: The Number of the under-five child mortality (per 1,000 live births) testing and forecast data in Bangladesh using both ARIMA (1,2,1) and Holt's Method

Year	Actual Value	ARIMA (1,2,1) Model		Holt's Method	
		Predicted Value	Difference	Predicted Value	Difference
2015	39	38	1	38	1
2016	37	37	0	37	0
2017	35	37	2	36	1
2018	34	37	3	33	1
2019	32	37	5	32	0
2020	31	38	7	30	1
2021	30	39	9	30	0
2022	29	41	12	29	0

Table 5 displays the values of the out-of-sample forecasts compared to the actual values for both methods. In comparison with ARIMA (1,2,1) model, the out-sample forecasting values of Double Exponential Smoothing Holt's method are closer to actual values. So, the Double Exponential Smoothing Holt's method gives a probable good forecast result.

3.2.4. Forecast Methods Comparison

In-sample forecast is the process of formally evaluating the predictive capabilities of the models developed using observed data to see how effective the algorithms are in reproducing data. It is kind of similar to a training set in a machine learning algorithm and the out-of-sample is similar to the test set. In this study, the total dataset period is from 1972 to 2022. The study considered the in-sample forecasting data period of 1972-2014 and out-sample forecasting data period of 2015–2022. To compare the performance of the best fitting models from the two methods, the ARIMA (1,2,1) model and Holt's Method, the study used three forecast error statistics: RMSE, MAPE, and MAE for this comparison. Hyndman (2006) mentioned that the out-of-sample period has smaller errors than the in-sample period because the in-sample period includes some relatively large observations. This error statistics are applied on the forecasts of the testing and training set (in-sample & out-of-sample forecasts) and are presented in Table 6.

Table 6: Forecast error statistics of the two forecast methods for the under-five child mortality (per 1,000 live births)

Model	In-Sample Forecasting Error			Out- Sample Forecasting Error		
	RMSE	MAPE	MAE	RMSE	MAPE	MAE
ARIMA (0,2,0)	0.632	0.504	0.469	0.567	0.444	0.276
Holt's Method	0.697	0.517	0.521	0.484	0.470	0.314

From Table 6, it could be seen that for both methods, in-sample forecasting errors are higher than out-sample forecasting errors.

3.2.5. Forecast Accuracy

The appropriate forecasting method is selected based on the accuracy of the measure set of values, whereby the method that has the smallest forecasting error is the best forecasting method. The general rule of thumb is the method with the lowest forecast error statistic is the best. The calculation of accuracy measure values of RMSE, MAE, and MAPE for those two selected methods is illustrated in Table 7.

Table 7: Result of comparison of forecast accuracy for forecasting of the under-five child mortality (per 1,000 live births)

Model	RMSE	MAPE	MAE
ARIMA (1,2,1)	0.661	0.713	0.513
Holt's Method	0.643	0.688	0.506

From Table 7, it is observed that compared to the ARIMA (1,2,1) model, the Double Exponential Smoothing Holt's method has the lowest values of RMSE, MAPE and MAE, which are 0.643%, 0.688% and 0.506%, respectively, whereas RMSE, MAPE and MAE for the ARIMA (1,2,1) model are 0.661%, 0.713% and 0.513% respectively. Hence, based on the findings, it is decided that the Double Exponential Smoothing Holt's method is the best and most appropriate forecasting method due to its high accuracy in forecasting the under-five child mortality (per 1,000 live births) in Bangladesh. Zainun and dan-Majid (2003) mentioned that if the MAPE value is less than 10%, the model performs extremely well; if the MAPE value is between 10% and 20%, the model performs well. The result displays in Table 7 that the MAPE value indicates the proportion of forecasting errors of 0.688% for the Double Exponential Smoothing Holt's method; this forecasting model performs very well with MAPE values below 10%.

3.2.6. Forecasting of Under-Five Child Mortality

The result of the forecasting plot of the under-five child mortality rate (per 1,000 live births) in Bangladesh is shown in Figure 5.

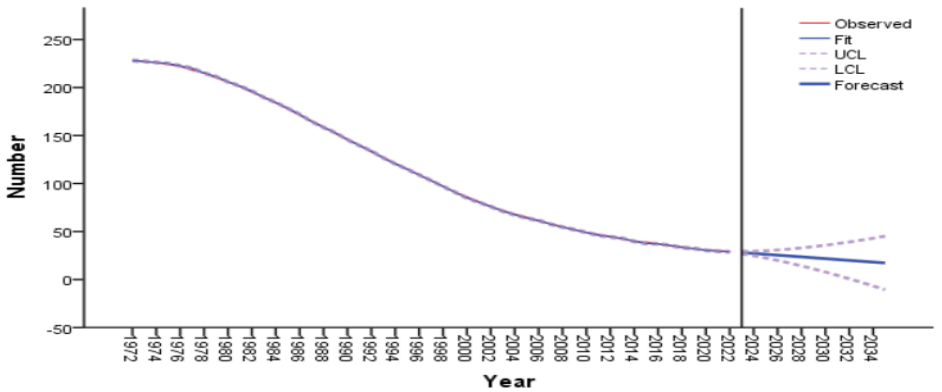


Figure 5: Predicted levels of the under-five mortality rate in Bangladesh

In Figure 5, the actual variables, fits, forecast, and 95% of confidence interval are shown. Alpha (α) of 0.700 and Gamma (γ) of 1.000 are the smoothing constants in this graph. The analysis of predicting models of the under-five child mortality (per 1,000 live births) is a data pattern with a linear trend; as a result, Holt's method is very appropriate for use. The forecasting values of the under-five child mortality (per 1,000 live births) using the Double Exponential Smoothing Holt's method for the period 2023–2035 are displayed in Table 8.

Table 8: The Result of forecasting of under-five child mortality (per 1,000 live births) for the period 2023–2035 using the Double Exponential Smoothing Holt's method

Years	**2023**	**2024**	**2025**	**2026**	**2027**	**2028**	**2029**	**2030**	**2031**	**2032**	**2033**	**2034**	**2035**
Forecasted value	28	27	26	25	24	24	23	22	21	20	19	18	17
GoB Targets			27										

From Table 8, it is observed that in Bangladesh, the value of under-five child mortality (per 1,000 live births) for the period 2023–2035 continues to decline year after year, with a trend linear data pattern. The result also shows that the average value of the under-five child mortality (per 1,000 live births) declined by 1 (per 1000 live births) during the period 2023–2035. Further, the value of the under-five child mortality (per 1,000 live births) will reach 26 (per 1000 live births) in 2025 and 22 (per 1000 live births) in 2030 whereas the government of Bangladesh set a target of 27 for the under-five child mortality rate (per 1,000 live births) by 2025 in the 8th five-year plan (GoB, 2020b). From the findings of this study, it is clearly understood that Bangladesh will achieve the SDGs target of the under-five child mortality by 2026 if the current declining trend is continued.

4. Conclusion and Recommendation

Child mortality is a crucial indicator of a nation's socio-economic advancement and the well-being of mothers, reflecting the overall quality of life within a society. Bangladesh, as one of the developing nations in Asia, has committed to pursuing the SDGs. The objective of achieving a reduction in under-five mortality to a level no higher than 25 per 1,000 live births by the year 2030 is commonly known as the third Sustainable Development Goal (SDG) (UN, 2015). The study explores the under-five child mortality trends and estimates projection by 2030 to make progress towards achieving the SDGs target regarding the under-five child mortality in Bangladesh. The yearly dataset regarding the under-five child mortality in Bangladesh utilized in this study is collected from the World Bank Databank (https://data.worldbank.org/indicator) for the period 1972–2022 by the indicator "Mortality rate, under-5 (per 1,000 live births)". For selecting the best-fitted model for the purpose of forecasting, the ARIMA model and the Double Exponential Smoothing Holt's Method are employed. As compared to the ARIMA (1, 2, 1) model, the Double Exponential Smoothing Holt's method is the best-fitted model for forecasting the estimate of future projections. The result depicts that the under-five child mortality rate in Bangladesh is a continuously declining trend in each and every year. The result also shows that the under-five child mortality rate will drop by 1 (per 1,000 live births) during 2023-2035. The under-five child mortality rate is expected to fall to 26 in 2025 and to 22 in 2030, whereas Bangladesh's national target is to reduce the under-five child mortality to 27 (per 1,000 live births) by 2025 in the 8th five-year plan. From the result of this study, it is clearly understood that Bangladesh will achieve the SDGs target regarding the under-five child mortality by 2026 if the current declining trend is continued. The factors that influence the under-five child mortality rate are education, awareness and health campaigning, access to health care, maternal health care, access to antennal care and prenatal care, vaccination, housing condition, earning patterns of household and various health and various strategies and initiatives of government and NGOs.

Abbreviations

Abbreviation	Details
ACF	Autocorrelation Function
AR	Autoregressive (AR)
ARIMA	Auto-Regressive Integrated Moving Average
BDHS	Bangladesh Demographic and Health Survey
BIC	Bayesian Information Criterion
GoB	Government of Bangladesh

(cont.)

Abbreviation	Details
LCL	Lower Confidence Limits
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MaxAE	Maximum Absolute Error
MaxAPE	Maximum Absolute Percentage Error
MDGs	Millennium Development Goals
MICS	Multiple Indicator Cluster Survey
PACF	Partial Autocorrelation Function
RMSE	Root Mean Squared Error
SDGs	Sustainable Development Goals
UCL	Upper Confidence Limits
UN	United Nations
UNICEF	United Nations International Children's Emergency Fund
UNIGM	United Nations Inter-agency Group for Child Mortality Estimation
WHO	World Health Organization

References

Ahmed, T., Mahfuz, M., Ireen, S., Ahmed, A. S., Rahman, S., Islam, M. M., et al., (2012). Nutrition of children and womenin Bangladesh: trends and directions for the future. *Journal of Health, Population, and Nutrition,* Vol. 30, No. 1, pp. 1-11.

BDHS, (2014). Bangladesh Demographic and Health Survey (BDHS), 2014. NIPORT, Dhaka, Bangladesh; Mitra andAssociates, Dhaka/ Bangladesh.

Box, G. E. P., Jenkins, G. M., (1976). Time series analysis, forecasting and control. Holden Day: San Francisco, California, USA, p. 625.

Box, G. E. P., Pierce, D., (1970). Distribution of Residual Autocorrelations in Auto-regressive Integrated Moving Average Time Series Models. *Journal of the American Statistical Association,* Vol. 65, pp. 1509–1526.

Chao, F., et al., (2018). National and regional under-5 mortality rate by economic status for low-income and middle income countries: a systematic assessment. *The Lancet Global Health*, Vol. 6, No. 5, pp. 535–547. https://doi.org/10.1016/S2214-109X(18)30059-7.

Chen, H. U., Guoa, S., Chong-Yu X. U., Singh, V. P., (2007). Historical temporal trends of hydro-climatic variables and runoff response to climate variability and their relevance in water resource management in the Hanjiang basin. *Jour. Hydrol.,* Vol. 344, pp. 171–184.

GoB, (2020a). Sustainable Development Goals: Bangladesh Progress Report 2020, SDGs Publication No. # 23 by Bangladesh Planning Commission, Ministry of Planning, Government of the People's Republic of Bangladesh (GoB), Dhaka.

GoB, (2020b). *8th Five Year Plan (July 2020-June 2025),* General Economics Division (GED), Bangladesh Planning Commission, Government of the People's Republic of Bangladesh (GoB), Dhaka.

Hesel, D. R., Hirsch, R. M., (1992). *Statistical Methods in Water Resources*, Elsevier, Amsterdam.

Holt, C. E., (1957). *Forecasting seasonals and trends by exponentially weighted averages* (O.N.R. Memorandum No. 52). Carnegie Institute of Technology, Pittsburgh USA. https://doi.org/10.1016/j.ijforecast.2003.09.015

Hyndman, R. J., (2006). Another Look At Forecast-Accuracy Metrics For Intermittent Demand. *Foresight, 4,* pp. 43–46.

Khan, J. R., Awan N., (2017). A comprehensive analysis on child mortality and its determinants in Bangladesh using frailty models. *Archives of Public Health,* Vol. 75, No. 1, p. 58. https://doi.org/10.1186/s13690-017-0224-6.

Kendall, M. G. (1975). Rank Correlation Methods. Charles Griffin: London.

Machiwal, D., Jha, M. K., (2008). Comparative evaluation of statistical tests for time series analysis: Application to hydrological time series. *Hydrological Sciences, Journal-des Sciences Hydrologiques,* Vol. 53, No. 3, pp. 353–366.

Mann, H. B., (1945). Nonparametric tests against trend. *Econometrica,* Vol. 13, pp. 245–259.

McGuire, J. W., (2006). Basic health care provision and under-5 mortality: a cross-national study of developing countries. *World Development*, Vol. 34, NO. 3, pp. 405–425.

Mega, F., (2003). *Strategi Bersama Masyarakat Sipil Indonesia: Empat Pilar Demokratisasi untuk Melawan Kemiskinan dan Pemiskinan.* Jakarta: Gerakan Anti Pemiskinan (GAPRI).

MICS, (2019). Progotir Pathey, *Bangladesh Multiple Indicator Cluster Survey*, Survey Findings Report. Dhaka, Bangladesh: Bangladesh Bureau of Statistics (BBS) and UNICEF Bangladesh.

Nino, F. S., (2015). *Sustainable Development Goals*. United Nations.

Nurkse, R., (1953). *Problems of Capital Formation in Underdeveloped Countries*.

UN, (2015). *The Millennium Development Goals Report*. The United Nations, New York, USA.

UNICEF, (2018). *2018–2021 Progress Report - Save Newborns*, UNICEF South Asia.

UNIGME, (2020). *'Levels & Trends in Child Mortality: Report 2020*. The United Nations Inter-agency Group for Child Mortality Estimation (UNIGME), United Nations International Children's Emergency Fund, New York.

WHO, (2018). *WHO | Under-five mortality*, World Health Organization (WHO).

WHO, (2016). *World health statistics 2016: monitoring health for the SDGs sustainable development goals*, World Health Organization (WHO).

Wilson, J. H., Keating, B., (2007). *Business Forecasting with Accompanying Excel-Based ForecastX Software*, 5th ed., Boston, Mass: McGraw-Hill.

Zainun, N. Y., dan Majid, Z. A., (2003). *Low Cost Demand Predictor*, Malaysia. Universitas Teknologi Malaysia.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 137–155, https://doi.org/10.59139/stattrans-2024-007
Received – 08.04.2022; accepted – 22.04.2024

On some statistical properties of a stationary Gaussian process in the presence of measurement errors

Kuntal Bera[1], **M. Z. Anis**[2]

Abstract

Process outputs of many production processes like chemical, food processing and pharmaceutical industry follow a stationary Gaussian process. Some amount of measurement error always present in the measured data due to inaccurate measuring processes. Throughout this paper, we discuss some statistical properties like the mean and variance of a stationary Gaussian process when observed data are affected by measurement errors. As a special case, we discuss a stationary autoregressive process of order one with Gaussian white noise where measurement error follows an independent Gaussian distribution.

Key words: autoregressive process, central moments, measurement errors, white noise.

1. Introduction

Measurement error is a general problem while collecting data and hence the usage of such data may lead to improper inference. The variable of interest, say X, cannot be measured accurately in the presence of measurement error. Koutsoyiannis (1977) presents interesting examples of measurement errors in economics. The presence of measurement error has a profound influence in almost every area dealing with the measurement of the sample. Maleki et al. (2017) pointed out that in spite of refined and sophisticated measuring devices, real-life data are contaminated with measurement errors. Hence, these measurement errors need to be taken into account while monitoring items. There is a large body of literature dealing with the effect of measurement error in many areas including statistical process control (SPC), economics, medical studies, environmental studies, agricultural studies and others. See, for example, Carroll (1998), Linna and Woodall (2001), Noor-ul Amin et al. (2022), Schennach (2016), Blackwell et al. (2017), Abay et al. (2023) and the references therein. Wu (2011) cautioned that if measurement error is ignored, it may lead to unreliable decisions for the process under study. Attention should first be paid to the measurement system to ascertain if the variability significantly increases due to the presence of measurement error in the data.

Autocorrelation is an inherent property of many processes, see, for example, Shumway and Stoffer (2017). The interval between process observations is decreasing due to the

[1] SQC & OR Unit, Indian Statistical Institute, Kolkata-700108, India. E-mail: bera.kuntal12345@gmail.com. ORCID: https://orcid.org/0000-0003-1217-5708.

[2] SQC & OR Unit, Indian Statistical Institute, Kolkata-700108, India. E-mail: zafar@isical.ac.in. ORCID: https://orcid.org/0000-0001-5546-0723.

rampant use of online data acquisition systems, leading to positive autocorrelation as noted by Runger and Willemain (1995). Such trends are more pronounced in the process and chemical industries, the practice of measuring every part produced induces positive autocorrelation even in discrete parts measurement. Statistical tests (e.g. the Durbin-Watson test and the Bartlett test) can be used to detect the presence of first-order and higher-order autocorrelation. Zhang (1998) has estimated the variance and expected value of the sample mean and the sample variance for a stationary Gaussian process.

In many industrial production processes it is found that quite a few of the quality characteristics of output products follow stationary Gaussian process, see Box et al. (2015). To control the production process economically, knowledge of the mean and variance of the respective quality characteristics is necessary. For instance, the knowledge of the mean and variance of a process is required for many statistical process control techniques like estimation of control limits, estimation of process capability, estimation of percentage of conforming quality, etc.

In this work, we discuss some statistical properties of such autocorrelated processes when the observed data is contaminated with measurement error and thereby extend the work of Zhang (1998). The paper is organized as follows. In Section 2, we discuss the statistical properties of a stationary Gaussian process. These results are extended to account for measurement errors in Section 3. The impact of measurement errors in estimating the mean and variance of a stationary $AR(1)$ process is shown graphically as a special case. Some simulation studies are reported and an industrial example is reported in Section 4. An industrial application is mentioned in Section 5 and Section 6 concludes the paper.

2. Statistical Properties of a Stationary Gaussian Process

Let $\{X_t, t \in \mathbb{Z}\}$ be a discrete-time stationary Gaussian process. A time series $\{X_t\}$ is said to be a stationary process if $Var(X_t) < \infty$, expectation $E(X_t)$ is independent of time t and auto-covariance function $Cov(X_{t+h}, X_t)$ depends only on difference between the time points; i.e., h and is independent of t. For a stationary time series X_t, let

$$E(X_t) = \mu_X$$

and

$$Cov(X_{t+h}, X_t) = \gamma_X(h).$$

A time series $\{X_t\}$ is said to be a Gaussian process if the joint distribution of any finite n number of random variables $\{X_1, X_2, X_3, \ldots, X_n\}$ from the process follows a multivariate normal distribution; see Kotz et al. (2019) for details about properties of multivariate normal distribution. A process that is stationary and Gaussian simultaneously is said to be a stationary Gaussian process, see Brockwell and Davis (2002) for details. A stationary Gaussian process is also strictly stationary. A time series is said to be *strictly stationary* if the probabilistic behaviour of every finite collection of values $\{X_{t_1}, X_{t_2}, \ldots, X_{t_n}\}$ is identical

with the time shifted values $\{X_{t_1+h}, X_{t_2+h}, \ldots, X_{t_n+h}\}$.

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of n consecutive observations from a stationary Gaussian process. Then, the sample mean and the sample variance of the process are, respectively, given by $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $S_X^2 = \frac{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right]}{n-1}$. Zhang (1998) found the expected value and variance of $\bar{X}$ and S_X^2. We shall stick to the same notation as used by Zhang (1998); but for completeness, we shall define the functions that will be used subsequently.

Let

$$\rho_i = \rho_X(i) = \frac{\gamma_X(i)}{\gamma_X(0)}$$

for $i = 1, 2, 3, \ldots, n$ be the autocorrelation of X_t at lag i. Define,

$$f(n, \rho_i) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{n-1} (n-i)\, \rho_i, \tag{1}$$

$$F(n, \rho_i) = n + 2\sum_{i=1}^{n-1} (n-i)\, {\rho_i}^2 + \frac{1}{n^2}\left[n + 2\sum_{i=1}^{n-1}(n-i)\rho_i\right]^2 - \frac{2}{n}\sum_{i=0}^{n-1}\sum_{j=0}^{n-i}(n-i-j)\,\rho_i\rho_j \tag{2}$$

and

$$g(n, \rho_i) = 1 + \frac{2}{n}\sum_{i=1}^{n-1}(n-i)\rho_i. \tag{3}$$

Observe that when the process $\{X_t\}$ is identically and independently normally distributed, then $\rho_i = 0$ for $i \geq 1$. In this case $f(n, \rho_i) = 1$, $g(n, \rho_i) = 1$ and $F(n, \rho_i) = (n-1)$.

Since $\{X_t\}$ is a Gaussian process, therefore $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2 g(n,\rho_i)}{n}\right)$. Hence, the $r-$th order central moment of $\bar{X}$ is

$$\mathbb{E}\,(\bar{X} - \mu_X)^r = \begin{cases} 0 & \text{when } r \text{ is an odd integer} \\ 1.3.5\ldots(2k-1)\left[\frac{\sigma_X^2 g(n,\rho_i)}{n}\right]^{2k} & \text{when } r = 2k \text{ for } k = 1, 2, 3, \ldots. \end{cases} \tag{4}$$

Here, we discuss the particular case of an $AR(1)$ process. $AR(1)$ process is very popular and often occurs in many chemical and process industries for modeling autocorrelation structures. Note that the $AR(1)$ process is a special case of stationary Gaussian process. An $AR(1)$ process with mean μ is defined as

$$X_t - \mu = \phi\,(X_{t-1} - \mu) + a_t, \qquad |\phi| < 1. \tag{5}$$

When a_t is a Gaussian white noise, then $\{X_t\}$ will be a stationary Gaussian process. Suppose $a_t \sim IIDN(0, \sigma_a^2)$. For an $AR(1)$ process defined by equation (5), $\rho_i = \phi^i$. So, the expected value and variance of the sample mean and the sample variance, in this case, is

$$\mathbb{E}\,(\bar{X}) = \mu_X,$$

$$Var(\bar{X}) = \frac{\sigma_X^2}{n} g(n,\phi),$$
$$\mathbb{E}\left(S_X^2\right) = \sigma_X^2 f(n,\phi),$$

and

$$Var\left(S_X^2\right) = \frac{2\sigma_X^4}{(n-1)^2} F(n,\phi)$$

where

$$\sigma_X^2 = \frac{\sigma_a^2}{(1-\phi^2)}, \tag{6}$$

$$f(n,\phi) = 1 - \frac{2}{n(n-1)} \frac{\phi\left[n-1-n\phi+\phi^n\right]}{(1-\phi)^2}, \tag{7}$$

$$F(n,\phi) = n + 2\sum_{i=1}^{n-1}(n-i)\,\phi^{2i} + \frac{1}{n^2}\left[n + 2\sum_{i=1}^{n-1}(n-i)\phi^i\right]^2 - \frac{2}{n}\sum_{i=0}^{n-1}\sum_{j=0}^{n-i}(n-i-j)\,\phi^{i+j}, \tag{8}$$

and

$$g(n,\phi) = 1 + \frac{2}{n} \frac{\phi\left[n-1-n\phi+\phi^n\right]}{(1-\phi)^2}. \tag{9}$$

Thus, the sample mean is an unbiased estimator of the population mean while the sample variance is a biased estimator of population variance. The variance of the sample mean can be expressed as

$$\begin{aligned} Var(\bar{X}) &= \frac{\sigma_a^2 g(n,\phi)}{n(1-\phi^2)} \\ &= \sigma_a^2 R_1(n,\phi) \end{aligned}$$

where

$$R_1(n,\phi) = \frac{g(n,\phi)}{n(1-\phi^2)}.$$

The bias and mean square error of the sample variance are given by

$$\begin{aligned} Bias(S_X^2) &= \sigma_X^2\left[f(n,\phi)-1\right] \\ &= \frac{\sigma_a^2[f(n,\phi)-1]}{(1-\phi^2)} \\ &= \sigma_a^2 R_2(n,\phi) \end{aligned}$$

$$\begin{aligned} MSE(S_X^2) &= Var\left(S_X^2\right) + \left\{Bias(S_X^2)\right\}^2 \\ &= \frac{2\sigma_a^4}{(1-\phi^2)^2}\left[\frac{F(n,\phi)}{(n-1)^2} + \frac{\{f(n,\phi)-1\}^2}{2}\right] \\ &= \sigma_a^4 R_3(n,\phi) \end{aligned}$$

where

$$R_2(n,\phi) = \frac{[f(n,\phi)-1]}{(1-\phi^2)},$$

$$R_3(n,\phi) = \frac{2}{(1-\phi^2)^2}\left[\frac{F(n,\phi)}{(n-1)^2} + \frac{\{f(n,\phi)-1\}^2}{2}\right].$$

Similarly, the $r-$th order central moment of $\bar{X}$ in this special case will be,

$$\mathbb{E}\left(\bar{X}-\mu_X\right)^r = \begin{cases} 0 & \text{when } r \text{ is an odd integer} \\ 1.3.5\ldots(2k-1)\left[\frac{\sigma_X^2 g(n,\phi)}{n}\right]^{2k} & \text{when } r = 2k \text{ for } k = 1,2,3,\ldots. \end{cases} \tag{10}$$

We graphically present $R_1(n,\phi)$, $R_2(n,\phi)$ and $R_3(n,\phi)$ to show the effect of autocorrelation and the sample size on estimating mean and variance when process observations are autocorrelated. Note that when sample observations are independent, then $\phi = 0$ and in this case $R_1(n,\phi)$, $R_2(n,\phi)$, $R_3(n,\phi)$ will become only the function of the sample size n.

From Figure 1 it is clear that the variance of the sample mean increases as the autocorrelation level increases. For small sample size ($n < 50$), this variance is significantly large when autocorrelation is high ($\phi \geq 0.50$). The variance of the sample mean decreases with the increase of the sample size as expected. From the graph of $R_2(n,\phi)$ in Figure 2 we notice that the sample variance is a biased estimate of population variance in the presence of autocorrelation. Note that sample variance is an unbiased estimate of population variance when sample observations are independent. Sample variance is underestimated in the presence of autocorrelation. The bias of the sample variance is significantly large when the sample size is small and autocorrelation is high. From the graph of $R_3(n,\phi)$ in Figure 3 we notice that the MSE of the sample variance increases as autocorrelation increases. This MSE is significantly large when the sample size is small. This happens because uncertainty within the sample increases due to an increase of autocorrelation and as a result, in such cases, a sample of small size cannot properly estimate the mean and variance of the population. Therefore, to avoid estimation error due to autocorrelation, a relatively large sample size is required for estimating mean and variance.

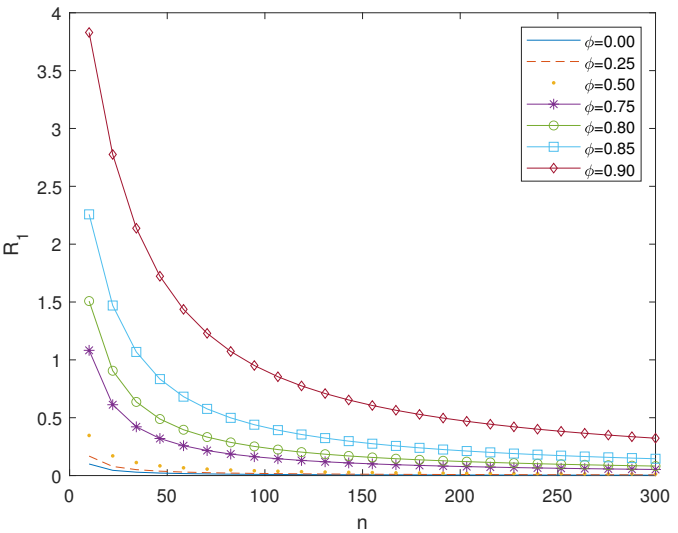


Figure 1: Graph of $R_1(n,\phi)$ corresponding to different values of ϕ.

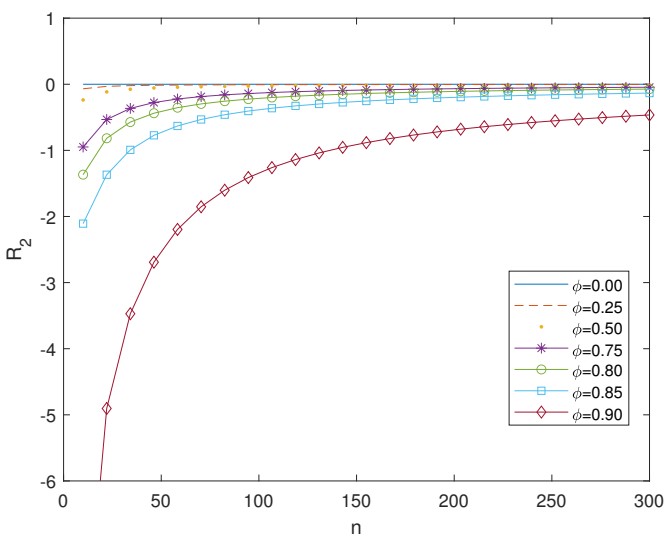


Figure 2: Graph of $R_2(n,\phi)$ corresponding to different values of ϕ.

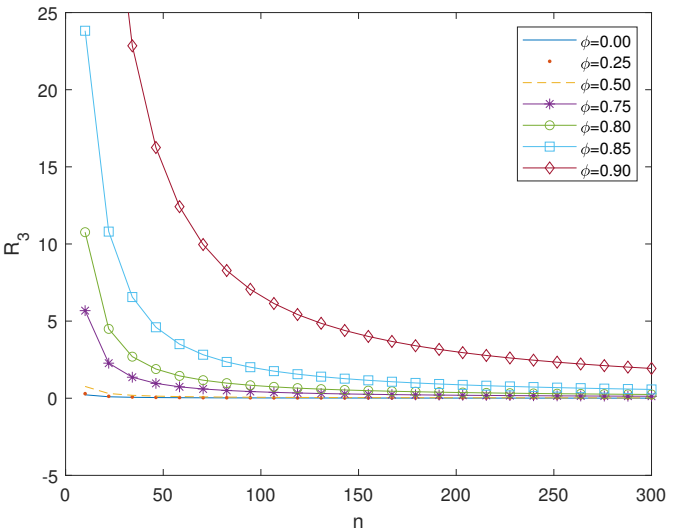


Figure 3: Graph of $R_3(n,\phi)$ corresponding to different values of ϕ.

3. Statistical Properties of a Stationary Gaussian Process in the Presence of Measurement Errors

In practical situations, true values of process outputs are often unobservable as they are contaminated by measurement errors. Instead of the true process $\{X_t\}$, we observe the process $\{Y_t\}$, where Y_t is defined by

$$Y_t = X_t + E_t. \tag{11}$$

Here, E_t is a random measurement error variable. Assume that $E_t \sim N(0,\sigma_E^2)$; X_t and E_t are stochastically independent. Let $\{Y_1, Y_2, \ldots, Y_n\}$ be a random sample of size n from the observable process $\{Y_t\}$. Thus, the sample mean and the sample variance, using the observable data, are given by $\bar{Y} = \frac{\Sigma_{i=1}^n Y_i}{n}$ and $S_Y^2 = \frac{\left[\Sigma_{i=1}^n (Y_i - \bar{Y})^2\right]}{n-1}$ respectively.

3.1. Statistical Analysis of Sample Mean

It is easy to see that

$$\mathbb{E}(\bar{Y}) = \mu_X; \tag{12}$$

and

$$Var(\bar{Y}) = \frac{\sigma_X^2}{n} g(n,\rho_i) + \frac{\sigma_E^2}{n}. \tag{13}$$

As $\bar{X} \sim N\left(\mu_X, \frac{\sigma_X^2 g(n,\rho_i)}{n}\right)$; $\bar{E} \sim N\left(0, \frac{\sigma_E^2}{n}\right)$ and $\bar{Y}$ is the sum of two normal variables, it follows that $\bar{Y} \sim N\left(\mu_X, \frac{\left[\sigma_X^2 g(n,\rho_i)+\sigma_E^2\right]}{n}\right)$. Hence, the $r-$th order central moment of $\bar{Y}$ is given by

$$\mathbb{E}(\bar{Y}-\mu_X)^r = \begin{cases} 0 & \text{when } r \text{ is an odd integer} \\ 1.3.5\ldots(2k-1)\left[\frac{\sigma_X^2 g(n,\rho_i)+\sigma_E^2}{n}\right]^{2k} & \text{when } r=2k \text{ for } k=1,2,3,\ldots. \end{cases} \tag{14}$$

In the particular case of an $AR(1)$ process defined by equation (5), we have

$$\begin{aligned} Var(\bar{Y}) &= \frac{\sigma_X^2}{n} g(n,\phi) + \frac{\sigma_E^2}{n} \\ &= \frac{\sigma_X^2}{n}\left[g(n,\phi) + (1-\phi^2){\tau_a}^2\right] \\ &= \sigma_a^2 R_1^e(n,\phi,\tau_a) \end{aligned}$$

where

$$R_1^e(n,\phi,\tau_a) = \frac{\left[g(n,\phi) + (1-\phi^2){\tau_a}^2\right]}{n(1-\phi^2)}. \tag{15}$$

Here, σ_a^2 is the variance of Gaussian white noise a_t and $R_1^e(n,\phi,\tau_a)$ is defined by the equation (15). Here, τ_a is the ratio between the standard deviation of measurement error E_t and the standard deviation of the Gaussian white noise a_t and is defined by

$$\tau_a = \frac{\sigma_E}{\sigma_a}. \tag{16}$$

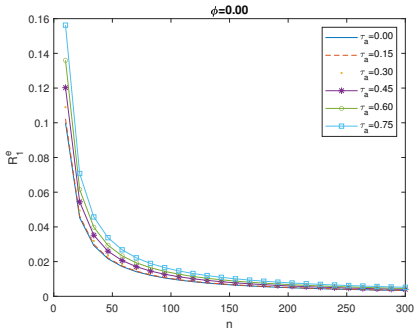
To see the combined effect of measurement error and autocorrelation on the estimation of mean, we graphically analyze the function $R_1^e(n,\phi,\tau_a)$. Note the $R_1^e(n,\phi,\tau_a)$ is a function of autocorrelation parameter ϕ, degree of error contamination τ_a and the sample size n. In case sample observations are free of measurement error then $\tau_a = 0$ and when independent then $\phi = 0$.

It is clear from Figure 4 and Figure 5 that the variance of the sample mean has some increment due to the presence of measurement error. But this increment is not significant. Hence, measurement error does not affect seriously the sample mean.

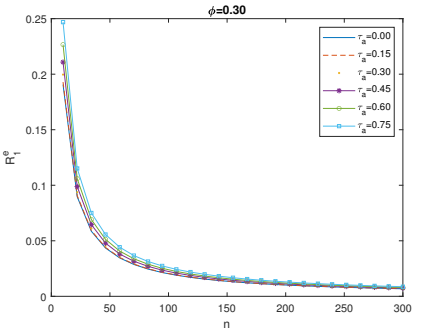
3.2. Statistical Analysis of Sample Variance

It is easy to see that

$$S_Y^2 = S_X^2 + \frac{2}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(E_i-\bar{E}) + S_E^2;$$

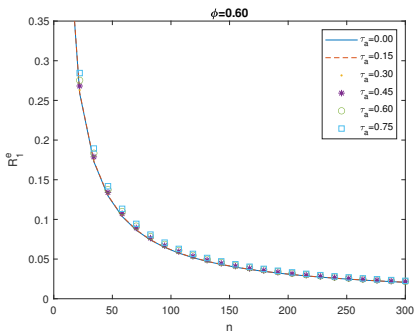


(a) Graph of $R_1^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.00$.

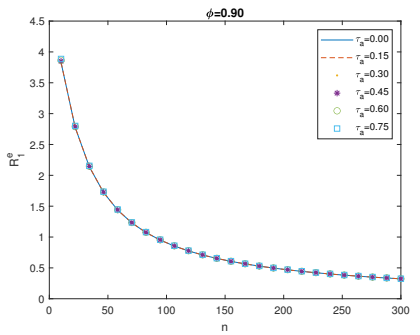


(b) Graph of $R_1^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.30$.

Figure 4



(a) Graph of $R_1^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.60$.



(b) Graph of $R_1^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.90$.

Figure 5

and on taking expectations we get

$$\mathbb{E}(S_Y^2) = \sigma_X^2 f(n,\rho_i) + \sigma_E^2; \tag{17}$$

If $\{X_t\}$ is an AR(1) process defined by equation (5), then

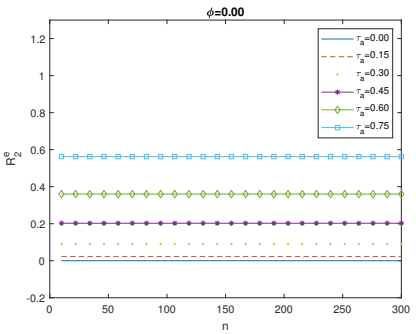
$$\mathbb{E}(S_Y^2) = \sigma_X^2 \left[f(n,\phi) + (1-\phi^2)\tau_a^2 \right].$$

In this case, sample variance is a biased estimator of population variance. The bias of the estimator with respect to the true process variance is given by

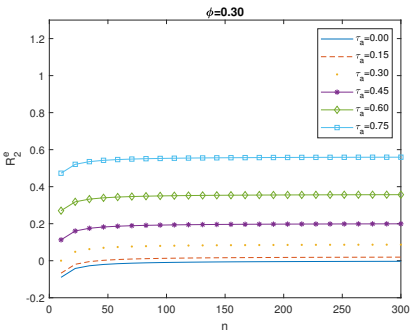
$$\begin{aligned} Bias(S_Y^2) &= E\left(S_Y^2\right) - \sigma_X^2 \\ &= \sigma_X^2 \left[\{ f(n,\phi) - 1 \} + (1-\phi^2)\tau_a^2 \right] \\ &= \sigma_a^2 R_2^e(n,\phi,\tau_a) \end{aligned}$$

where

$$R_2^e(n,\phi,\tau_a) = \frac{\left[f(n,\phi) - 1 + (1-\phi^2){\tau_a}^2\right]}{(1-\phi^2)}. \tag{18}$$

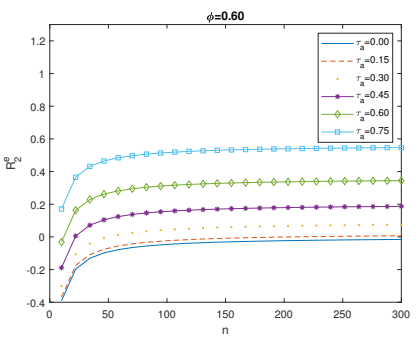


(a) Graph of $R_2^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.00$.

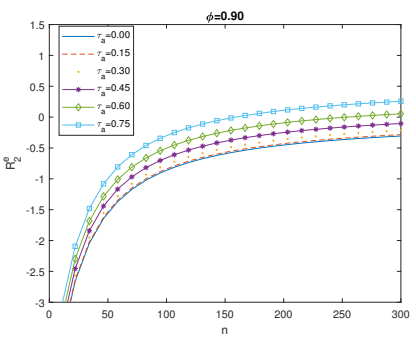


(b) Graph of $R_2^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.30$.

Figure 6



(a) Graph of $R_2^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.60$.



(b) Graph of $R_2^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.90$.

Figure 7

We have observed from Figure 2 of Section 2 that the sample variance is underestimated in the presence of autocorrelation. Also, we have noticed that the sample variance is unbiased when the sample is independent. From Figure 6(a) we notice that the sample variance is overestimated in the presence of the measurement error when sample observations are independent. However, a mixed effect is observed in the presence of both autocorrelation and measurement error. Autocorrelation tries to underestimate the sample variance while measurement error attempts to overestimate it. As the effect of underestimation is large when autocorrelation is present and the sample size is small, a relatively small value of $R_2^e(n,\phi,\tau_a)$ is visible for a small sample size.

The expression for the variance of the sample variance is given in the equation (19). Detailed calculation is provided in the Appendix.

$$Var(S_Y^2) = \frac{2\sigma_X^4}{(n-1)^2}\left\{F(n,\rho_i) + (n-1)\frac{\sigma_E^4}{\sigma_X^4} + 2(n-1)f(n,\rho_i)\frac{\sigma_E^2}{\sigma_X^2}\right\}. \tag{19}$$

Now, if $\{X_t\}$ is an $AR(1)$ process defined by equation (5), then we can write the expression (19) as:

$$Var(S_Y^2) = \frac{2\sigma_X^4}{(n-1)^2}\left\{F(n,\phi) + (n-1)(1-\phi^2)^2\tau_a^4 + 2(n-1)(1-\phi^2)\tau_a^2 f(n,\phi)\right\}.$$

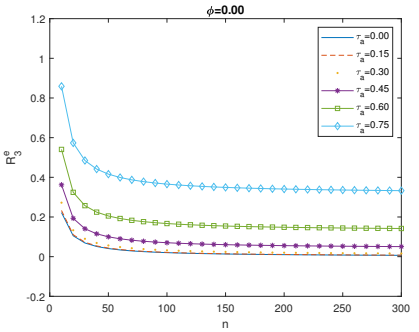
The mean square error of the estimator of the sample variance about the true process variance, in this case, is

$$\begin{aligned} MSE(S_Y^2) &= \frac{2\sigma_X^4}{(n-1)}\left\{\frac{\left[F(n,\phi) + (n-1)(1-\phi^2)^2\tau_a^4 + 2(n-1)(1-\phi^2)\tau_a^2 f(n,\phi)\right]}{(n-1)}\right. \\ &\qquad \left. + \frac{(n-1)[\{f(n,\phi)-1\} + (1-\phi^2){\tau_a}^2]^2}{2}\right\} \\ &= \sigma_a^4 R_3^e(n,\phi,\tau_a) \end{aligned} \tag{20}$$

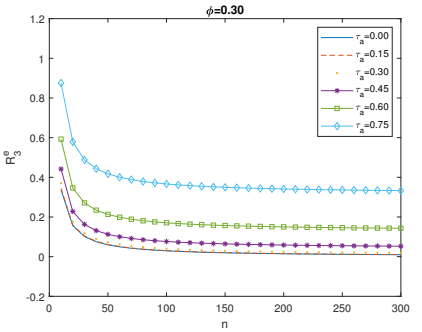
where

$$\begin{aligned} R_3^e(n,\phi,\tau_a) &= \frac{2}{(1-\phi^2)^2}\left\{\frac{\left[F(n,\phi) + (n-1)(1-\phi^2)^2\tau_a^4 + 2(n-1)(1-\phi^2)\tau_a^2 f(n,\phi)\right]}{(n-1)^2}\right. \\ &\qquad \left. + \frac{[\{f(n,\phi)-1\} + (1-\phi^2){\tau_a}^2]^2}{2}\right\}. \end{aligned} \tag{21}$$

The different behaviour of $MSE(S_Y^2)$ for different values of ϕ, τ_a and n is observable from Figure 8 and Figure 9. We can see from these figures that as measurement error increases, the MSE value also increases. But for highly autocorrelated data ($\phi = 0.90$), as can be seen from Figure 9(b), the value of MSE relatively decreases as measurement error increases.

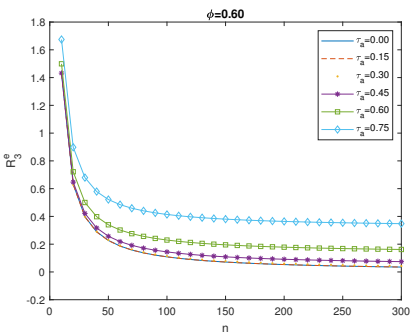


(a) Graph of $R_3^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.00$.

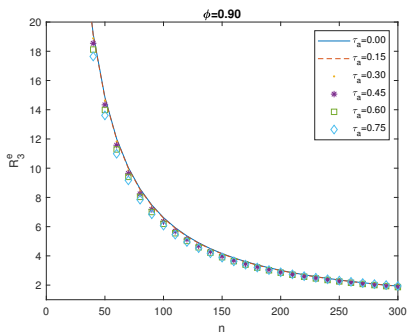


(b) Graph of $R_3^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.30$.

Figure 8



(a) Graph of $R_3^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.60$.



(b) Graph of $R_3^e(n,\phi,\tau_a)$ corresponding to different values of τ_a for $\phi = 0.90$.

Figure 9

4. Simulation

We carried out a simulation exercise to study the behavior of the mean and variance of the sample mean and the sample variance. We simulate 5000 random samples of size 50(25)200 respectively from each of an IID normal process, a stationary $AR(1)$ process and a stationary $AR(1)$ process with random measurement errors separately. In each of these cases, we compare the sample estimated value of the mean and variance of the estimators with the corresponding theoretical values. Here, `hat` (^) indicates the sample estimated value.

4.1. Case of an IID normal Process

Here, we simulate a model of iid random normal variables $X_t \sim N(\mu,\sigma^2)$ where $\mu = 5$ and $\sigma^2 = 4$.

Table 1: Results based on iid normal model X_t.

n	$E(\bar{X})$	$\hat{E}(\bar{X})$	$Var(\bar{X})$	$\hat{Var}(\bar{X})$	$E(S_X^2)$	$\hat{E}(S_X^2)$	$Var(S_X^2)$	$\hat{Var}(S_X^2)$
50	5	5.0068	0.0800	0.0778	4	3.9759	0.6531	0.6700
75	5	5.0058	0.0533	0.0545	4	3.9784	0.4324	0.4337
100	5	5.0056	0.0400	0.0404	4	3.9824	0.3232	0.3235
125	5	5.0045	0.0320	0.0317	4	3.9865	0.2581	0.2580
150	5	5.0038	0.0267	0.0264	4	3.9898	0.2148	0.2145
175	5	4.9967	0.0229	0.0227	4	4.0015	0.1839	0.1833
200	5	4.9968	0.0200	0.0201	4	4.0008	0.1608	0.1598

4.2. Case of a stationary $AR(1)$ process

Here, we simulate a model from a stationary $AR(1)$ model defined by equation (5) where $\mu = 5$, $\phi = 0.5$ and white noise $a_t \sim N(0,1)$.

Table 2: Results based on an $AR(1)$ model X_t.

n	$E(\bar{X})$	$\hat{E}(\bar{X})$	$Var(\bar{X})$	$\hat{Var}(\bar{X})$	$E(S_X^2)$	$\hat{E}(S_X^2)$	$Var(S_X^2)$	$\hat{Var}(S_X^2)$
50	5	4.9981	0.0779	0.0758	1.2811	1.2729	0.1233	0.1072
75	5	4.9980	0.0524	0.0510	1.2983	1.2939	0.0812	0.0746
100	5	4.9989	0.0395	0.0390	1.3069	1.3038	0.0605	0.0575
125	5	4.9992	0.0317	0.0313	1.3122	1.3116	0.0482	0.0461
150	5	4.9995	0.0264	0.0261	1.3157	1.3139	0.0400	0.0386
175	5	4.9997	0.0227	0.0225	1.3182	1.3163	0.0343	0.0330
200	5	4.9997	0.0199	0.0197	1.3201	1.3201	0.0299	0.0294

4.3. Case of an $AR(1)$ process in the presence of measurement errors

Here, we simulate a stationary $AR(1)$ process in the presence of measurement errors. The model is defined by equation (11). In this model X_t is same as defined in the previous cases and $E_t \sim iid\ N(0, \sigma_E^2)$ where $\sigma_E^2 = 4$.

Table 3: Results based on an $AR(1)$ model in the presence of measurement error, Y_t.

n	$E(\bar{Y})$	$\hat{E}(\bar{Y})$	$Var(\bar{Y})$	$\hat{Var}(\bar{Y})$	$E(S_Y^2)$	$\hat{E}(S_Y^2)$	$Var(S_Y^2)$	$\hat{Var}(S_Y^2)$
50	5	4.9990	0.1579	0.1545	5.2811	5.2856	1.1947	1.1494
75	5	4.9999	0.1057	0.1039	5.2983	5.2996	0.7943	0.7836
100	5	5.0001	0.0795	0.0782	5.3069	5.2945	0.5949	0.5851
125	5	4.9986	0.0637	0.0630	5.3122	5.3116	0.4756	0.4718
150	5	5.0020	0.0531	0.0526	5.3157	5.2947	0.3961	0.3938
175	5	5.0027	0.0455	0.0452	5.3182	5.3093	0.3394	0.3367
200	5	5.0017	0.0399	0.0395	5.3201	5.3159	0.2969	0.2944

5. An industrial application

As an application, we will discuss the case of estimation of the control limits of the mean chart in statistical process monitoring. Let, the true process $\{X_t\}$ follow a stationary Gaussian process and the observable process $\{Y_t\}$ modeled by equation (11). Then, the upper and lower control limit for the process sample mean based on the observable sample is given by,

$$\text{UCL/LCL} = \mu \pm \text{K}\sqrt{Var\left(\bar{Y}\right)} \tag{22}$$

where $\text{K} > 0$ is a real positive constant, often taken as $\text{K} = 3$ and $Var\left(\bar{Y}\right)$ is given by equation (13). In particular, when the process is an AR(1) process

$$\text{UCL/LCL} = \mu \pm \text{K}\sigma_a\sqrt{R_1^e\left(n,\phi,\tau_a\right)} \tag{23}$$

where $R_1^e\left(n,\phi,\tau_a\right)$ is given by equation (15). As measurement errors increase, the value of $R_1^e\left(n,\phi,\tau_a\right)$ also increases, widening the control limits. The probability P of detecting the mean shift from in-control mean μ_0 to out-of-control mean μ_1 is equal to

$$\text{P} = \Phi\left(-K - \delta/\sqrt{R_1^e\left(n,\phi,\tau_a\right)}\right) + \Phi\left(-K + \delta/\sqrt{R_1^e\left(n,\phi,\tau_a\right)}\right) \tag{24}$$

and the average run length (ARL) of the control chart is equal to

$$\text{ARL} = 1/\text{P} \tag{25}$$

where $\delta = |(\mu_0 - \mu_1)/\sigma_a|$ measures the mean shift. From Figure 4 and Figure 5 it can be noticed that as autocorrelation and measurement error increases, $R_1^e\left(n,\phi,\tau_a\right)$ increases. As a result, the power of the control chart decreases and the ARL increases. It can be visualized from Figure 10. Therefore, the performance of the control chart decreases in the presence of both autocorrelation and measurement errors. Some techniques are used to reduce the autocorrelation and measurement error. For example, serial autocorrelation can be reduced by using the s-skip strategy and measurement error can be reduced by taking several measures $(m \geq 1)$ of each observed item or by improving the gauge performance, see, for example, Costa and Castagliola (2011), Shongwe et al. (2021), Shongwe et al. (2019), Shongwe and Malela-Majika (2021), Garza-Venegas et al. (2018).

6. Conclusions

Many industrial process data are autocorrelated and at the same time, the existence of measurement errors due to inadequate measuring devices is common. To estimate some inferential results based on the collected sample, sometimes the mean and variance of the sample are required to be estimated. Hence, statistical properties of the sample mean and the sample variance are required to ascertain how reliable the estimated values are.

Our discussion in Section 2 indicates that the variance of the sample mean and the sample variance increases as autocorrelation increases. Also, the sample variance is biased

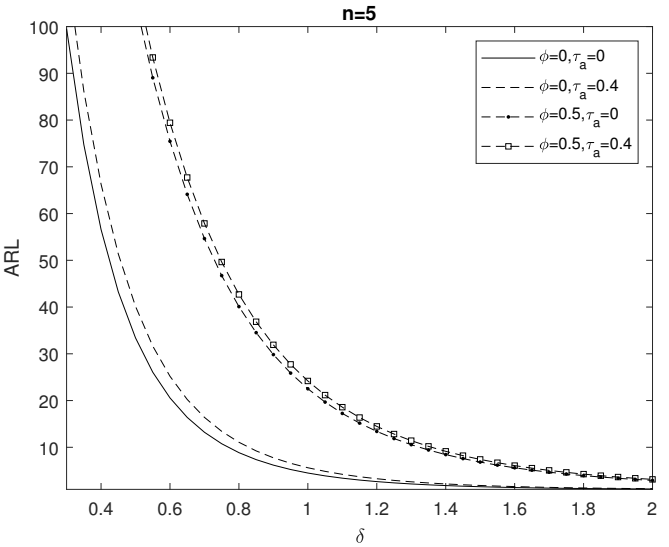


Figure 10: The effect of autocorrelation and measurement error on the ARL of the mean control chart, for K $= 3$ and $n = 5$.

and underestimated in the presence of autocorrelation. It has been shown that the variance of the sample mean and the sample variance is significantly large when autocorrelation is high and the sample size is small. Therefore, a relatively large sample size is recommended for estimating mean and variance depending on the level of autocorrelation.

Similarly, our analysis in Section 3 shows that the combined effect of autocorrelation and measurement error is found in estimating mean and variance for autocorrelated samples contaminated by measurement errors. Measurement error increases the variance of the sample variance. Sample variance is underestimated in the presence of autocorrelation and overestimated in the presence of measurement error. Therefore, one should be very careful when taking measurements and measuring devices should also be good enough to provide adequate confidence. In Section 4 we compare theoretical values of mean and variance of the estimator of the sample mean and the sample variance with the sample estimated value on the basis of simulated data. These results show that the theoretical values based on our obtained results and the estimated sample values are reasonably close. Here, we have considered a particular case of a stationary Gaussian process, namely an $AR(1)$ process. But there are more stationary Gaussian processes other than the $AR(1)$ process. Also, in this paper, we have assumed that the measurement error follows an independent Gaussian distribution but there are some situations where the measurement error does not follow Gaussian distribution. These situations can be considered in future research.

Disclosure statement

There is no potential conflict of interest involved with this submission.

Acknowledgements

Thanks to anonymous reviewers whose suggestions helped to improve the paper.

References

Abay, K. A., Wossen, T., Abate, G. T., Stevenson, J. R., Michelson, H., & Barrett, C. B., (2023). Inferential and behavioral implications of measurement error in agricultural data. *Annual Review of Resource Economics*, 15(1), pp. 63–83.

Blackwell, M., Honaker, J., & King, G., (2017). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research*, 46(3), pp. 303–341.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M., (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Brockwell, P. J., Davis, R. A. (Eds.), (2002). *Introduction to time series and forecasting*. New York, NY: Springer New York.

Carroll, R. J., (1998). Measurement error in epidemiologic studies. *Encyclopedia of biostatistics*, 3, pp. 2491–2519.

Costa, A. F., Castagliola, P., (2011). Effect of measurement error and autocorrelation on the X chart. *Journal of Applied Statistics*, 38(4), pp. 661–673.

Garza-Venegas, J. A., Tercero-Gómez, V. G., Lee Ho, L., Castagliola, P., & Celano, G., (2018). Effect of autocorrelation estimators on the performance of the X control chart. *Journal of Statistical Computation and Simulation*, 88(13), pp. 2612–2630.

Kotz, S., Balakrishnan, N., & Johnson, N. L., (2019). *Continuous multivariate distributions, Volume 1: Models and applications* (Vol. 334). John Wiley & Sons.

Koutsoyiannis, A., (1977). Theory of econometrics: an introductory exposition of econometric methods. (No Title).

Linna, K. W., Woodall, W. H., (2001). Effect of measurement error on Shewhart control charts. *Journal of Quality technology*, 33(2), pp. 213–222.

Maleki, M. R., Amiri, A., & Castagliola, P., (2017). Measurement errors in statistical process monitoring: A literature review. *Computers & Industrial Engineering*, 103, pp. 316–329.

Noor-ul-Amin, M., Javaid, A., Hanif, M., & Dogu, E., (2022). Performance of maximum EWMA control chart in the presence of measurement error using auxiliary information. *Communications in Statistics-Simulation and Computation*, 51(9), pp. 5482–5506.

Runger, G. C., Willemain, T. R., (1995). Model-based and model-free control of autocorrelated processes. *Journal of Quality Technology*, 27(4), pp. 283–292.

Schennach, S. M., (2016). Recent advances in the measurement error literature. *Annual Review of Economics*, 8(1), pp. 341–377.

Shongwe, S. C., Malela-Majika, J. C., (2021). A new double sampling scheme to monitor the process mean of autocorrelated observations using an AR (1) model with a skip sampling strategy. *Computers & Industrial Engineering*, 153, p. 107084.

Shongwe, S. C., Malela-Majika, J. C., & Castagliola, P., (2021). A combined mixed-s-skip sampling strategy to reduce the effect of autocorrelation on the X scheme with and without measurement errors. *Journal of Applied Statistics*, 48(7), pp. 1243–1268.

Shongwe, S. C., Malela-Majika, J. C., & Molahloe, T., (2019). One-sided runs-rules schemes to monitor autocorrelated time series data using a first-order autoregressive model with skip sampling strategies. *Quality and Reliability Engineering International*, 35(6), pp. 1973–1997.

Shumway, R. H., Stoffer, D. S., (2017). *Time series analysis and its applications: with R examples*. Springer

Wu, C. W., (2011). Using a novel approach to assess process performance in the presence of measurement errors. *Journal of Statistical Computation and Simulation*, 81(3), pp. 301–314.

Zhang, N. F., (1998). Estimating process capability indexes for autocorrelated data. *Journal of Applied Statistics*, 25(4), pp. 559–574.

Appendix

Derivation of $V\left(S_Y^2\right)$

Here, the observable process is $\{Y_t\}$ where $Y_t = X_t + E_t$, $E_t \sim IID\, N(0, \sigma_E^2)$. Sample variance is,

$$\begin{aligned} S_Y^2 &= \frac{1}{(n-1)} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \\ &= S_X^2 + \frac{2}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})(E_i - \bar{E}) + S_E^2 \end{aligned} \tag{26}$$

where,

$$S_E^2 = \frac{1}{n-1} \sum_{i=1}^{n} (E_i - \bar{E})^2, \quad \bar{E} = \frac{1}{n} \sum_{i=1}^{n} E_i.$$

Therefore,

$$Var(S_Y^2) = Var(S_X^2) + \frac{4}{(n-1)^2} Var\left(\sum_{i=1}^{n} (X_i - \bar{X})(E_i - \bar{E}) \right) + Var(S_E^2). \tag{27}$$

Note that other covariance terms in Equation (27) will become zero as X_t, E_t are independent. From Zhang (1998) we get,

$$Var(S_X^2) = \frac{2\sigma_X^4}{(n-1)^2} F(n, \phi). \tag{28}$$

Also, we have,

$$Var(S_E^2) = \frac{2\sigma_E^4}{(n-1)}. \tag{29}$$

Now,

$$
\begin{aligned}
& Var\left(\sum_{i=1}^{n}(X_i-\bar{X})(E_i-\bar{E})\right) \\
&= \mathbf{E}\left(\sum_{i=1}^{n}(X_i-\bar{X})(E_i-\bar{E})\right)^2 \\
&= \mathbf{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\bar{X})(X_j-\bar{X})(E_i-\bar{E})(E_j-\bar{E})\right) \\
&= \sum_{i=1}^{n}\mathbf{E}(X_i-\bar{X})^2\mathbf{E}(E_i-\bar{E})^2+\frac{\sigma_E^2}{n}\sum_{i=1}^{n}\mathbf{E}(X_i-\bar{X})^2 \\
&\quad -\frac{\sigma_E^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\left((X_i-\bar{X})(X_j-\bar{X})\right) \\
&= \sigma_E^2\sum_{i=1}^{n}\mathbf{E}(X_i-\bar{X})^2 \\
&= (n-1)\sigma_E^2\sigma_X^2 f(n,\rho_i).
\end{aligned}
\tag{30}
$$

Substituting Equations (28), (29) and (30) in Equation (27) we finally get,

$$
Var(S_Y^2) = \frac{2\sigma_X^4}{(n-1)^2}\left\{F(n,\phi)+(n-1)\frac{\sigma_E^4}{\sigma_X^4}+2(n-1)f(n,\rho_i)\frac{\sigma_E^2}{\sigma_X^2}\right\}. \tag{31}
$$

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 157–178, https://doi.org/10.59139/stattrans-2024-008
Received – 16.04.2023; accepted – 06.08.2024

A method of estimating the Return on Housing Investment (ROHI)

Arkadiusz J. Derkacz[1]

Abstract

The aim of the study was to develop a method for estimating the profitability of housing investments. Market practice shows that the profitability of this type of investment is influenced by specific determinants that are absent in the classical approach to profitability analysis. The most commonly used method is the Return on Equity (ROE) ratio, which is dedicated to enterprises. However, housing investments are becoming increasingly popular among individuals, while the classical ROE method proved suboptimal for such ventures (i.e. those involving the purchase of residential property and its subsequent rental to third parties). In this context, we made an attempt to develop a method that would make the estimation of the profitability level of this type of investment possible. Through the decomposition of the ROE ratio, a model for the Return on Housing Investment (ROHI) was created. This model was verified using real market data. Ultimately, we found that the ROHI method allows the estimation of the profitability level while taking into consideration the most important determinants characteristic of this type of investment.

Key words: multiple equation models, multiple variables, verification of the ROE ratio, apartment rentals, Return on Housing Investment.

1. Introduction

Housing investments are becoming an increasingly popular form of capital investment for both real estate sector companies and private individuals (Christophers, 2022; Krulický & Horák, 2019). However, housing investments should be understood as the purchase of a residential property that is intended for rental to third parties (Davis & Van Nieuwerburgh, 2015). This is, of course, the most general definition of this form of investment. It is also possible for an investor to acquire a property in another way, such as through inheritance or donation (primarily applicable to private individuals). It is increasingly common for development companies to allocate some of their apartments specifically for rental rather than sale (Antczak-Stępniak,

[1] Institute of Management and Quality Sciences, University of Kalisz, Kalisz, Poland & National Bank of Poland, Poland. E-mail: a.derkacz@uniwersytetkaliski.edu.pl. ORCID: https://orcid.org/0000-0003-1363-9551.

2019; Tomal, 2017). Of course, there are various ways to get ownership rights to a property, but what is most important is that housing investments require the investor to make an investment decision involving risk in the real estate market. By owning a specific property, the investor rents it out to others. Each such investment requires some investment outlay (greater or lesser), but each investment is made to generate optimal income. As a result, every investor is interested in achieving the highest possible return on their investment.

Currently, there is no longer any doubt that the rental housing market is subject to the phenomenon of financialization (Gadowska-dos Santos, 2018; Piętka, 2023). The 'ideology of homeownership', which has been researched and described by R. Ronald, is considered being the root cause of this phenomenon. According to this concept, residential real estate went from being a good for satisfying housing needs to an investment good (Ronald, 2008). Thus, this type of good has become less and less a social function and more and more an asset to investment portfolios (Fernandez & Aalbers, 2017). In this context, it is important to state that housing-centered financialization is a phenomenon that involves the increasing role – and even dominance – of financial actors, markets, valuations and mechanisms that result in the real transformation of entire economies, companies and, perhaps most importantly, households. In this paper, we will not deal with the aspect of short-term renting (Simcock, 2023) treating only the aspect of investment for housing. In this context, a distinction must also be made between professional/institutional investors and private investors. In both cases, the purpose of the investment is the same, which is to rent one's own flat for a specific profit. However, there are more differences. Their common denominator seems to be the way these two groups of investors operate in the rental market (Byrne, 2020). An obvious difference is also the fact that professional investors are companies with the right knowledge, experience, and tools to optimize housing investments. Private investors entered the residential rental market because of the 'generational annuity' (Aalbers et al., 2021). The development of private housing investment has also been influenced by the privatization processes of public housing, as shown by G. Wijburg and M. Aalbers in the example of Germany (2017). Recent years also reveal a trend of increasing interest in the purchase of housing, which is regarded by private individuals as a form of capital investment. This makes, among other things, the question of the profitability of investments in the residential property market an important argument for making appropriate decisions. For private individuals, such investments are characterized by high capital intensity. Not infrequently, in order to realize their investments, they decide to take out a loan, which increases the real costs of the investment.

In this context, it should be stated that one of the most important indicators is the return on equity (ROE) for residential developments. It is an indicator whose main function is to reveal the efficiency of the business, which is based primarily on the principle of economy and rationality of operation (Almagribi et al., 2023; Dufrénot,

2023). From the perspective of professional or institutional investors, this knowledge is almost taken for granted. However, the situation is different among the Polish public. Here, economic and financial knowledge is at a very low level. Research results confirm that most of the population lacks knowledge of proper management of the household budget and sensible planning of their expenses (Koćwin, 2021). The low level of economic knowledge may cause suboptimal functioning in a market economy (Szczechowiak, 2020). This state of affairs may therefore affect Poles who become private investors in the residential property market. The low level of economic knowledge may cause their decisions being marked by bounded rationality (Lejarraga & Pindard-Lejarraga, 2020). These limitations are, in this case, the result of incomplete information and suboptimal data in investment management (de Clippel & Rozen, 2021). This can cause the expected profitability of housing investments to diverge significantly from the actual profitability. The housing investment profitability calculators that are available on the Internet are also not helpful in addressing this problem. This is one of the most important sources where individuals seek economic knowledge (Bączar et al., 2024). In residential investment, it should be noted that ready-made profitability calculators can be found on the Internet. In the table below, we have presented the range of variables that determine - according to the authors of the individual calculators - the level of profitability of the invested capital. Because of this experiment, it can be concluded that the estimated levels of profitability of investments in rental housing are very simplified. However, these are publicly available tools that most often provide private investors with knowledge in this area.

Table 1: List of selected residential investment calculators available on the Internet and variables determining the rate of return

Sources/Internet address of calculators	Rental income	Fees/administrative rent	Monthly total/operating costs	Purchase price of the flat	Loan costs	Renovation/finishing costs	Furnishing costs	PCC tax	Flat tax 8.5%	real estate agency commission	Notary fees	Annual rental occupancy rate
realtytools.pl	X	-	X	X	-	X	-	X	-	X	X	X
eportal-nieruchomosci.pl	X	-	X	X	-	-	-	-	-	-	-	-
listaprzetargow.pl	X	X		-	-	X	X	X	X	X	-	X
calkoo.com	X	-	X	X	X	-	-	-	-	-	-	X
planhipoteczny.pl	X	X		X	X	X	-	-	X	X	-	X
infakt.pl	X	-	X	-	-	-	-	-	-	-	-	-
ftathub.pl	X	-	X	X	-	X	-	-	-	-	-	-
invest.rentujemy.pl	X	-	X	-	-	-	-	-	X	-	-	-

Source: own study.

A key issue arises for the presented research - Return on Investment. If investments are made by companies, the situation seems straightforward. Every development company, deciding on investment rental of its own apartments, is based on detailed financial analyses (Laszek & Olszewski, 2015). Financial departments are created for this purpose, which estimates the optimal level of the ROE ratio. Analysis of the rental housing segment now suggests some observations. Is the classical method of estimating the ROE ratio optimal for the profitability of housing investments? The second question seems even more important. Does this method allow for the analysis of the profitability of this type of investment by private individuals? These questions became the basis of the main research aim. In this context, an attempt was made to verify the classical method of estimating Return on Equity and adapt it to housing investments. This verification aims to optimally adapt the profitability estimation method to market realities (Konowalczuk, 2018). This method should be useful also – or perhaps primarily – for private investors who do not have their own specialized financial departments.

Working on developing an optimal method for estimating the Return of Housing Investments, the availability of tools for this type of analysis was investigated. Currently, such instruments are mostly sought on the Internet. It turned out that practically these types of advice boil down to analyzing the relationship between profit and investment expenses or the value of the property. This means that the profitability of housing investments relies on calculating the Return on Investment (ROI) for the investor – which is the classic approach to the ROE ratio. The fundamental rules of financial analysis, of course, in terms of the Return on Equity ratio, allow us to answer the question – what profits does the company generate from the capital invested/equity in its operations? Thus, the Return on equity is calculated as the ratio of net profit to equity (Ichsani & Suhardi, 2015). The ROE ratio calculated in this way reveals how much profit the company's capital has generated. Of course, the method for estimating the ROE ratio presented here is very general. There are many sophisticated methods for calculating and interpreting levels of ROE in a company. It is possible to analyses the dynamics of changes in individual profitability determinants, relate net profits to changes in equity levels, or compare changes in profitability in different reporting periods. There are also methods that enable a very detailed decomposition of the ROE ratio, such as the DuPont decomposition model. For this study, a detailed analysis of these methods, which are used to manage the company, is not required. However, the fundamental assumptions of the ROE ratio method relate to the profits and equity of the company. In some tools dedicated to housing investments, the investment return ratio also appears (Gertsekovich et al., 2019). The most significant difference is that, in this case, the analysis suggests examining the level of net profits in relation to investment expenses. From the perspective of corporate finance management, this distinction seems quite obvious. However, the situation is slightly different when an

individual becomes an investor and the investment involves buying an apartment and renting it out to third parties.

Here, the key question regarding the level of profitability should be related to the realities of the real estate market (Gołąbeska, 2017; Haran et al., 2016) and adjusted to the level of usefulness of private investors. Therefore, a critical analysis of the profitability of housing investments should provide an answer to a slightly different question. What is the level of return on investment based on renting an apartment, even for a private investor who does not have their own financial analysis department? This means that the tool for analyzing this type of profitability should also have an educational character, which will enlighten private investors about the optimally wide range of factors that affect the final level of profitability of their investment. The method for estimating the profitability of housing investments should consider those phenomena that directly determine the level of costs and revenues of this type of investment. The statement that the profitability of this type of investment depends on the purchase price of the apartment and the level of rental income is at least suboptimal. It is regrettable that these types of approaches often appear on the Internet as "guides" for investors in the rental segment of apartments in Poland.

Other methods of assessing profitability are also used in financial management, including in housing market operators. What they have in common is that they are intended to inform the soundness of the housing investment. It is sufficient to mention just a few of them here. These methods are often based on a basic cash flow model and Net Operating Income (NOI) calculation, which is the difference between operating income and operating expenses. The Gross Rent Multiplier (GRM) method is based on this. According to the principles adopted here, the value of the investment is estimated by multiplying the projected potential rental income from the property annually by the gross rent multiplier value determined by the investor (Schmidt, 2014). Another method is the so-called direct capitalization. In this case, the investment result is calculated by dividing the Net Operating Income (NOI) by the capitalization rate (Cap Rate – CR). This one is usually determined by the investor (investment value – IRV) or is based on market value. However, the capitalization rate is not the same as the yield (YIELD) or return on investment (ROI). It is also worth mentioning the Cash-on-Cash Return Equity (ROI) method. For residential property investments, it is most often calculated by dividing the pre-tax cash flow forecast by the initial outlay. The value in the numerator most often refers to the first year of the investment. The initial outlay, on the other hand, determines the investor's share without external financing, such as an investment or housing loan (Crosby et al., 2020; Sherman et al., 1933). Another method worth mentioning here is the Internal Rate of Return (IRR). It reports the percentage return on the money invested over each of the investment periods. It is primarily based on the initial outlay, periodic flows, and sales income. The Net Present Value (NPV) method can also be used to evaluate residential investments. This can be

assumed to be the value of the investment, which is the sum of all cash flows minus the initial outlay. However, all these values are discounted values. An even more extensive method of analyzing an investment is the Capital Accumulation Comparison Method (CACM). Its main objective is to identify the investment that will provide the maximum level of accumulation of invested capital in the future. Each of these methods has its advantages and disadvantages. Some are simple, others are more complex. Some are based on a few elements, others are multifactorial (Brealey et al., 2014; Brueggeman & Fisher, 2018). The choice of one method is sometimes determined by external factors or non-financial arguments of investors. Although this is not the purpose of this article, a hypothesis can be made on this basis. We suppose that it would be very valuable if research could consist of comparing different methods of estimating the value/profitability/justification for making investments in the residential property market. This opens up a new avenue for future research. In the research presented here, the return on equity ratio and its decomposition, according to the Du Pont concept, were chosen as the base method.

These observations have led to an attempt to verify and adjust the Return on Equity (ROE) indicator method to the real conditions prevailing in the rental housing market. Previous studies by real estate market experts reveal a huge diversity of factors influencing phenomena occurring in the housing and rental housing market segments (Laszek et al., 2021; Renigier-Bilozor et al., 2017). This causes defining a specific set of determinants for this segment, which also affects the profitability of housing investments (Wójcik, 2016). This attempt is significant because the results and analyses can create a tool that will optimize decision-making processes for private investors in this segment. This attempt is expected to optimize scientific research on the profitability of such investments. Market research exists that estimates profitability levels based on limited determinants and many theoretical assumptions. To some extent, this is because of difficulties in accessing statistical data. However, there are factors that can be estimated but are not considered in these studies. This attempt can also allow for more detailed and optimal scientific research in this area. These observations became the main determinants for conducting research and analysis that ultimately led to the definition of the Return on Housing Investments (ROHI) indicator.

2. Method

The classical approach to calculating the ROE indicator assumes a dependency:

$$ROE_t = \frac{NP_t}{CE_t} \times 100 \tag{1}$$

where:
NP_t – net profit, CE_t – capital expenditures.

This method assumes that investment expenditures consist only and only of equity capital. In the situation where a portion of investment expenditures comes from external financing sources, the value of CE should be adjusted by a coefficient of $1 - LTV$; $LTV \in \{0,1\}$, where LTV denotes the percentage of external capital involvement in financing investment expenditures. Aspects characteristic of housing investments are introduced into the classical method of estimating the ROE indicator. Therefore, from this point on, the term "Return on Housing Investment – ROHI" will be used. In economic reality, expenses related to this type of investment do not concern only the cost of purchasing an apartment (PD). Additional costs arise for finishing the apartment, its renovation, or adaptation for final use (PFD). The first group of costs usually applies to apartments in a developer state. The following types of costs concern apartments from the secondary market. The process of implementing residential investments requires other costs as well. Therefore, initial costs (IC) were also introduced into the analysis of the ROHI indicator, which may include costs such as a notary or commission for an intermediary. The cost of insuring the apartment (IP) should also be included in the analysis. However, these types of costs are usually financed with equity capital. Nonetheless, they constitute real costs that should be treated as part of investment expenditures. In this context, it is also necessary to introduce a value that is significant from the perspective of financing investments. This is the equity contribution, which is often a condition for obtaining external financing (MLC). At the same time, this is capital that can be treated as a potential source of income from alternative investments.

From this, a portion of total investment outlays (CE) is the sum of capital associated with financing a housing loan ($CEFL$) and capital not related to external financing (CE''). It should be emphasized that this division is not equivalent to the division into internal and external capital. This is because $CEFL$ is the sum of the loan value (LV) and the amount of equity (MLC). The latter value will be referred to as $MLC = CEFL \times (1 - LTV)$. Therefore, the value of equity (CE') allocated to housing investment should be determined as:

$$\begin{aligned} CE'_t = MLC_t + CE''_t = CEFL_t \times (1 - LTV) + IC_t + IP_t = \\ = (PD_t + PFD_t) \times (1 - LTV_t) + IC_t + IP_t \end{aligned} \tag{2}$$

The value of CE'_t determined in this way will represent the total investment expenditures that have been covered by the investor's own capital. This also means that the total value of external financing is $LV = (PD_t + PFD_t) \times LTV_t$, and the amount of own contribution to the loan is $MLC = CEFL - LV$. Therefore, equation (2) can be written as:

$$CE'_t = MLC_t + IC_t + IP_t \tag{3}$$

The final form of equation (1) for residential investments will be:

$$ROHI_t = \frac{NP_t}{MLC_t + IC_t + IP_t} \times 100 \tag{4}$$

In the analysis, we can move on to decomposing the net profit value (NP). In the most general form, it can be written as:

$$NP_t = GP_t - TAX_t \tag{5}$$

where:

GP – gross profit, TAX – value of taxes.

Next, it is necessary to analyze the detailed factors determining the value of gross income ($GP_t = EBIT_t - FC_t$), which is the difference between the value of operating profit ($EBIT$) and financial costs (FC). This can be presented as a set of successive detailed dependencies:

$$\begin{gathered} EBIT_t = EBITDA_t - DV_t \\ DV_t = RIV_t \times \%dv_t \\ EBITDA_t = NOI_t - HOC_t \\ NOI_t = EGI_t - OC_{t_t}^- \\ EGI_t = PGI_t \times \%rei_t \\ PGI_t = GRI_t + OC_{t_t}^+ \end{gathered} \tag{6}$$

where:

$EBITDA$ – operating profit before deducting interest on borrowed interest-bearing liabilities, taxes, amortization; DV – the value of amortization; RIV – the value of the residential investment subject to amortization; NOI – net operating income; HOC – housing operating costs; EGI – effective gross income; OC^+/OC^- – operational/ administrative costs as payable (+) or investor's obligation (−); PGI – potential gross income; $\%rei$ – rental efficiency ratio; GRI – gross rental income.

A comment on the introduced variable of housing operating costs is required. This variable replaced the value of the company's operating costs, which appear in the classical decomposition of ROE. This is because the profitability of residential investments usually concerns individual investors, and it is difficult to speak in this context about the costs of a company. However, it seems reasonable to introduce such a cost item that will reflect the result more realistically. Based on this, the value of operating profit before tax ($EBIT$) can be decomposed as follows:

$$\begin{gathered} EBIT_t = EBITDA_t - DV_t = EBITDA_t - RIV_t \times \%dv_t = NOI_t - HOC_t - RIV_t \times \%dv_t = \\ = EGI_t - OC_{t_t}^- - HOC_t - RIV_t \times \%dv_t = PGI_t \times \%rei_t - OC_{t_t}^- - HOC_t - RIV_t \times \%dv_t = \\ = \left(GRI_t + OC_{t_t}^+\right) \times \%rei_t - OC_{t_t}^- - HOC_t - RIV_t \times \%dv_t \end{gathered} \tag{7}$$

where:

$\%dv$ – depreciation rate per year, RIV – initial value of assets subject to depreciation.

After simplifying equation (7), we get the following form:

$$EBIT_t = GRI_t \times \%rei_t + OC_{t_t}^+ \times \%rei_t - OC_{t_t}^- - HOC_t - RIV_t \times \%dv_t \tag{8}$$

The next value to be decomposed concerns financial costs (FC). In simplified calculations of Return on Housing Investments, the cost of a housing loan is assumed in this case. However, in reality, this aspect is not so simple. The cost of a mortgage loan, which is intended for the purchase of a property (LTV_t^{PD}), is obviously present. There are also situations where the buyers of a property obtain external capital for the renovation or finishing of a rental property (LTV_t^{PFD}). One should also not forget about the so-called own contribution, which also generates certain costs (LTV_t^{MLC}). On the one hand, these may be costs of equity. On the other hand, there are situations where equity capital also comes from external sources of financing. In this context, it is assumed that investment expenditures financed by a loan are equal to:

$$CEFL_t = PD_t + PFD_t + MLC_t = LV_t + MLC_t \tag{9}$$

This implies that the share of external capital in financing a housing investment should be defined as:

$$LTV = LTV_t^{PD} + LTV_t^{PFD} + LTV_t^{MLC} = 1 \tag{10}$$

The above equation requires some explanation. This concerns the situation when investors get external financing from several loans. That $LTV = 1$ should therefore refer to the total amount of loans. The value of LTV_t^{PD} in equation (10) will therefore be the share of the value of the loan intended for the purchase of the apartment in relation to the total debt of the investor. Similarly, the value of LTV_t^{PFD} should be interpreted. The value of LTV_t^{MLC} will determine the share of equity in the total debt amount. In this context, the decomposition of the financing costs of the housing investment can be continued.

$$\begin{gathered} FC_t = CEFL_t \times \%lv_t = (PD_t + PFD_t + MLC_t) \times LTV \times \%lv_t = \\ = (PD_t \times LTV_t^{PD} \times \%lv_t^{PD}) + (PFD_t \times LTV_t^{PFD} \times \%lv_t^{PFD}) + (MLC_t \times LTV_t^{MLC} \times \%lv_t^{MLC}) \end{gathered} \tag{11}$$

where:

$\%lv$ – the interest rate of the loan or the cost of own contribution.

Based on equations (8) and (11), the gross revenue (GP) can be expressed as follows:

$$\begin{gathered} GP_t = GRI_t \times \%rei_t + OC_{t_t}^{+} \times (\%rei_t - 1) - HOC_t - RIV_t \times \%dv_t - \\ -(PD_t \times LTV_t^{PD} \times \%lv_t^{PD} + PFD_t \times LTV_t^{PFD} \times \%lv_t^{PFD} + MLC_t \times LTV_t^{MLC} \times \%lv_t^{MLC}) \end{gathered} \tag{12}$$

At this stage of the analysis, it is possible to move on to a detailed decomposition of equation (5) in terms of tax values (TAX). Assuming that $TAX_t = NOI_t \times \%tax_t$, where $\%tax$ represents the tax rate, and using some of the dependencies from equations (6), the decomposition of tax values takes the following form:

$$\begin{gathered} TAX_t = \left(EGI_t - OC_{t_t}^{-}\right) \times \%tax_t = \left(PGI_t \times \%rei_t - OC_{t_t}^{-}\right) \times \%tax_t = \\ = \left(\left(GRI_t + OC_{t_t}^{+}\right) \times \%rei_t - OC_{t_t}^{-}\right) \times \%tax_t = \\ = \left(GRI_t \times \%rei_t + OC_{t_t}^{+} \times \%rei_t - OC_{t_t}^{-}\right) \times \%tax_t = \\ = (GRI_t \times \%rei_t + OC_{t_t}^{+} \times \%rei_t - OC_{t_t}^{-}) \times \%tax_t \end{gathered} \tag{13}$$

Using equations (2), (12), and (13), it is possible to determine the value of $ROHI$. Equation (4) takes the form:

$$\begin{aligned} ROHI_t = {} & GRI_t \times \%rei_t + OC_t^+ \times \%rei_t - OC_t^- - HC_t - RIV_t \times \%dv_t \\ & - PD_t \times LTV_t^{PD} \times \%lv_t^{PD} - PFD_t \times LTV_t^{PFD} \times \%lv_t^{PFD} \\ & - MLC_t \times LTV_t^{MLC} \times \%lv_t^{MLC} - (GRI_t \times \%rei_t \\ & + OC_t \times (\%rei_t - 1)) \times \%tax_t \\ & \times \frac{1}{(PD_t + PFD_t) \times (1 - LTV_t) + IC_t + IP_t} \times 100 \end{aligned} \tag{14}$$

To simplify the above equation, the following relationships can be introduced:

- actual gross rental income: $GRI_t^r = GRI_t \times \%rei_t$;
- actual operating costs: $OC_t^b = OC_{t_t}^+ \times \%rei_t - OC_{t_t}^-$;
- actual depreciation: $RIV_t^r = RIV_t \times \%dv_t$;
- actual gross rental income tax: $GRI_t^{tax} = GRI_t \times \%rei_t \times \%tax_t$;
- actual operating costs tax: $OC_t^{tax} = (OC_t^+ \times \%rei_t - \mathrm{OC}_t^-) \times \%tax_t$;
- actual costs of financing the purchase of the housing: $FC_t^{PD} = PD_t \times LTV_t^{PD} \times \%lv_t^{PD}$;
- actual costs of financing the finishing/renovation of the housing: $FC_t^{PFD} = PFD_t \times LTV_t^{PFD} \times \%lv_t^{PFD}$;
- actual costs of financing the own contribution: $FC_t^{MLC} = MLC_t \times LTV_t^{MLC} \times \%lv_t^{MLC}$.

Based on this, equation (14) can be written as:

$$ROHI_t = \frac{GRI_t^r - GRI_t^{tax} + OC_t^r - OC_t^{tax} - HOC_t - RIV_t^r - FC_t^{PD} - FC_t^{PFD} - FC_t^{MLC}}{(PD_t + PFD_t) \times (1 - LTV_t) + IC_t + IP_t} \times 100 \tag{15}$$

Further simplification of the above equation is possible. Two new dependencies have been introduced:

- $GRI_t^r - GRI_t^{tax} = GRI_t \times \%rei_t - GRI_t \times \%rei_t \times \%tax_t = GRI_t \times \%rei_t \times (1 - \%tax_t) = GRI_t^r \times (1 - \%tax_t)$;
- $OC_t^r - OC_t^{tax} = OC_{t_t}^+ \times \%rei_t - OC_{t_t}^- - (OC_{t_t}^+ \times \%rei_t - OC_{t_t}^-) \times \%tax_t = (OC_{t_t}^+ \times \%rei_t - OC_{t_t}^-) \times (1 - \%tax_t) = OC_t^b \times (1 - \%tax_t)$

Based on the above, it is possible to present the final form of the equation that determines the value of $ROHI$:

$$ROHI_t = \frac{GRI_t^r \times (1 - \%tax_t) + OC_t^b \times (1 - \%tax_t) - HOC_t - RIV_t^r - FC_t^{PD} - FC_t^{PFD} - FC_t^{MLC}}{(PD_t + PFD_t) \times (1 - LTV_t) + IC_t + IP_t} \times 100 \tag{16}$$

In this way, an equation has been derived that determines the level of Return on Housing Investments, taking into account optimal factors drawn from the reality of the real estate market.

3. Research

The proposed method of calculating the Return on Housing Investment in residential properties was verified using statistical data from the Polish economy. To this end, the following dataset was used:

- the average transaction rate per 1 sqm of purchased apartment on the secondary market in Q4 2022, source: NBP;
- interest rate on new housing loans, on average from 2022, source: NBP;
- interest rate on new consumer loans, on average from 2022, source: NBP;
- notarial fees according to the real estate purchase cost calculator[2];
- apartment insurance costs according to the adopted criteria[3];
- average rental prices for apartments with an area of 40−59 sqm in February 2023, source: Otodom Analytics.

Additionally, the following assumptions were adopted for the calculation:

- apartment size of 50 sqm;
- share of the loan for purchasing the apartment in three options, LTV=0%, LTV=60%, and LTV=80%;
- costs of finishing/renovating the apartment based on city groups:
  - Warsaw – 3000 PLN/year;
  - group of 5 cities[4] – 2500 PLN/year;
  - group of 10 cities[5] – 2000 PLN/year;
- costs of maintaining the apartment based on city groups:
  - Warsaw – 1500 PLN/year;
  - group of 5 cities – 1200 PLN/year;
  - group of 10 cities – 1000 PLN/year;
- cost of equity – 0% (excluded from the calculation);
- tax rate on rental income – 8.5% annually;
- number of months of renting the apartment – 12 months;
- value of operating costs based on city groups:
  - Warsaw – 1200 PLN/year;
  - group of 5 cities – 1000 PLN/year;
  - group of 10 cities − 800 PLN/year;
- apartment depreciation coefficient – 1.5% annually.

Based on the above, the level of Return on Housing Investments was estimated for individual voivodeship cities in Poland using equation (16). Calculations were performed for three options based on the level of involvement of a housing loan (see Figures 1–3).

[2] https://www.bankier.pl/narzedzia/kupno-nieruchomosci

[3] https://www.ubezpieczeniemieszkania.pl

[4] The group of 5 cities included Gdańsk, Kraków, Łódź, Poznań and Wrocław.

[5] The group of 10 cities included Białystok, Bydgoszcz, Katowice, Kielce, Lublin, Olsztyn, Opole, Rzeszów, Szczecin and Zielona Góra.

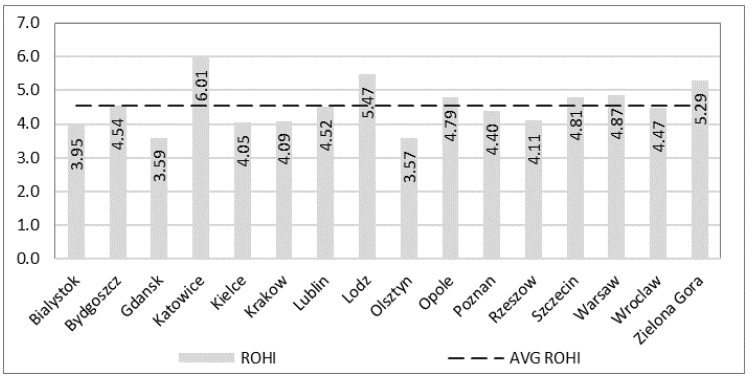


Figure 1: Levels of Return on Housing Investment by cities in %, LTV=0%.

Source: own study.

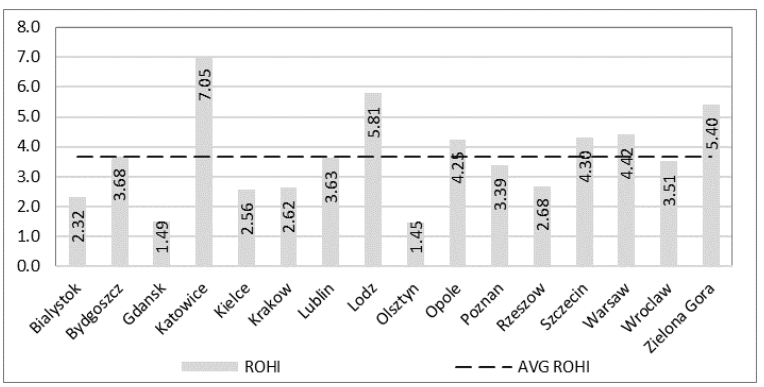


Figure 2: Levels of Return on Housing Investment by cities in %, LTV=60%.

Source: own study.

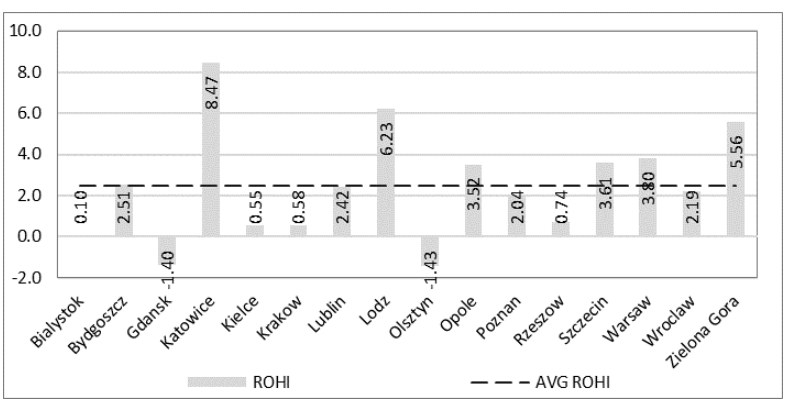


Figure 3: Levels of Return on Housing Investment by cities in %, LTV=80%.

Source: own study.

The presented method for estimating the Return on Housing Investments also allows for diagnosing the strength of the impact of individual determinants that appear on the right-hand side of equation (16). For this comparison, the expression from the denominator of the equation (16), $(PD_t + PFD_t) \times (1 - LTV_t)$, was adopted as the value of MLC_t in accordance with equations (2) and (3). In addition, independent variables were standardized within individual voivodeship cities. This made it possible to compare the impact of 10 determinants of the Return on Housing Investments under specified assumptions (see Figures 4–6). It should be emphasized that MLC_t values were plotted on the right-hand axis.

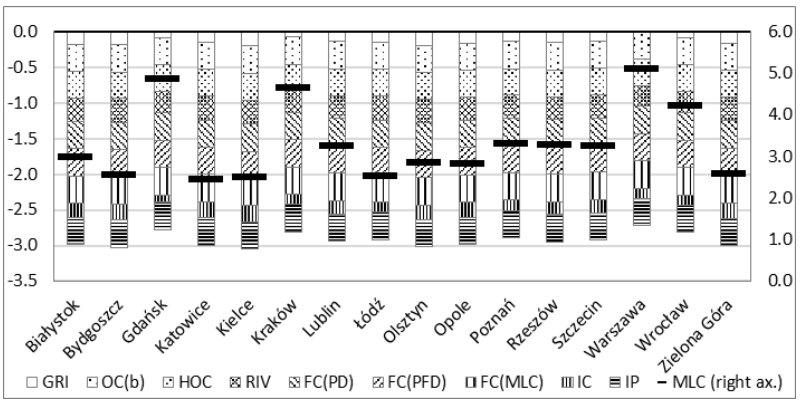


Figure 4: Strength of the determinants' impact on the level of Return on Housing Investments (ROHI) by cities in %, LTV=0%.

Source: own study.

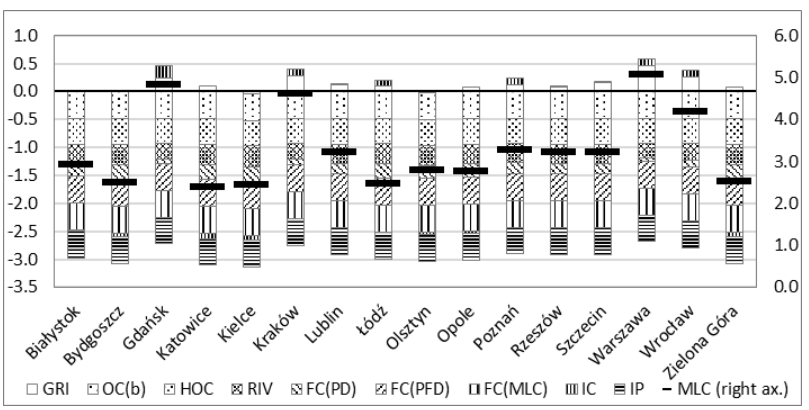


Figure 5: Strength of the determinants' impact on the level of Return on Housing Investments (ROHI) by cities in %, LTV=60%.

Source: own study.

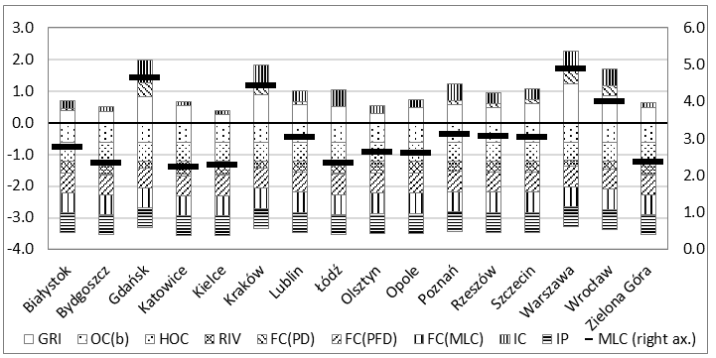


Figure 6: Strength of the determinants' impact on the level of Return on Housing Investments (ROHI) by cities in %, LTV=80%.

Source: own study.

The scope of the research should not only be limited to measurements using the new ROHI index. It is also important to compare these results with measurements according to the simplified method. This will reveal discrepancies in profitability results that may mislead investors. For this purpose, the profitability of the investment was estimated by rejecting the individual components that were entered the HRAI index model. One by one, the items that were rejected were (1) the costs of finishing/ renovating the flat, (2) the costs of insuring the flat, (3) the costs of operating the flat, (4) the taxation of rental income and (5) depreciation. Finally, a level of profitability was achieved, which was referred to as simplified profitability. The following three figures show the results. It turned out that the simplified profitability is on average approximately 10% higher than the profitability estimated according to the ROHI method. This means that the real income for tenants per year is several thousand zloty lower than that calculated according to the simplified profitability. These differences, especially for individual tenants, are significant from the perspective of their budgets.

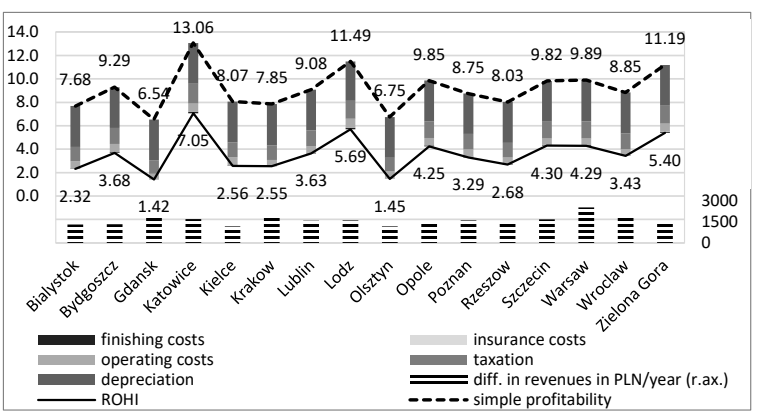


Figure 7: Comparison of ROHI index values with simple profitability by city in %, LTV=0%.

Source: own study.

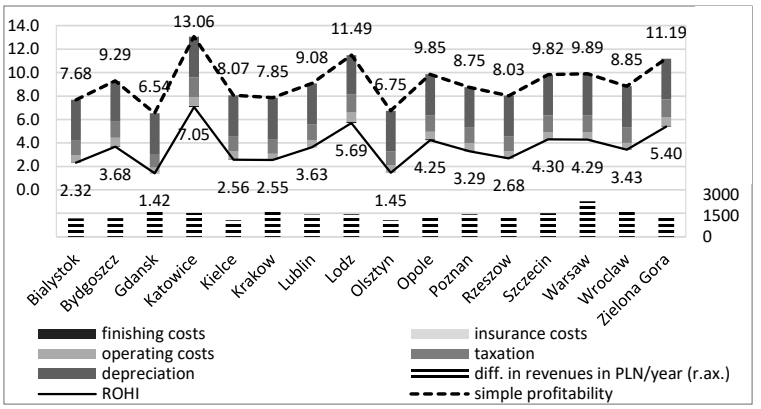


Figure 8: Comparison of ROHI index values with simple profitability by city in %, LTV=60%.
Source: own study.

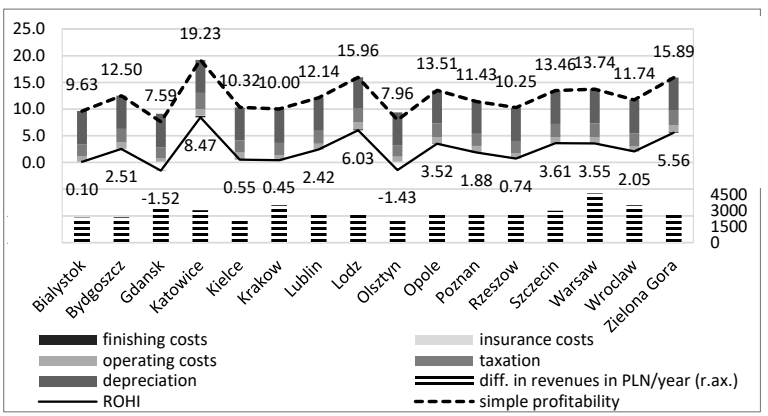


Figure 9: Comparison of ROHI index values with simple profitability by city in %, LTV=80%.
Source: own study.

4. Discussion

The results of the decomposition of the $ROHI$ index should be discussed. This requires, first and foremost, the interpretation of equation (16) along with the disclosed determinants. The value in the denominator determines the level of equity capital involvement in the implementation of the housing investment. Some of this capital is treated as a contribution to the loans taken out, while the remaining part is investment expenditures solely derived from own capital. In this context, the value in the denominator of equation (16) can be expressed in the following form:

$$\frac{1}{CE_t'} = \frac{1}{(PD_t + PFD_t) \times (1 - LTV_t) + IC_t + IP_t} \tag{17}$$

This shows that an increase in external capital involvement (LTV_t) results in a decrease in equity capital involvement (CE_t'), which appears to be a favorable relationship. An increase in investment expenditures in the IC_t and IP_t areas directly increases the investor's equity capital involvement, reducing the value of $ROHI$ index.

Equation (16) also requires interpretation in terms of the values that have emerged in its numerator. Here, seven factors appear that affect the final value of $ROHI$ in different ways, some of which are corrected by the tax rate ($\%tax_t$). First, the actual net rental income value needs to be described. It takes the form of $NRI_t^r = GRI_t^r \times (1 - \%tax_t)$, which represents the actual rental income adjusted for the rental efficiency ratio and tax costs.

The second factor increasing the level of ROHI is the real value of net operating costs, which can be expressed as $NOC_t^b = OC_t^b \times (1 - \%tax_t)$. This is the value of operating/administrative costs also adjusted for the rental efficiency ratio and tax costs. However, this variable requires further explanation. Operating costs are defined as costs that arise in the relationship between the owner of the apartment and the community or cooperative housing administration. In the case of renting an investment apartment, this type of cost is usually passed on to the tenant. However, two options may arise, which are revealed in the rental efficiency ratio. If the apartment is rented for the whole year, the $\%rei$ ratio equals 1. This means that the value of $OC_t^b = 0$ and thus $NOC_t^b = 0$. However, a different situation arises when the investment apartment is rented for, for example, 10 months a year. This means that the $\%rei$ ratio equals $10/12 = 0.83$. This means that the real value of operating costs becomes negative. This results in the owner of the investment apartment being obliged to cover them from their own capital. The occurrence of an additional operating/administrative cost directly and negatively affects the level of investment profitability. It should also be emphasized that in reality, the $\%rei$ ratio can take values from 0 to 1.

In this context, it is important to describe a situation where the owner of a rental property charges the tenant for operating costs (OC) higher than their obligations to the community administration or property manager. If the landlord generates a real value of balanced operating costs ($OC_t^r > 0$), it will increase real income. This follows from the decomposition of the $EBIT$ indicator (formula 8). This value will increase the taxable income base (formula 13). This way, a real value of balanced net operating costs will appear, which will increase the value of the $ROHI$ index.

The remaining values in the numerator of equation (16) seem to be significantly less complicated. It is characteristic that each of them is preceded by a minus sign, showing that an increase in their value would cause a decrease in the level of Return on Housing Investments. In the above considerations, equation (16) can be written in its final form:

$$ROHI_t = \frac{NRI_t^r + NOC_t^b - HOC_t - RIV_t^r - FC_t^{PD} - FC_t^{PFD} - FC_t^{MLC}}{(PD_t + PFD_t) \times (1 - LTV_t) + IC_t + IP_t} \times 100 \qquad (18)$$

The results of calculating the Return on Housing Investments according to the proposed method generate two main benefits. First, it should be emphasized that the estimated ROHI ratio provides an optimal picture of profitability due to a wide range of interacting factors. The presented research results clearly show that financial costs are only a part of the factors that affect the level of profitability of this type of investment. These costs are usually the only ones taken into account in simplified calculators of the profitability of housing investments. Second, the unquestionable advantage of using this type of method is that a detailed analysis of the factors affecting the final level of Return on Housing Investments is possible. Such knowledge at the stage of planning or investment decision-making can be a source of additional benefits for investors and investment optimization. These are among the elements that have sparked interest in the business of the real estate sector in the proposed method of estimating the ROHI ratio. Currently, the implementation of the described method in the form of a profitability calculator for housing investments, which will be publicly available on websites, is planned.

The proposed method also has its limitations. Firstly, it can be mentioned that ROHI is a static measure. This means that it does not take into account the change in the time value of money. Although this is a factor that is often analyzed in the investment decision-making process, to present the ROHI method, this approach has been abandoned. This is due to two assumptions. On the one hand, the future value can be estimated in a simplified manner based on an annual interest rate. In such situations, the present value can also be estimated based on the discount rate and the effective interest rate. These issues, although statistically and economically sound, give rather general results. To calculate the time value of money in a very precise manner, one would have to take into account many factors that influence this value. Suffice it to mention macroeconomic factors, i.e. inflationary processes or currency exchange rate changes, and microeconomic factors, i.e. investors' consumption preferences, risks or opportunity costs. These types of factors are numerous. It does not seem reasonable to include them in the proposed method. This would make the ROHI method very complicated. It should be emphasized at this point that the method was primarily developed as a decision support tool for individual investors and not for specialized think tanks.

Another limitation of the presented method is the accrual perspective used. This leaves the cash flow perspective, which is relevant to the cash method, out of consideration. This limitation, however, is the consequence of having to choose one of the possible options. Of course, it is possible to create two ROHIs separately for each financial perspective. However, would such a solution generate additional added value? At the same time, it should be made clear that the accrual method and the cash method have their advantages and disadvantages. The main advantage of the second method is

that it shows the result based on real money flows. Another advantage often mentioned by accounting experts is that it is simpler than the accrual method. However, the most important disadvantage of the cash method is its limitation in the long-term analysis of the financial health of a business. This may be the main argument for choosing the accrual method. This method provides the opportunity for this type of analysis. In addition, the accrual method reveals a broad picture of assets and liabilities, provides an analytical view of formal accounting records and is favored in International Accounting Standards (IAS). This kind of argumentation can give rise to a very wide debate on the subject. Which method is better and in which situations should they be used? For the development of the ROHI method, no such discussions were held. It was decided that return on equity and the Du Pont decomposition would be the foundation. This makes the accrual method a far better choice here.

As the issue of return on investment is very broad and multifaceted, many other limitations of this method can be written about. Certainly, there may be criticism that the estimation of the return on investment does not take into account the possibility of reselling the flat in the future, which will increase the capital inflow for the owner. Full agreement. There will certainly also be proponents of the claim that the sale of investment flats is no longer an investment in itself, but a mere disposal of fixed assets. It is a mere sale and not a rental investment. Such behavior is closer to the profit-making activity of 'buying cheaper and reselling more expensively'. In addition, it can be noted that the ROHI is estimated for a specific time horizon and does not take into account rolling profitability, albeit on a compound interest basis. Furthermore, the proposed method does not take into account alternative/potential profits and costs, the investor's income issues and the investor's creditworthiness. According to it, the anchoring aspect and the investor's perspective are also not taken into account. Is, for example, 15% profitability a lot or a little? For which investor is this the optimal level? Does the answer to this question lie in the perspective of, for example, the investor's income or perhaps in economies of scale? If an individual investor owns only one flat, for example, he expects a 15% yield. If he already owns five flats, 5% of each might be enough for him. Discussions on this and similar topics can be held for a very long time. One thing is certain. It is not possible to create a one-size-fits-all ROHI that will include 'the whole world' on the 'right side of the equation'. The proposed ROHI method was intended to allow individual investors to estimate their residential investments more optimally. More optimal is one that takes into account a larger catalogue of factors than the simple models we wrote about earlier. On the other hand, it cannot be a method that, with its above-average complexity, becomes inaccessible to this group of investors. This is one of the reasons why it was decided that return on equity and the Du Pont decomposition method would be its substantive foundation.

The method of calculating the ROHI indicator also allows for examining changes in the profitability of residential real estate investments in the long run. However, the results of such research were not presented in this paper, as it was not the aim of the analyses. The proposed method can fill an analytical gap or optimize existing research on the profitability of this type of investment. By optimization, we mean expanding the catalog of significant determinants affecting changes in the ROHI level. This may cause result in such research more realistically showing the levels of profitability of housing investments. The proposed method can create a profitability calculator for housing investments. From the investor's point of view, such an analysis has the potential to reveal various factors that ultimately affect the level of investment profitability. This, in turn, can enable more conscious investment decision-making.

The proposed method for estimating the ROHI indicator can also facilitate further research on the rental housing segment. By analyzing the identified factors, which are recorded as independent variables, it can be assumed that they "hide" institutions that affect the level of Return on Housing Investments. It seems that research into social, economic, and legal-political institutions could be a source of new valuable insights from the perspective of economics, private investors, or even the legislature, which co-shapes the environment of such investment instruments in the real economy.

Acknowledgement

Special thanks to Marcin Krason, Business Growth Manager of Otodom Analytics, and Katarzyna Kuniewicz, Head of Research of Otodom Analytics, for the opportunity to have a substantive discussion on the proposed method for estimating the ROHI indicator.

References

Aalbers, M. B., Hochstenbach, C., Bosma, J. and Fernandez, R., (2021). The Death and Life of Private Landlordism: How Financialized Homeownership Gave Birth to the Buy-To-Let Market. *Housing, Theory and Society*, *38*(5), pp. 541–563. https://doi.org/10.1080/14036096.2020.1846610.

Almagribi, M. K., Lukviarman, N. and Setiany, E., (2023). Financial Determinants of Corporate Cash Holdings: Evidence from Property and Real Estate Companies in Indonesia. *Review of Integrative Business and Economics Research*, *12*(3), pp. 251–260.

Antczak-Stępniak, A., (2019). *Determinants of Development Activity in the Polish Housing Market*. Wydawnictwo Uniwersytetu Łódzkiego.

Bączar, A., Dąbrowska, A., Polak, M. and Sekścińska, K., (2024). *Poziom wiedzy finansowej Polaków 2024. VII*.

Byrne, M., (2020). Generation rent and the financialization of housing: A comparative exploration of the growth of the private rental sector in Ireland, the UK and Spain. *Housing Studies*, *35*(4), pp. 743–765. https://doi.org/10.1080/02673037.2019.1632813.

Christophers, B., (2022). Mind the rent gap: Blackstone, housing investment and the reordering of urban rent surfaces. *Urban Studies*, *59*(4), pp. 698–716.

Davis, M. A., Van Nieuwerburgh, S., (2015). Housing, Finance, and the Macroeconomy. In G. Duranton, J. V. Henderson, & W. C. Strange (Eds.). *Handbook of Regional and Urban Economics,* Vol. 5A, pp. 753–811. Elsevier. https://doi.org/10.1016/B978-0-444-59531-7.00012-0.

de Clippel, G., Rozen, K., (2021). Bounded rationality and limited data sets. *Theoretical Economics*, *16*(2), pp. 359–380. https://doi.org/10.3982/TE4070.

Dufrénot, G., (2023). Interest Rates, Financial Markets, and Macroeconomics. In G. Dufrénot, New Challenges for Macroeconomic Policies. S*pringer International Publishing*, pp. 195–257. https://doi.org/10.1007/978-3-031-15754-7_5.

Fernandez, R., Aalbers, M. B., (2017). Housing and Capital in the Twenty-first Century: Realigning Housing Studies and Political Economy. *Housing, Theory and Society*, *34*(2), pp. 151–158. https://doi.org/10.1080/14036096.2017.1293379.

Gadowska-dos Santos, D., (2018). Istota i skala finansjalizacji rynku nieruchomości mieszkaniowych w Polsce w latach 2000–2015. *Studia i Prace Kolegium Zarządzania i Finansów*, *165*, pp. 111–140. https://doi.org/10.33119/SIP.2018.165.7.

Gertsekovich, D., Gorbachevskaya, L., Grigorova, L. and Peshkov, V., (2019). Return on investment in REIT real estate funds. *IOP Conference Series: Materials Science and Engineering*, *667*(1), 012025. https://doi.org/10.1088/1757-899X/667/1/012025.

Gołąbeska, E., (2017). Współczesne trendy na rynku nieruchomości mieszkaniowych. In E. Broniewicz (Ed.), Gospodarowanie przestrzenią w warunkach rozwoju zrównoważonego. *Oficyna Wydawnicza Politechniki Białostockiej,* pp. 85–106.

Haran, M., McCord, M., Davis, P., McCord, J., Lauder, C. and Newell, G., (2016). European emerging real estate markets: Re-examining investment attributes and framing opportunities. *Journal of Property Investment & Finance*, *34*(1), pp. 27–50. https://doi.org/10.1108/JPIF-04-2015-0024.

Ichsani, S., Suhardi, A. R., (2015). The Effect of Return on Equity (ROE) and Return on Investment (ROI) on Trading Volume. *Procedia – Social and Behavioral Sciences*, *211*(25), 896–902. https://doi.org/10.1016/j.sbspro.2015.11.118.

Koćwin, J., (2021). Ethics in the financial services market. Knowledge, consumer education and morality. *Wrocławsko-Lwowskie Zeszyty Prawnicze*, *12*(1), pp. 37–52.

Konowalczuk, J., (2018). Households Investments in the Residential Real Estate Market. *Świat Nieruchomości*, *1*(103), pp. 21–28.

Krulický, T., Horák, J., (2019). Real estate as an investment asset. *SHS Web of Conferences*, *61*(010011). https://doi.org/10.1051/shsconf/20196101011.

Laszek, J., Augustyniak, H. and Olszewski, K., (2021). The development of the rental market in Poland. In Real Estate at Exposure. New Challenges, Old Problems, *Oficyna Wydawnicza SGH*, pp. 263–274.

Laszek, J., Olszewski, K., (2015). The behaviour of housing developers and aggregate housing supply. *NBP Working Paper*, *206*. http://dx.doi.org/10.2139/ssrn.2642756.

Lejarraga, J., Pindard-Lejarraga, M., (2020). Bounded Rationality: Cognitive Limitations or Adaptation to the Environment? The Implications of Ecological Rationality for Management Learning. *Academy of Management Learning & Education*, *19*(3), pp. 289–306. https://doi.org/10.5465/amle.2019.0189.

Piętka, A., (2023). Residential private rental market in Poland – prospects and challenges. *Studia BAS*, *4*, pp. 9–25.

Renigier-Bilozor, M., Wisniewski, R. and Bilozor, A., (2017). Rating attributes toolkit for the residential property market. *International Journal of Strategic Property Management*, *21*(3), 307–317. https://doi.org/10.3846/1648715X.2016.1270235.

Ronald, R., (2008). *The ideology of home ownership: Homeowner societies and the role of housing*. Palgrave Macmillan. https://books.google.com/books?hl=pl&lr=&id=gCuHDAAAQBAJ&oi=fnd&pg=PR1&dq=R.+Ronald,+The+Ideology+of+Home+Ownership.+Homeowner+Societies+and+the+Role+of+Housing,+Houndmills,+Palgrave+Macmillan,+UK+2008&ots=h73qvHBfjv&sig=5psTs31ue8t3Y2xJ0Ve6cKz25kg.

Simcock, T., (2023). Home or hotel? A contemporary challenge in the use of housing stock. *Housing Studies*, *38*(9), pp. 1760–1776. https://doi.org/10.1080/02673037.2021.1988063.

Szczechowiak, E., (2020). Economic education and the entrepreneurial attitudes of students. Report on research on saving and spending money. In W. Truszkowski (Ed.), *Uwarunkowania budowy bezpieczeństwa prawnego, ekonomicznego i społecznego w Polsce* (p. 89), Uniwersytet Warmińsko-Mazurski w Olsztynie.

Tomal, M., (2017). Identifying the factors that significantly influence decisions to make housing development investments in the communes of Małopolska province. *Folia Pomeranae Universitatis Technologiae Stetinensis*, *88*(337), pp. 67–76.

Wijburg, G., Aalbers, M. B., (2017). The alternative financialization of the German housing market. *Housing Studies*, *32*(7), pp. 968–989. https://doi.org/10.1080/02673037.2017.1291917.

Wójcik, A., (2016). Residential Real Estate as an Investment Instrument. *Annales Universitatis Mariae Curie-Skłodowska*, *L*(3), pp. 195–203.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 179–189, https://doi.org/10.59139/stattrans-2024-009
Received – 22.01.2024; accepted – 16.08.2024

Estimation of the Cox model with grouped lifetimes

Piotr Bolesław Nowak[1]

Abstract

This paper presents how random numbers can be used to transform grouped lifetimes into a pseudo-complete sample. The aim of the study is to investigate the Fisher consistency of the partial likelihood estimator of the regression parameters in the Cox model based on the restored sample. It has been proven that for elliptical-type distributional assumptions about explanatory variables the estimators of the regression parameters in the Cox model based on the pseudo-complete sample are consistent up to a scaling factor. A simulation study illustrates the asymptotic properties of the estimates. In addition, real data case analysis is presented.

Key words: Cox model, grouped data, Fisher consistency, elliptical distribution.

1. Introduction

Let T be a random variable denoting survival time and $X = (X_1, \ldots, X_p)^\top$ be a vector of covariates having cumulative distribution function H. The Cox proportional hazard model is a common technique for analysis of censored survival data which assumes that the hazard function of time t, given the covariate value $X = x$ is of the form

$$\lambda(t|x) = \lambda_0(t)\exp(\beta^\top x),$$

where $\lambda_0(t)$ is the baseline hazard function and $\beta \in \mathbb{R}^p$ denotes unknown regression parameters. It implies that the conditional survival function of T given $X = x$ takes the form $S(t|x) = P(T > t|x) = \exp\left(-\Lambda(t)\exp(\beta^\top x)\right)$, where $\Lambda(t) = \int_0^t \lambda_0(s)ds$ is the baseline cumulative hazard function.

Given a random sample $\{(T_i \wedge C_i, (X_{i1}, \ldots, X_{ip}), \delta_i)\}_{i=1}^n$, where $\delta_i = 1(T_i \leq C_i)$ and the censoring variable C is independent of T given the value of $X = x$, Cox (1972) introduced a method of estimating β without considering Λ, which is known as the partial likelihood method. The partial likelihood estimator for the Cox model solves the equation

$$\int \left[y - \frac{\int 1(t \wedge c \geq w) x \exp(\beta^\top x) dF_n(t,c,x)}{\int 1(t \wedge c \geq w)\exp(\beta^\top x) dF_n(t,c,x)} \right] 1(w \leq a)\, dF_n(w,a,y) = 0, \qquad (1)$$

where $F_n(t,c,x)$ denotes the empirical distribution function of the random sample and 1 denotes the indicator function.

[1]Institute of Economic Sciences, University of Wrocław, Wrocław, Poland. E-mail: piotr.nowak2@uwr.edu.pl. ORCID: https://orcid.org/0000-0002-7404-2946.

Assume now that time is partitioned into k intervals $A_j = [a_{j-1}, a_j)$, $j = 1, \ldots, k$ and $a_0 = 0$, $a_k = \infty$. For each individual the exact value of X is known but the underlying variable T is unobserved due to grouping mechanism. We only know in what interval each individual died or was censored.

In practice, it is impossible to measure time with infinite precision. For instance, in constructing life tables age is rounded to the nearest year. Moreover, sample elements are often classified into disjoint subsets, like intervals, rectangles, etc. It means that it is not possible to give individual sample values but only the numbers of observations in each specified class (for more examples see Haitovsky (1982)). Inference methods for the grouped survival data can be found in Kalbfleich and Prentice (1973), Thompson (1977), Prentice and Gloeckler (1978). Kalbfleich and Prentice (1973) obtained a generalized linear model with a complementary log-log link function while Thompson (1977) used the logistic model. A comprehensive lecture on discrete hazard models can be found in Fahrmeir and Tutz (2001), in particular, see Chapter 9 for the methods for modelling of discrete survival data. McKeague and Zhang (1996) obtained a Sheppard correction for grouping in the Cox model.

The aim of this paper is to present a different approach from the one mentioned above to estimate the conditional survival function, which we describe in the next section. In the sequel we show that estimating equation (1) can be used for inference about β even when lifetimes coming from the Cox model are grouped into intervals. In the final section, we present simulation study concerning scale Fisher consistency of the proposed estimators and give examples with a real data set.

2. The estimator of the parameters for the grouped Cox model

Since the presented considerations hold also in the case of censoring, in order to simplify notations, we first consider the case without censoring.

Recall, that for grouped data instead of the sample $\{(T_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$ we observe $\{(z_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$, where z_i is a $1 \times k$ vector indicating the grouping interval. Thus, $\sum_{i=1}^n z_i = (n_1, \ldots, n_k)$, where n_i is the total number of deaths in the ith interval.

The estimation of the distribution parameters based on the grouped data is often more difficult than for ungrouped data. For data divided into intervals the most straightforward approach to estimation is to assume that all observations within each finite interval are assigned to its midpoint.

The presented method of estimation in the case of the grouped data is based on the idea that an unobserved lifetime T for given $X = x$ in the interval $A_j = [a_{j-1}, a_j)$ may be replaced by a random variable $\tilde{T}$ generated independently according to some distribution on this set with cumulative distribution function (cdf), namely G_j. Therefore, instead of sample $\{(T_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$ we have $\{(\tilde{T}_i, (X_{i1}, \ldots, X_{ip})\}_{i=1}^n$ and hence the estimating equation (1) can be applied. Throughout this paper we will call this sample the *pseudo-complete* sample generated by the grouped Cox model. The term of the pseudo-complete sample was also used by Whitten et al. (1988) for the restoration of incomplete samples, but their method was applied only to censored samples.

Observe that the density of the random variable $[\tilde{T}|X = x]$ is given by the formula $\tilde{f}(t|x) = \sum_{i=1}^k g_i(t) 1_{A_i}(t) P(T \in A_i|x)$, where g_i is the density function over the set A_i. Now,

denote the conditional survival function of this distribution by $\tilde{S}(t|x)$. From the above description, we conclude that

$$\tilde{S}(t|t \in A_j, x) = P(\tilde{T} > t | t \in A_j, x) = S(a_j|x) + [1 - G_j(t)][S(a_{j-1}|x) - S(a_j|x)].$$

The uniform distribution over $[a_{j-1}, a_j)$ is the most natural choice of G_j for each finite A_j. It corresponds to the piecewise linear approximation of the survival function S. For the last set A_k, it is reasonable to consider shifted exponential distribution or distribution of random variable with probability one at the point a_k. When $1 - G_j(t) = (a + h - t)/h$ is the survival function of the uniform distribution over $[a, a+h)$ and if the interval length h approaches 0, then we have $\tilde{S}(t|t \in [a, a+h), x) \approx S(a|x) + (t-a)S'(a|x)$.

In the next chapter we prove that the described reconstruction of the sample leads to estimators which are consistent up to some positive scale, which is explained below.

3. Scaled Fisher consistency

In statistics, most estimators are defined as solutions to the estimating equations based on the empirical distribution. We say that the estimating equation is Fisher consistent at the model (or in short, the estimator being its solution is Fisher consistent) if the solution to this equation coincides with the true parameter when the empirical distribution is replaced by the true model distribution. For instance, Fisher consistency for the Cox model means that if F_n in (1) is substituted by a joint distribution of (T, C, X), where (T, X) is from the Cox's model with parameter β_0, then $\beta = \beta_0$ is its only solution. Proving Fisher consistency is a primary step in examining the asymptotic properties of M-estimators (see, e.g. Huber and Ronchetti (2009)). This notion was used by Bednarski (1993) in robust method of estimation of regression coefficients based on a modification of partial likelihood estimator.

The scaled Fisher consistency means that solutions to the estimating equation, if the empirical distributions are replaced by the true model distributions, are scaled regression parameters, i.e. $\beta = \alpha\beta_0$ for some scaling factor $\alpha > 0$.

The problem of scaled Fisher consistency for some regression models was considered by Ruud (1983) and Stoker (1986), among others. Another recent important account in such studies is due to Bednarski and Skolimowska-Kulig (2018), who showed that the maximum likelihood estimator for the regression parameters in the classical exponential regression model is scaled Fisher consistent for the extended model. Recently, Bednarski and Nowak (2021), Bednarski, Nowak and Skolimowska-Kulig (2022) have showed that in the Cox model with arbitrary frailty the partial likelihood estimator is also Fisher consistent up to a scaling factor under elliptic type distributional assumptions on explanatory variables.

For further considerations replace the empirical distribution function F_n in (1) by the joint distribution of $(\tilde{T}, X)$, i.e. $\tilde{F}_{\beta_0}(t,x) = \tilde{F}_{\beta_0}(t|x)H(x)$. We always use the subscript β_0 to emphasize that the distribution of (T, X) is under the true value of the parameter β. Thus,

equation (1) becomes

$$\sum_{j=1}^{k}\int_{A_j}\left[y-\frac{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)xe^{\beta^\top x}dH(x)}{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)e^{\beta^\top x}dH(x)}\right]d\tilde{F}_{\beta_0}(w,y)=0. \tag{2}$$

We have the following definition.

Definition 1. *The scaled Fisher consistency of the partial likelihood estimator of β in the Cox model based on a pseudo-complete sample means that equation (2) is satisfied for $\beta=\alpha\beta_0$, where $\alpha>0$ is some scaling factor.*

Remark 1. *Reduction of equation (2)*

Observe that equation (2) can be reduced with an assumption that $EX=0$.

Denoting by μ_0 the expectation of X and after performing some simple algebra this equation can be transformed as follows: $H(x)$ is replaced by $H(x+\mu_0)$ and $\Lambda(w)$ by $e^{\beta_0^\top\mu_0}\Lambda(w)$.

In the view of the above remark our aim is to show that $\tilde{L}(\beta,\beta_0)=0$ is satisfied for $\beta=\alpha\beta_0$, $\alpha>0$, where

$$\tilde{L}(\beta,\beta_0)=\sum_{j=1}^{k}\int_{A_j}\left[\frac{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)xe^{\beta^\top x}dH(x)}{\int \tilde{S}_{\beta_0}(w|w\in A_j,x)e^{\beta^\top x}dH(x)}\right]d\tilde{F}_{\beta_0}(w). \tag{3}$$

The main idea of proving scaled Fisher consistency is based on the construction of an auxiliary function $f_\beta:[0,\infty)\to\mathbb{R}$ defined as follows:

$$f_\beta(\alpha)=\sum_{j=1}^{k}\int_{A_j}\left[\frac{\int(\beta^\top x)\tilde{S}_\beta(w|w\in A_j,x)e^{\alpha\beta^\top x}dH(x)}{\int \tilde{S}_\beta(w|w\in A_j,x)e^{\alpha\beta^\top x}dH(x)}\right]d\tilde{F}_\beta(w). \tag{4}$$

The behavior of the function f_β is described in the following lemma. Its proof is omitted as it is similar to the proof of Lemma 3.1 in Bednarski and Nowak (2021).

Lemma 1. *For any β and any continuous $G_1,\ldots,G_k$ the function f_β has the following properties:*

1. *It is continuous and strictly increasing on $[0,\infty)$.*
2. *$f_\beta(0)<0$.*
3. *$\lim_{\alpha\to\infty} f_\beta(\alpha)>0$.*

Now, let us recall that a p-dimensional random vector X is spherically symmetric distributed if for every orthogonal matrix Γ of size p (i.e. $\Gamma\Gamma^\top=\Gamma^\top\Gamma=I$) the random vector ΓX is distributed as X. Then, the random vector $Y=\mu+AX$ is said to be elliptically symmetric distributed with parameters $\mu\in\mathbb{R}^p$ and covariance matrix Σ_Y, where $\Sigma_Y=AA^\top$. It is known that conditional expectation of Y given $\beta^\top Y=c$ is a linear function with respect to c. In fact, the following lemma can be proved (see also Bednarski and Nowak (2021)).

Lemma 2. *Let Y be a p-dimensional random vector which has an elliptically symmetric distribution with parameters $\mu \in \mathbb{R}^p$ and Σ_Y. Then, for any $\beta \in \mathbb{R}^p$ and any $c \in \mathbb{R}$ it holds*

$$E[Y|\beta^\top Y = c] = \mu + (c - \beta^\top \mu)\frac{\Sigma_Y \beta}{\beta^\top \Sigma_Y \beta}.$$

Now, we are ready to formulate the main theorem, which gives sufficient conditions for the scaled Fisher consistency when the partial likelihood estimator is used for the grouped Cox model based on the pseudo-complete sample.

Theorem 1. *Let the vector of explanatory variables $X = (X_1, \ldots, X_p)^\top$ be elliptically symmetric distributed. Then for any continuous distributions $G_1, \ldots, G_k$ the partial likelihood estimator for the grouped Cox based on the pseudo-complete sample is Fisher consistent up to a scale factor.*

Proof. We show that the equation $\tilde{L}(\alpha\beta_0, \beta_0) = 0$ is satisfied for some scaling factor $\alpha > 0$. Observe that an immediate conclusion from Lemma 1 is that there exists $\alpha_0 > 0$ such that $f_{\beta_0}(\alpha_0) = 0$. Putting $\beta = \alpha_0\beta_0$ we can write the inner integral from the numerator in (3) as the expectation $E(\tilde{S}_{\beta_0}(w|w \in A_j, X)Xe^{\alpha_0\beta_0^\top X})$. Conditioning it on $\beta_0^\top X$ and applying Lemma 2 for X with $\mu = 0$ we have

$$E(\tilde{S}_{\beta_0}(w|w \in A_j, X)Xe^{\alpha_0\beta_0^\top X}) =$$
$$E(E(\tilde{S}_{\beta_0}(w|w \in A_j, X)Xe^{\alpha_0\beta_0^\top X}|\beta_0^\top X)) = \frac{\Sigma_X\beta_0}{\beta_0^\top\Sigma_X\beta_0}E\left((\beta_0^\top X)\tilde{S}_{\beta_0}(w|w \in A_j, X)e^{\alpha_0\beta_0^\top X}\right).$$

Hence, $\tilde{L}(\alpha_0\beta_0, \beta_0) = \frac{\Sigma_X\beta_0}{\beta_0^\top\Sigma_X\beta_0}f_{\beta_0}(\alpha_0) = 0$, which ends the proof. □

Remark 2. *The presence of a censoring variable.*

In the case of the presence of a censoring variable we observe $(T_1 \wedge C_1, X_1, \delta_1), \ldots, (T_n \wedge C_n, X_n, \delta_n)$, where X denotes covariate vector and $\delta = 1(T \leq C)$. Let $F(t,c,x)$ denote the joint distribution of time T, censoring variable C and covariates X under the Cox model. Under the conditional independence of T and C given X one can factorize $dF_{\beta_0}(t,c,x) = dF_{\beta_0}(t|x)dC(c|x)dH(x)$. Now, we replace the random variable $T \wedge C$ by $\tilde{T}$ as follows: when $T \wedge C$ takes the values from A_j then $\tilde{T}$ follows the distribution on the set A_j with cdf G_j on this set. Thus, the pseudo-sample generated by the grouped Cox model consists of $(\tilde{T}_1, X_1, \delta_1), \ldots, (\tilde{T}_n, X_n, \delta_n)$. Then, the Fisher scaled consistency for the grouped Cox model based on the pseudo-complete sample means that the equation $L(\beta, \beta_0) = 0$ is satisfied for $\beta = \alpha\beta_0$, $\alpha > 0$, where

$$L(\beta, \beta_0) =$$
$$\sum_{j=1}^{k}\int_{A_j}\left[y - \frac{\int \tilde{S}_{\beta_0}(w|w \in A_j, x)[1 - C(w|x)]xe^{\beta^\top x}dH(x)}{\int \tilde{S}_{\beta_0}(w|w \in A_j, x)[1 - C(w|x)]e^{\beta^\top x}dH(x)}\right][1 - C(w|y)]d\tilde{F}_{\beta_0}(w|y)dH(y) = 0.$$

From the above it follows that Lemma 1 and Theorem 1 remain applicable in the presence of a censoring variable.

4. Numerical examples

This section presents computational examples for selected distributions and an application of presented method for real and simulation data.

Example 1. (Monte Carlo simulation)

A Monte Carlo experiment for 5000 runs was conducted to investigate properties of the partial likelihood estimation under the pseudo-complete sample when data were generated from the Cox model. The S-Plus programing language was used to generate lifetimes coming from the Cox model. For the true parameter value $\beta_0 = (1, -0.5, 0.5)^\top$, two types of cumulated baseline intensities, $\Lambda(t) = t^{1/2}$ and $\Lambda(t) = t^2$ were used. The vector X was either elliptically distributed with standard normal distributions or non-elliptically distributed with exponential marginals. The sample size was taken $n = 500$ and the grouping was performed for $k = 2, 5, 15, 20$. For each grouping the class $A_k = [a_k, \infty)$ was chosen so that $P_{\beta_0}(T > a_k) = 0.1$ and the group limits $0, a_1, \ldots, a_{k-1}$ were equidistant. After grouping of lifetimes the pseudo-complete samples were created. The uniform distribution on each finite interval and the shifted exponential distribution on the tail were applied. Table 1 shows the results of this experiment.

Table 1: Results of simulation experiment for true parameter $\beta_0 = (1, -0.5, 0.5)^\top$. The first vector in each cell refers to the means of ratios of components of estimates and the true parameters. The second one refers to the standard deviations of the vector estimates of true parameter values.

grouping	Regressors normally distributed		Regressors non-elliptically distributed	
	$\Lambda(t) = t^{1/2}$	$\Lambda(t) = t^2$	$\Lambda(t) = t^{1/2}$	$\Lambda(t) = t^2$
$k = 2$	(3.220, 3.231, 3.214) (0.066, 0.065, 0.065)	(3.105, 3.113, 3.116) (0.065, 0.064, 0.066)	(6.262, 2.382, 4.255) (0.045, 0.041, 0.042)	(6.288, 2.386, 4.257) (0.045, 0.041, 0.044)
$k = 5$	(2.056, 2.060, 2.053) (0.068, 0.066. 0.066)	(1.216, 1.213, 1.210) (0.075, 0.070, 0.069)	(3.500, 1.848, 2.492) (0.054, 0.041, 0.047)	(1.557, 1.420, 1.343) (0.090, 0.045, 0.056)
$k = 15$	(1.544, 1.542, 1.545) (0.073, 0.066, 0.066)	(1.089, 1.091, 1.088) (0.079, 0.072, 0.072)	(2.465, 1.624, 1.829) (0.067, 0.042, 0.051)	(1.098, 1.223, 1.073) (0.075, 0.047, 0.051)
$k = 20$	(1.462, 1.463, 1.468) (0.074, 0.069, 0.068)	(1.084, 1.087, 1.081) (0.077, 0.072, 0.071)	(2.292, 1.581, 1.732) (0.071, 0.042, 0.053)	(1.066, 1.203, 1.054) (0.068, 0.048, 0.050)

Simulations indicate good asymptotic performance of the estimator under normally distributed covariates. Note, that each elliptically distributed vector X can be chosen as a member of a large family of probability distributions like multivariate normal distributions, multivariate t-distributions, multivariate Logistic and Laplace distributions and many others. For this case components of the first vectors in each cell are almost the same, which

shows that we have the estimation of the regression parameter up to the same scaling factor. It is interesting that even for grouping for $k = 2$ we can estimate the regression parameter up to a scaling constant which is approximately equal to 3.2 and 3.1 for $\Lambda(t) = t^{1/2}$ and $\Lambda(t) = t^2$, respectively. The scaling factors decrease as the number of classes increase. For grouping with $k = 15, 20$ and $\Lambda(t) = t^2$ a scaling factor is near one, which corresponds to the consistent estimation of the regression parameter. On the other hand, we observe bad performance of the estimators under departure from the elliptical type distributional assumption of explanatory variables when the number of grouping classes is small, especially for $k = 2$. As the number of classes increase the estimators may approach to the true parameter despite non-elliptically distributed regressors, for instance, see the case for $k = 15$ and $\Lambda(t) = t^2$, where the estimation seems to be correct.

Example 2. (Life table)

Another example presented here compares two estimation methods for the life table for gender and race (see Table 2 based on article by Arias (2007)). These data were also considered by Agresti (2010) on page 127.

The first method of the estimation is applied in order to reconstruct the entire sample from the grouped sample using random numbers generated according the uniform distribution on each interval. We assumed that $A_7 = (95, 120)$, because according to the International Database on Longevity the longest-lived person ever form the United States died at the age of 119 years and 97 days, see also Kestenbaum and Ferguson (2010).

For two explanatory variables, gender g (1 = female; 0 = male) and race r (1 = black; 0 = white), the Cox model was fitted to the pseudo-complete sample of size 1000 for each of the four groups.

As a second model, we used the generalized linear model (GLM) with complementary log-log link function, i.e.

$$\log(-\log(1 - P(Y \leq j))) = \theta_j + \beta_1 g + \beta_2 r, \; j = 1, 2, \ldots, 6.$$

Table 2 contains fitted distributions, the first value in each parenthesis corresponds to the Cox model and the second value in each parenthesis to the GLM. For each of the four distributions and for each of the estimation methods, we calculated the dissimilarity index, which is the half the sum of absolute differences between the fitted and estimated population distributions. This index takes values (in percent) 2.2, 6.6, 7.2, 3.5 for the Cox model and 2.7, 6.8, 6.8, 3.3 for the GLM. The differences in estimates of two mentioned methods are very small.

Example 3. (Veteran data)

The next example compares the two estimation methods for the Veteran's Administration lung cancer data, see Kalbfleisch and Prentice (1980). This data set is frequently used to test different estimation. There were continuous covariates: Karnofsky rating, disease duration and age whereas binary ones are prior therapy (yes=1 or no=0), treatment (standard=1 or test=0) and four cell types (squamous, small, large and adeno). Because of colinearity of these cell types, we take into consideration in this model only three of them, namely squamous, small and adeno.

Table 2: Observed and fitted (in parentheses) life-length distributions of U.S. residents, as percentages. The first value in each parenthesis corresponds to the Cox model based on pseudo-complete sample, the second one to the GLM with complementary log-log link function.

Gender	Race	Life Length						
		0-20	20-40	40-50	50-65	65-80	80-95	over 95
Female	Black	1.8 (1.5, 1.5)	2.4 (2.6, 2.6)	3.7 (3.4, 3.3)	12.9 (12.6, 12.4)	29.6 (30.0, 29.9)	39.3 (40.8, 41.5)	10.3 (9.1, 8.8)
	White	0.9 (1.2, 1.2)	1.3 (2.2, 2.0)	1.9 (2.7, 2.6)	8.0 (10.4, 9.9)	25.9 (26.3, 25.3)	49.7 (43.1, 43.5)	12.3 (14.1, 15.5)
Male	Black	2.6 (2.1, 2.2)	4.9 (3.6, 3.8)	5.6 (4.6, 4.8)	20.2 (16.6, 17.3)	34.7 (35.2, 36.1)	27.8 (34.5, 33.2)	4.2 (3.4, 2.6)
	White	1.3 (1.7, 1.7)	2.8 (2.9, 2.9)	3.2 (3.7, 3.8)	12.2 (13.9, 14.0)	32.8 (32.1, 32.2)	42 (39.2, 39.3)	5.7 (6.5, 6.1)

In order to present the result for the pseudo-complete sample we grouped lifetimes into twenty equidistance classes, i.e. $k = 20$. The range of lifetime is 1–999. Each grouped lifetime was replaced by a random number according to the uniform distribution on the corresponding interval. Scaled values of estimates for complete and pseudo-complete sample are presented in Table 3. The differences in scaled estimates are very small, i.e. the maximum absolute difference is no more than 0.05.

Table 3: Comparison of partial likelihood estimation for complete and pseudo-complete sample for the Veteran's Administration lung cancer data.

	complete sample			pseudo-complete sample		
covariates	ple	scaled ple	p-value	ple	scaled ple	p-value
karnofsky	-0.0328	-0.0314	0.0000	-0.0327	-0.0299	0.0000
diag time	0.0001	0.0001	0.9929	-0.0040	-0.0036	0.6683
age	-0.0087	-0.0083	0.3492	-0.0015	-0.0014	0.7807
prior	0.0072	0.0068	0.7579	0.0018	0.0017	0.8328
squamous	-0.4013	-0.3839	0.1557	-0.4072	-0.3720	0.1574
small	0.4603	0.4403	0.0838	0.4404	0.4024	0.1105
adeno	0.7948	0.7604	0.0087	0.8498	0.7764	0.0076
tratment	0.2946	0.2818	0.1558	0.3390	0.3098	0.1131

Example 4. (Estimation in the Cox model with rounded data)

Let us recall that the Cox model is based on several restrictive assumptions and one of them says that there were no tied values among the observed survival times. When constructing a new partial-likelihood function we must assume that the roundings for particular survival times appear by imprecision in the measurements of survival time. Therefore, when we have d values rounded to the one value, in fact they could have been observed in any of the $d!$ possible orders. The exact form of the partial-likelihood function is obtained by modification of the partial-likelihood function to include all possible arrangements. Then, we get expressions inconvenient for further calculations, therefore we use approximations.

Approximation, both introduced by Breslow (1974) and Efron (1977), provides simpler expressions than an exact function, but still include effect of rounded data.

In order to apply randomization procedure to rounded data we replace each tied survival time, namely t, by a random variable according to uniform distribution on the interval $(t-\varepsilon, t+\varepsilon)$, $\varepsilon > 0$.

Now, we illustrate this randomization procedure by considering data on HMO examination of patients infected with HIV (see Hosmer and Lemeshow (1999)), where 100 patients participated in the study, with 31 different survival times. The number of people with the same time survival rates ranged between 1 and 17. For the simulation, we assumed that $\varepsilon = 0.5$.

Table 4: Comparison of estimation results for different estimation methods.

Method	AGE		DRUG	
	Coeff.	Sd.Err.	Coeff.	Sd.Err.
Exact	0.0977	0.0187	1.0226	0.2572
Breslow	0.0915	0.0185	0.9414	0.2555
Efron	0.0971	0.0186	1.0167	0.2562
Random	0.0976	0.0186	1.0307	0.2577

Table 4 shows estimation results for methods mentioned above. Note that using the randomization method (see the last row of Table 4), we get the results that are very close to the exact one. In fact, estimators calculated by all four methods are close to each other and their standard errors are almost identical.

5. Final conclusions

The Cox model is based on several restrictive assumptions and one of them assumes continuous survival time. This assumption may not be fulfilled in many situations, e.g. when the data are rounded and then at least two events may occur at one point in time. Another situation concerns the case when survival data are grouped. In general, data grouping is a frequently used data presentation mechanism in practical applications. There are well-known methods of inference based on grouped data in the statistical literature, but due to data compression, the resulting estimators may be less effective or more biased than those obtained on the basis of the full sample.

This paper presents a method of estimation of the regression parameters in the Cox model, when lifetimes are grouped into a set of intervals. We showed how using random numbers, which are easily available in statistical packages, one can obtain a reconstruction of a simple sample, called here a pseudo-complete sample, and hence the classic Cox estimator can be still used. We noticed that in case of using a uniform distribution on each grouping interval, the described randomization method corresponds to the approximation of the survival function by a piecewise linear function. We proved that for the pseudo-complete sample the partial likelihood method of estimation leads to the consistent estimation of regression parameters up to a scaling factor if covariates are elliptically distributed. By stan-

dard asymptotic argumentation it means that solutions to equation (1) for restored samples converge to scaled regression parameters as the sample increases and they are asymptotically normal.

The problem discussed in this paper is important because initial data are often aggregated and then classical methods based on the assumption of continuity of the dependent variable are limited. Therefore, the presented randomization method can also be used in other regression models, where a dependent variable is grouped.

Acknowledgements

The author thanks the referees for comments leading to important improvements in the paper.

References

Agresti, A., (2010). *Analysis of Ordinal Categorical Data, 2nd Edition*, John Wiley and Sons.

Arias, E., (2007). United States life tables, 2004. *National Vital Statistics Reports*, 56(9), pp. 1–40.

Bednarski, T., (1993). Robust estimation in Cox's regression model. *Scandinavian Journal of Statistics*, 20(3), pp. 213–225.

Bednarski, T., Nowak, P. B., (2021). Scaled Fisher consistency of partial likelihood estimator in the Cox model with arbitrary frailty. *Probability and Mathematical Statistics*, 41(1), pp. 77–87.

Bednarski, T., Nowak, P. B., Skolimowska-Kulig, M., (2022). Scaled Fisher consistency for the partial likelihood estimation in various extensions of the Cox model. *Statistics in Transition new series*, 23(2), pp. 185–196.

Bednarski, T., Skolimowska-Kulig, M., (2018). Scaled consistent estimation of regression parameters in frailty models. *Acta Universitatis Lodziensis. Folia Oeconomica*, 5(338), 133–142.

Breslow, N., (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1), pp. 89–99.

Cox, D. R., (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34(2), pp. 187–220.

Efron, B., (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, 71(359), pp. 557–565.

Fahrmeir, L., Tutz, G., (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd Edition*, Springer-Verlag, New York.

Haitovsky, Y., 1982. Grouped data, in *Encyclopedia of Statistical Sciences 3*, John Wiley and Sons, pp. 527–536.

Hosmer, D., Jr, Lemeshow, S., (1999). *Applied Survival Analysis. Regression Modeling of Time to Event Data*, John Wiley and Sons.

Huber, P. J., Ronchetti, E. M., (2009). *Robust Statistics, 2nd Edition*, John Wiley and Sons.

Kalbfleich, J. D., Prentice, R. L., (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2), pp. 267–278.

Kalbfleich, J. D., Prentice, R. L., (1980). *The Statistical Analysis of Failure Time Data*, John Wiley and Sons.

Kestenbaum, B., Ferguson, R., (2010). Supercentenarians in the United States, in *H. Maier et al. (eds.), Demographic Research Monographs*, Springer, Berlin, Heidelberg, pp. 43–58.

McKeague, I. W., Zhang, M. J., (1996). Fitting Cox's Proportional Hazards Model Using Grouped Survival Data, in *N.P. Jewell et al. (eds.), Lifetime Data: Models in Reliability and Survival Analysis*, Kluwer, Boston, pp. 227–232.

Prentice, R. L., Gloeckler, L. A., (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34(1), pp. 57–67.

Ruud, P., (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica*, 51(1), pp. 225–228.

Stoker, T., (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54(6), pp. 1461–1481.

Thompson, W. A., (1977). On the treatment of grouped observations in live studies. *Biometrics*, 33(3), pp. 463–470.

Whitten, B. J., Cohen, A. C., Sundaraiyer, V., (1988). A pseudo-complete sample technique for estimation from censored samples. *Communications in Statistics - Theory and Methods*, 17(7), pp. 2239–2258.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4 pp. 191–202, https://doi.org/10.59139/stattrans-2024-010
Received – 09.09.2023; accepted – 20.07.2024

Comparison between classical and Bayesian estimation with joint Jeffrey's prior to Weibull distribution parameters in the presence of large sample conditions

Ahmed Mahdi Salih[1], **Murtadha Mansour Abdullah**[2]

Abstract

Weibull distribution has been considered one of the most common and valuable distributions for building and analyzing good models for lifetime data. Many researchers have studied the properties of Weibull distribution, also in search of the best method to estimate both parameters. In this paper, we proposed a comparison of Weibull distribution parameters under large sample conditions. We chose to study the classical estimation methods of Weibull distribution parameters, including the maximum likelihood estimator and moments estimation (ME). Next, we compared these methods with the Bayesian estimation method (BE) with Jeffrey's prior function. We validated the proposed study via simulation using both small and large samples. We used mean square errors (MSE) to determine the best estimation method. Our simulation findings suggest that maximum likelihood estimators are reasonably effective when using small sample sizes. In addition, in cases where the sample size is larger, the BE performed more effectively for both scale and shape parameters of the Weibull distribution function.

Key words: Weibull distribution, classic estimation, Bayesian estimation, Jeffrey's prior, large sample.

1. Introduction

Weibull distribution was first introduced by Waloddi Weibull (1951), and it has been widely used in reliability and life data analysis. Also, the Weibull distribution function can be used as a model for various life behaviors depending on the values of its parameters. Estimation of parameters for the Weibull distribution function is fundamental. There are two parameters of the Weibull distribution function; the first

[1] Statistics Department, College of Administration and Economics, Wasit University, Iraq. E-mail: amahdi@uowasit.edu.iq. ORCID: https://orcid.org/0000-0002-6109-224X.

[2] Statistics Department, College of Administration and Economics, Wasit University, Iraq. E-mail: mabdullah@uowasit.edu.iq. ORCID: https://orcid.org/0000-0002-3341-0415.

is the Shape Parameter β, which marks the behavior of the distribution. Different values of β give the Weibull distribution function variety.

Moreover, the Shape Parameter affects the failure rate of the distribution function in life data analysis. The second Weibull distribution function is the Scale Parameter α, which determines the probability density function's figure and peak. The height of the probability density function will decrease as α increases.

Many approaches have been submitted to estimate the two parameters of the Weibull distribution function, and many were considered classical methods of estimation, such as the Maximum Likelihood Estimator that depends on finding the values of parameters that maximize the joint probability function of observed data over the parameter space. In addition, the Moments Estimator starts by expressing the population as a function of the parameters. It is set to be equal to the sample moments to get equations and solve them by finding an estimator for the parameters. Other modern estimation methods were submitted, such as the Bayes Estimation method, which depends on knowledge about the prior distribution of the parameters to develop new and efficient estimators using Bayes Theory. Bayes estimation depends on selecting recently developed functions, such as Jeffrey's prior and Gamma-Gamma prior functions.

Over the years, many prior functions were submitted to be Informative and Non-Informative prior functions to get good quality estimators for Weibull distribution function parameters. All estimation methods agree on minimizing the difference between the observed value and the fitted value provided by the distribution function.

This paper is organized as follows: Section 2 discusses related work, Section 3 presents the Maximum Likelihood Estimation, Section 4 describes the Moment Estimator, Section 5 presents the Bayes Estimator, Section 6 presents comparison methods, Section 7 presents the discussion and results, and Section 8 presents conclusions.

2. Related Work

Weibull distribution has been widely studied by many researchers, such as:

(Mann, Schafer, & Singpurwalla, 1974) proposed a study of the analysis of reliability and life data that used the Weibull distribution function. They estimated the parameters in many graphical and analytical methods then studied the effect of parameter estimation methods in the reliability function.

(Popocikova & Sedliackova, 2014) compared different estimators for the Weibull distribution function for both shape and scale parameters. They studied the parameters' performance using the Weighted Least Square (WLS), Maximum Likelihood Estimator, and Moments Estimation methods (ME). The comparisons were based on Monte Carlo Simulation data using Minimum Square Errors as a comparison tool.

The following researchers discussed the excellent qualities of the Bayesian estimator using different functions:

(Aslam, Kazmi, Ahmad, & Shah, 2014) estimated the Weibull distribution function's shape and scale parameters using the Bayes Estimation method. They considered Informative and Non-Informative prior functions for both parameters. They also used many loss functions to improve the estimation. A comprehensive simulation was used to make a fair comparison among different Bayes estimates.

(Guure, Ibrahim, & Ahmed, 2012) proposed a study to estimate the Weibull distribution function parameters using classical Maximum Likelihood Estimation, Moments Estimation (ME), and Bayes Estimation method. They used Jeffrey's prior function as a Non-Informative prior function, and three types of loss functions to improve the Bayes estimates. Also, Minimum Square Errors were used to compare estimates of shape and scale parameters.

In this paper, we estimated the parameters of the Weibull distribution under the presence of a large sample size under study. We studied the performance of three estimation methods by changing the shape parameter under different sample sizes.

Estimating Weibull distribution parameters is a significant process used in different areas, including reliability and modeling lifetime data for engineering and medicine applications.

3. Maximum Likelihood Estimator

One of the most commonly used methods to estimate the parameter of any known distribution function is the Maximum Likelihood Estimation method, which utilizes the log-likelihood function to estimate parameters.

Let $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}$ be a random sample with n the Weibull Distribution pdf, which will be given as (Guure, Ibrahim, & Ahmed, 2012).

$$\boldsymbol{f(x)} = \left(\frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}}\right)\left(\frac{\boldsymbol{x}}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta-1}} \boldsymbol{exp}\left[-\left(\frac{\boldsymbol{x}}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta}}\right]. \tag{1}$$

where $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are the scale and shape parameters as $\boldsymbol{\alpha}, \boldsymbol{\beta} > \mathbf{0}$. Then, the likelihood of the pdf in (1) $\boldsymbol{f(x_1, x_2, \dots x_n \setminus \alpha, \beta)}$ will be (Nwobi & Ugomma, 2014).

$$\boldsymbol{L(x_1, x_2, \dots x_n \setminus \alpha, \beta)} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left\{\left(\frac{\boldsymbol{\beta}}{\boldsymbol{\alpha}}\right)\left(\frac{\boldsymbol{x_i}}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta-1}} \boldsymbol{exp}\left[-\left(\frac{\boldsymbol{x_i}}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta}}\right]\right\}. \tag{2}$$

The log-likelihood function is (Johnson, Kotz, & Balakrishnan, 1994), (Kundu, 2008).

$$\mathbf{ln}(\boldsymbol{L}) = \boldsymbol{n}\,\mathbf{ln}(\boldsymbol{\beta}) - \boldsymbol{n\beta}\,\mathbf{ln}(\boldsymbol{\alpha}) + (\boldsymbol{\beta}-\mathbf{1})\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \mathbf{ln}(\boldsymbol{x_i}) - \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \left(\frac{\boldsymbol{x_i}}{\boldsymbol{\alpha}}\right)^{\boldsymbol{\beta}} \tag{3}$$

By differentiating (3) respectively to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and equating to zero we get

$$\frac{\partial \ln L}{\partial \alpha} = -\boldsymbol{n}\left(\frac{\beta}{\alpha}\right) + \left(\frac{\beta}{\alpha}\right)\sum_{i=1}^{n}\left(\frac{x_i}{\alpha}\right)^{\beta} = \mathbf{0}. \tag{4}$$

$$\frac{\partial \ln L}{\partial \beta} = \left(\frac{n}{\beta}\right) + \sum_{i=1}^{n}\left(\frac{x_i}{\alpha}\right) - \sum_{i=1}^{n}\left(\frac{x_i}{\alpha}\right)^{\beta}\ln\left(\frac{x_i}{\alpha}\right) = \mathbf{0} \tag{5}$$

From (4) we can obtain (Zhang & Meeker, 2005)

$$\widehat{\alpha} = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]^{1/\beta} \tag{6}$$

It is useful here to use numerical methods to find the value of $\widehat{\boldsymbol{\beta}}$. If we consider that $\boldsymbol{f(\beta)}$ is the same in (5) we can use the Newton-Raphson method by taking the first differential of $\boldsymbol{f(\beta)}$ as below (Lawless, 2011).

$$\boldsymbol{f'(\beta)} = -\left(\frac{n}{\beta^2}\right) - \sum_{i=1}^{n}\left(\frac{x_i}{\alpha}\right)^{\beta}\ln^2\left(\frac{x_i}{\alpha}\right) \tag{7}$$

Substituting (6) in (5).

$$\boldsymbol{f(\beta)} = \left(\frac{\boldsymbol{n}}{\boldsymbol{\beta}}\right) + \sum_{i=1}^{n}\left[\frac{(x_i)}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]^{1/\beta}}\right] - \sum_{i=1}^{n}\left[\frac{(x_i)^{\beta}}{\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}}\right]\ln\left[\frac{(x_i)}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]^{1/\beta}}\right] \tag{8}$$

Substituting (6) in (7).

$$\boldsymbol{f'(\beta)} = -\left\{\left(\frac{n}{\beta^2}\right) + \sum_{i=1}^{n}\left[\frac{(x_i)^{\beta}}{\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}}\ln^2\frac{x_i}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]^{1/\beta}}\right]\right\}. \tag{9}$$

Then, by choosing an initial value for $\boldsymbol{\beta_i}$ we can obtain $\widehat{\boldsymbol{\beta}}$ by iterating the formula below until it converges to the MLE for $\boldsymbol{\beta}$.

$$\boldsymbol{\beta_{i+1}} = \boldsymbol{\beta_i} - \frac{\left(\frac{n}{\beta}\right)+\sum_{i=1}^{n}\left[\frac{(x_i)}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]^{1/\beta}}\right] - \sum_{i=1}^{n}\left[\frac{(x_i)^{\beta}}{\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}}\right]\ln\left[\frac{(x_i)}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]^{1/\beta}}\right]}{-\left\{\left(\frac{n}{\beta^2}\right)+\sum_{i=1}^{n}\left[\frac{(x_i)^{\beta}}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]}\ln^2\frac{x_i}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i)^{\beta}\right]^{1/\beta}}\right]\right\}}. \tag{10}$$

4. Moments Estimator

The method of moments is commonly used depending on obtaining the K_{th} Moment *Mk* of the Weibull distribution function and equating it with the sample K_{th} moments given by (Pobočıková & Sedliačková, 2014).

$$\boldsymbol{M_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k} \tag{11}$$

Then, $\boldsymbol{M_1 = \overline{x} = E(x)}$, which is equal to the expected value of the Weibull distribution function.

$$\boldsymbol{\beta\Gamma\left(1+\frac{1}{\alpha}\right) = \overline{x}} \tag{12}$$

$$\boldsymbol{\beta^2\Gamma\left(1+\frac{2}{\alpha}\right) = \frac{1}{n}\sum_{i=1}^{n} x_1^2} \tag{13}$$

And by dividing (13) on the square of (12)

$$\boldsymbol{\frac{\Gamma\left(1+\frac{2}{\alpha}\right)}{\Gamma^2\left(1+\frac{1}{\alpha}\right)} = \frac{\frac{1}{n}\sum_{i=1}^{n} x_1^2}{\overline{x}^2}} \tag{14}$$

There is no analytical solution for (14), so we can use numerical methods to estimate $\boldsymbol{\alpha}$. Ramerez and Carta (2005) gave a starting point for $\boldsymbol{\alpha}$ such that:

$$\boldsymbol{\widehat{\alpha} = \left(\frac{\overline{x}}{S_x}\right)^{1.086}} \tag{15}$$

where $\boldsymbol{\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i}$ and $\boldsymbol{S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ then from (12).

$$\boldsymbol{\widehat{\beta} = \frac{\overline{x}}{\Gamma\left(1+\frac{1}{\widehat{\alpha}}\right)}} \tag{16}$$

5. Bayes Estimator

The Bayesian estimation method is critical and has attracted much attention lately. The Bayesian approach begins with determining a prior distribution function for the parameters under study. When we have information about the parameters, we may use the Informative prior distribution function or the Non-Informative prior distribution function. We have no knowledge about the parameters here, so we used the Non-Informative prior function. The most common prior distribution function is Jeffrey's prior. Jeffrey (Guure, Ibrahim, & Ahmed, 2012) suggested using the square root of the determinant of the Fisher information matrix as a prior distribution function for the parameters such that $\boldsymbol{u(\alpha,\beta) = \sqrt{\det(I_{(\alpha,\beta)})}}$ where:

$$\boldsymbol{I_{(\alpha,\beta)} = \begin{bmatrix} E\left(\frac{\partial^2 \log(f_{(x)})}{\partial^2\alpha^2}\right) & E\left(\frac{\partial^2 \log(f_{(x)})}{\partial\alpha\partial\beta}\right) \\ E\left(\frac{\partial^2 \log(f_{(x)})}{\partial\alpha\partial\beta}\right) & E\left(\frac{\partial^2 \log(f_{(x)})}{\partial^2\beta^2}\right) \end{bmatrix}} \tag{17}$$

According to (Guure, Ibrahim, & Ahmed, 2012), the final result for the prior distribution function will be as follows:

$$\boldsymbol{u(\alpha,\beta)=\frac{1}{\alpha\beta}} \tag{18}$$

Since we have satisfied the prior distribution function, we can now compute the posterior distribution function according to Bayes theory, in which the joint density function of $\boldsymbol{\alpha,\beta}$ is:

$$\boldsymbol{f(\alpha,\beta\backslash x_1,x_2,\ldots.x_n)=\frac{f(x_1,x_2,\ldots x_n\backslash\alpha,\beta)u(\alpha,\beta)}{\int_0^\infty\int_0^\infty f(x_1,x_2,\ldots x_n\backslash\alpha,\beta)u(\alpha,\beta)d\beta d\alpha}} \tag{19}$$

By using the likelihood function:

$$\boldsymbol{f(\alpha,\beta\backslash x_1,x_2,\ldots.x_n)=\frac{L(x_1,x_2,\ldots x_n\backslash\alpha,\beta)u(\alpha,\beta)}{\int_0^\infty\int_0^\infty L(x_1,x_2,\ldots x_n\backslash\alpha,\beta)u(\alpha,\beta)d\beta d\alpha}} \tag{20}$$

Then:

$$\boldsymbol{f(\alpha,\beta\backslash x_1,x_2,\ldots.x_n)=\frac{\frac{1}{\alpha\beta}\prod_{i=1}^n\left\{\left(\frac{\beta}{\alpha}\right)\left(\frac{x_i}{\alpha}\right)^{\beta-1}exp\left[-\left(\frac{x_i}{\alpha}\right)^\beta\right]\right\}}{\int_0^\infty\int_0^\infty\frac{1}{\alpha\beta}\prod_{i=1}^n\left\{\left(\frac{\beta}{\alpha}\right)\left(\frac{x_i}{\alpha}\right)^{\beta-1}exp\left[-\left(\frac{x_i}{\alpha}\right)^\beta\right]\right\}d\beta d\alpha}}. \tag{21}$$

Therefore, the Bayes estimators for the parameters will be:

$$\boldsymbol{BE_\alpha=E(\alpha\backslash x_1,x_2,\ldots.x_n)=\int_0^\infty\alpha f(\alpha,\beta\backslash x_1,x_2,\ldots.x_n)d\alpha} \tag{22}$$

$$\boldsymbol{BE_\beta=E(\beta\backslash x_1,x_2,\ldots.x_n)=\int_0^\infty\beta f(\alpha,\beta\backslash x_1,x_2,\ldots.x_n)d\beta} \tag{23}$$

In addition, we suppose that $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are independent.

6. Comparison Method

Different statistical tools can be used to make a fair comparison among estimators, and here we selected the MSE Mean Squared Errors, which is given by.

$$\boldsymbol{MSE=\frac{1}{n}\sum_{i=1}^n\left[\widehat{F}(x_i)-F(x_i)\right]^2} \tag{24}$$

where $\boldsymbol{F(x_i)}$ is the cumulative distribution function for the Weibull distribution as follows (Pobočıková & Sedliačková, 2014):

$$\boldsymbol{F(x_i)=\begin{cases}1-exp\left\{-\left(\frac{x_i}{\alpha}\right)^\beta\right\}, & x\geq 0\\ 0 \quad , & otherwise\end{cases}} \tag{25}$$

Thus, we can use parameters and their estimators to substitute in (25).

7. Simulation Study

In this section, we used a MATLAB program to make a Monte Carlo simulation to generate samples of random variables with the Weibull distribution; a Monte Carlo simulation method depends on generating initial random variables with Uniform Distribution and then generating the Weibull Distribution according to its cumulative distribution function. We can summarize the Monte Carlo Simulation for Weibull Distribution in the following steps:

1. Generating Normal z by using Lehmer's recursion simple random generator which is $\boldsymbol{z = az_0 \; mod \; m}$ where $\boldsymbol{z_0 = 1}$, $\boldsymbol{a = 3}$, $\boldsymbol{m = 31}$ where *a* and *m* can be changed to have a cycle of random numbers.
2. Normalizing z to obtain a random variable (u) with a value between zero and one $\boldsymbol{u = z/m}$
3. Obtaining t from the cumulative distribution function of the Weibull Distribution by equalizing it to u, i.e. $\boldsymbol{u = 1 - e^{-\left(\frac{t}{\alpha}\right)^{\beta}}}$ then $= \boldsymbol{\alpha \sqrt[\beta]{-\ln u}}$.

Then, t is a random variable with the Weibull Distribution, and we repeat these steps for 300 times in order to get random samples. Here, we chose only three values for shape parameter $\boldsymbol{\beta = 1.5}$, $\boldsymbol{2}$, $\boldsymbol{2.5}$ and we chose one value for scale parameter $\boldsymbol{\alpha = 0.5}$. These selections for the parameters' values were used to generate random samples [9], and here we chose sample sizes $\boldsymbol{n = 10, 20, \ldots 120}$ to cover both small and big samples.

Table 1: The estimated values of $\hat{\alpha}$, $\hat{\beta}$ and the obtained MSE with three estimation methods: Maximum Likelihood Estimation Method, Moment Estimation and Bayes Estimation, when $\alpha = 0.5$, $\beta = 1.5$

n	MLE			ME			BE		
	$\hat{\alpha}$	$\hat{\beta}$	MSE	$\hat{\alpha}$	$\hat{\beta}$	MSE	$\hat{\alpha}$	$\hat{\beta}$	MSE
10	0.7821	2.2763	2.7820	0.9013	2.7011	3.3212	0.9480	2.3631	2.8821
20	0.7612	2.2283	2.7723	0.8877	2.6699	3.2772	0.8931	2.3231	2.8271
30	0.7537	2.1843	2.7423	0.8721	2.6278	3.2242	0.8430	2.3141	2.7741
40	0.7411	2.1463	2.7024	0.8600	2.5790	3.1811	0.7921	2.2451	2.7231
50	0.7201	2.1133	2.6623	0.8511	2.5401	3.1404	0.7420	2.2071	2.6741
60	0.7015	2.0856	2.6128	0.8489	2.5154	3.1002	0.6961	2.1711	2.6291
70	0.6714	2.0456	2.5523	0.8401	2.4853	3.0598	0.6622	2.1276	2.5731
80	0.6421	2.0421	2.4732	0.8362	2.4644	3.0209	0.6180	2.0926	2.4921
90	0.6001	2.0412	2.4123	0.8302	2.4501	2.9849	0.6000	2.0576	2.4111
100	0.5598	2.0213	2.3576	0.8277	2.4400	2.9509	0.5381	2.0221	2.3301
120	0.5342	1.9963	2.3041	0.8223	2.4306	2.9199	0.5221	1.9891	2.2491
140	0.5210	1.9743	2.2598	0.8204	2.4214	2.8791	0.5091	1.9561	2.1701
160	0.5189	1.9633	2.2087	0.8188	2.4123	2.8341	0.5042	1.9251	2.0951
180	0.5045	1.9243	2.1665	0.8161	2.4037	2.8021	0.5011	1.8932	2.0241
200	0.5001	1.9153	2.0912	0.8155	2.3944	2.7711	0.5000	1.6551	1.9541

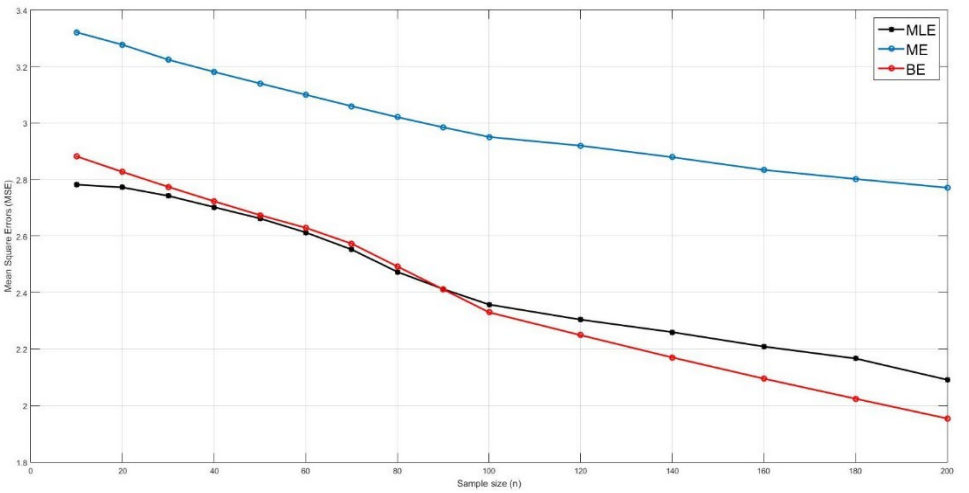


Figure 1: The values of MSE for the three estimation methods: Maximum Likelihood Estimation Method, Moment Estimation, and Bayes Estimation, when $\alpha = 0.5\,, \beta = 1$

While we have $\boldsymbol{\beta = 1.5\,, \alpha = 0.5}$, Table 1 and Figure 1 show the effectiveness of the MLE estimator with a small sample size, and this continues until we get to sample sizes of 90 as shown in Figure 1. The BE estimator showed great performance and improved as the sample size increased. The ME estimator showed no priority with both small and large sample sizes.

Table 2: The estimated values of $\hat{\alpha}\,, \hat{\beta}$ and the obtained MSE with three estimation methods: Maximum Likelihood Estimation Method, Moment Estimation and Bayes Estimation when $\alpha = 0.5\,, \beta = 2$

n	MLE			ME			BE		
	$\hat{\alpha}$	$\hat{\beta}$	MSE	$\widehat{\alpha}$	$\widehat{\beta}$	MSE	$\widehat{\alpha}$	$\widehat{\beta}$	MSE
10	0.8371	2.6216	3.7820	0.9013	2.9011	4.3321	0.8780	2.7031	3.8755
20	0.8161	2.6006	3.7611	0.8801	2.8801	4.2121	0.8510	2.6721	3.8445
30	0.7951	2.5796	3.7411	0.8601	2.8601	4.1521	0.8351	2.6421	3.8145
40	0.7761	2.5606	3.7221	0.8411	2.8411	4.0931	0.8001	2.6131	3.7855
50	0.7581	2.5426	3.7041	0.8231	2.8231	4.0751	0.7761	2.5851	3.7575
60	0.7411	2.5256	3.6871	0.8016	2.8061	4.0571	0.7531	2.5581	3.7305
70	0.7251	2.5096	3.6721	0.7901	2.7901	4.0411	0.7311	2.5331	3.7045
80	0.7101	2.4936	3.6581	0.7751	2.7751	4.0261	0.7101	2.5091	3.6785
90	0.6961	2.4786	3.6451	0.7611	2.7611	4.0121	0.6901	2.4861	3.6525
100	0.6831	2.4684	3.6341	0.7481	2.7481	3.9991	0.6711	2.4641	3.6275
120	0.6731	2.4469	3.6231	0.7361	2.7361	3.9871	0.6531	2.4441	3.6035
140	0.6581	2.4366	3.6131	0.7251	2.7251	3.9761	0.6361	2.4241	3.5805
160	0.6461	2.4236	3.6041	0.7151	2.7151	3.9661	0.6201	2.4051	3.5585
180	0.6361	2.4116	3.5951	0.7061	2.7061	3.9571	0.6041	2.3671	3.5375
200	0.6231	2.4006	3.5871	0.6981	2.6981	3.9491	0.5901	2.3701	3.5165

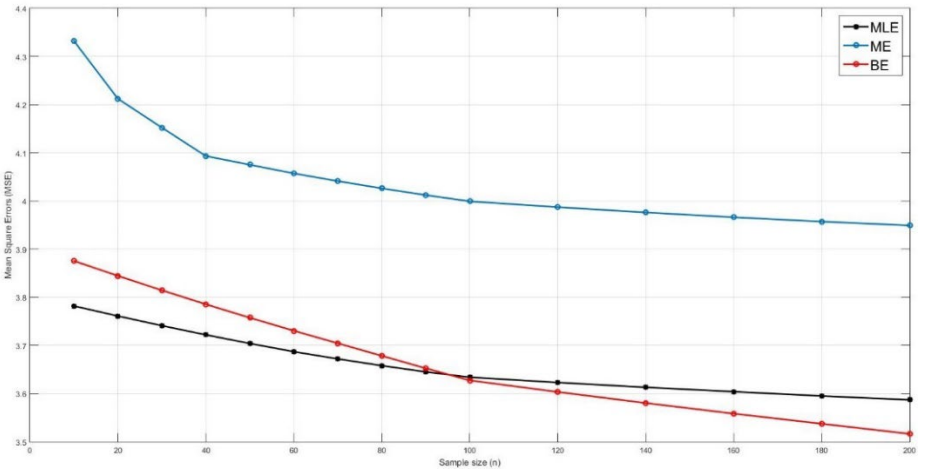


Figure 2: The value of MSE for the three estimation methods: Maximum Likelihood Estimation Method, Moment Estimation, and Bayes Estimation, when $\boldsymbol{\alpha = 0.5}$, $\boldsymbol{\beta = 2}$

By changing $\boldsymbol{\beta = 2}$ and from Table 2 and Figure 2, we see similar results, but here the MLE estimator's performance is good only to sample sizes of 100. The BE estimator becomes a better estimator until sample sizes reach 200, where the ME estimator's performance is not good as the sample size changes.

Table 3: The estimated values of $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ and the obtained MSE with three estimation methods: Maximum Likelihood Estimation Method, Moment Estimation and Bayes Estimation $\boldsymbol{\alpha = 0.5}$, $\boldsymbol{\beta = 2.5}$

n	MLE			ME			BE		
	$\hat{\alpha}$	$\hat{\beta}$	MSE	$\widehat{\alpha}$	$\widehat{\beta}$	MSE	$\widehat{\alpha}$	$\widehat{\beta}$	MSE
10	0.7181	3.2216	3.9120	0.9913	3.9211	4.3721	0.7980	3.5531	4.0055
20	0.6971	3.2006	3.8910	0.9703	3.9001	4.3511	0.7701	3.5241	3.9765
30	0.6771	3.1806	3.8711	0.9503	3.8801	4.3311	0.7431	3.4961	3.9485
40	0.6581	3.1616	3.8521	0.9313	3.8611	4.3121	0.7171	3.4691	3.9205
50	0.6401	3.1436	3.8341	0.9133	3.8431	4.2941	0.6911	3.4431	3.8935
60	0.6231	3.1266	3.8171	0.8963	3.8261	4.2771	0.6661	3.4171	3.8665
70	0.6071	3.1096	3.8011	0.8813	3.8101	4.2611	0.6411	3.3921	3.8405
80	0.5921	3.0936	3.7861	0.8673	3.7951	4.2461	0.6171	3.3681	3.8145
90	0.5781	3.0776	3.7731	0.8543	3.7811	4.2311	0.5941	3.3451	3.7885
100	0.5651	3.0626	3.7611	0.8423	3.7671	4.2181	0.5721	3.3231	3.7635
120	0.5531	3.0486	3.7501	0.8313	3.7541	4.2051	0.5511	3.3011	3.7385
140	0.5411	3.0356	3.7401	0.8203	3.7421	4.1931	0.5471	3.2801	3.7145
160	0.5301	3.0236	3.7311	0.8103	3.7301	4.1811	0.5321	3.2601	3.6905
180	0.5191	3.0126	3.7231	0.8013	3.7191	4.1701	0.5110	3.2411	3.6665
200	0.5101	3.0026	3.7161	0.7933	3.7091	4.1591	0.5001	3.2231	3.6425

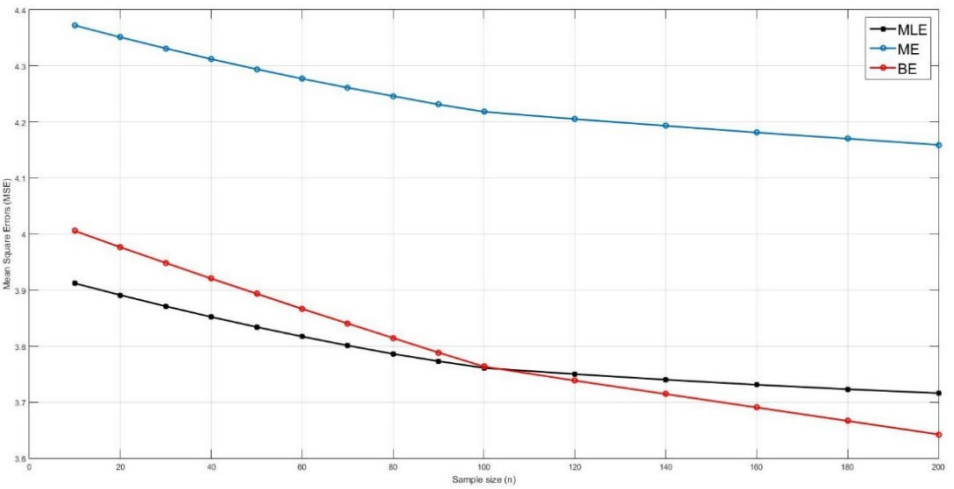


Figure 3: The values of MSE for the three estimation methods: Maximum Likelihood Estimation Method, Moment Estimation and Bayes Estimation, when $\boldsymbol{\alpha = 0.5}$, $\boldsymbol{\beta = 2.5}$

By changing $\boldsymbol{\beta = 2.5}$ Table 3 and Figure 3 show no change except that the BE estimator is better than the MLE estimator when the sample size is 120 and MLE is better when the sample size is smaller than 120, while the ME estimator is the same as above.

8. Discussion

We believe that estimating the parameters of the Weibull Distribution function is an essential procedure in many statistical applications. In Section 7, we made a simulation by fixed shape parameter and many scale parameter values to cover multiple cases of this distribution function by increasing the scale parameter, $\boldsymbol{\alpha}$ will increase the peak of the probability density function. From Figures 1, 2 and 3, we can see that the Bayes Estimator will have better performance as it increases in scale parameter $\boldsymbol{\alpha}$ and sample size either there is good performance of the Maximum likelihood Estimator MLE when both scale parameter $\boldsymbol{\alpha}$ and the sample size are small. By increasing the scale parameter $\boldsymbol{\alpha}$, we can see the MLE preference decreases to smaller sample sizes. As for the Moment Estimator ME, we can see from all figures that changing the scale parameter $\boldsymbol{\alpha}$ did not result in a good quality estimator in all sample sizes.

9. Conclusions

This paper compares and demonstrates three methods for estimating parameters: Maximum Likelihood Estimator, Moment Estimator, and Bayesian Estimator with a non-informative prior (weak prior with minimal influence).

The paper uses Monte Carlo simulations and analyzes results from tables and figures (not included here).

The Maximum Likelihood Estimator outperforms the Moment Estimator and Bayesian Estimator for small sample sizes. Concerning larger sample sizes: Maximum Likelihood Estimator and Bayesian Estimator perform better than Moment Estimator.

Finally, for large sample sizes, Bayesian Estimator becomes significantly better than the Maximum Likelihood Estimator and Moment Estimator for parameter estimation.

Acknowledgment

Our warm thanks are due to our colleagues in the Wasit University Statistics Department, who shared this wonderful experience with us.

References

Aslam, M., Kazmi, S., Ahmed, I. and Shah S., (2014). Bayesian Estimation for Parameters of the Weibull Distribution. *Science International,* Vol. 5, No. 1089 Lahore, Pakistan, pp. 1915–1920.

Berger, J. O., Sun D., (1993). Bayesian analysis for the poly-Weibull distribution. *Journal of the American Statistical Association*, Vol. 88, Issue 424, pp. 1412–1418.

Guure, C., Ibrahim, N. and Omari, M. Al., (2012). Bayesian Estimation of Two-Parameter Weibull Distribution Using Extension of Jeffreys' Prior Information with Three loss Functions. *Mathematical Problems in Engineering*, Article Id 58940.

Johnson, N. L., Kotz, S. and Balakrishnan, N., (1994). *Maximum Likelihood Estimation for Weibull Distribution,* John Wiley & Sons, New York.

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

Lawless, J. F., (2003). *Statistical Models and Methods for Lifetime Data*. 3ed. Edition, John Wiley and Sons, New York.

Mann, N., Schafer, R. and Singpurwalla, N., (1974). *Methods of Statistical Analysis of Reliability and Life Data*, John Wiley & Sons, New York.

Nwobi, F., Ugomma, C., (2014). A Comparison of Methods for the Estimation of Weibull Distribution Parameters. *Metodoloski zvezki*, Vol. 11, No.1, pp. 65–78.

Pobocikova, I., Sedliackova, Z., (2014). Comparison of Four Methods for Estimating the Weibull Distribution Parameters. *App Math Sci*, Vol. 8, No. 83, pp. 4137–4149.

Zhang, Y., Meeker, W. Q., (2005). Bayesian life testing planning for the Weibull distribution with given shape parameter. *Metrika,* Vol. 61, pp. 237–249.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 203–206

About the Authors

Abdullah Murtadha Mansour is an Assistant Professor at the Department of Statistics, College of Administration and Economics, Wasit University. At the same time, he is a skilled trainer of SPSS and leading the program of developing the skills of higher studies students in Wasit University. Moreover, he is leading studies for the Iraqi Meteorological Organization & Seismology. His main areas of interest include weather data analysis, categorical time series analysis, agrometeorological data analysis. He has more than 15 published papers and studies, and is a member of editorial boards: Al Kut Journal of Economics and Administrative Sciences (KJEAS), Wasit university press.

Agyemang Edmund Fosu is a Presidential Research Fellow at the University of Texas Rio Grande Valley, USA, where he serves as a graduate research assistant at the School of Statistical and Mathematical Science. His primary research interest is applied statistics with interdisciplinary applications. Besides this, he has interests in a range of statistical fields including machine learning in health, time series analysis of health data, design and analysis of experiments, and computational statistics. Edmund has published close to fifteen (15) research papers in international and national peer-reviewed journals. He currently serves as an editorial board member for academic journals including the Journal of HIV/AIDS Research, Journal Multidisciplinar (Montevideo), and Global Open Share Publishing. Additionally, he serves as a peer reviewer for more than ten (10) international journals.

Anis Mohammed Zafar is a Member of the Faculty of the Indian Statistical Institute (ISI), Calcutta. He received the prestigious Mary G. & Joseph Natrella Award from the American Statistical Association in 2012 and was elected as a Member of the International Statistical Institute in 2023. He has presented his works in many international conferences abroad and has authored more than 50 papers in referred internationally reputed journals. He has also served the statistical community by reviewing papers for many journals; and was the Vice-President (Membership & Outreach) of the International Society for Business & Industrial Statistics during 2021-23. He is Member-Secretary of the International Statistical Education Center at the ISI. His research interests are in reliability, SPC and applied statistics.

Bera Kuntal is currently a PhD student at the Statistical Quality Control and Operations Research Division, Indian Statistical Institute, Kolkata, India. He received the Bachelor of Science degree in Mathematics in 2014 from Calcutta University and

the Master of Science degree in Mathematics in 2016 from Presidency University, Kolkata, India. His research interests include process capability analysis, statistical quality control and statistical inference. He has published some papers in international journals.

Biswas Soma Chowdhury is a Full Professor of Statistics at the University of Chittagong, Chattogram, Bangladesh. She holds a BSc (Hons) and MSc in Statistics from the University of Dhaka, Bangladesh; an MA in Demography from ANU, Canberra, Australia, and a PhD in Biostatistics from RCMPS, University of Chittagong, Bangladesh. Her research expertise includes advanced statistical methods like Markov Chains, GLMs, and multivariate analysis, with a focus on health data. Professor Soma has authored many articles, have conference papers, and a published book. She is an active member of several scientific and professional organizations.

Derkacz Arkadiusz J., economist, economic analyst and specialist in international financial, economic and legal relations. Long-time manager and chief financial officer in multinational companies. Associate Professor at the Institute of Management and Quality Sciences of the University of Kalisz. Economic expert of the National Bank of Poland. His research interests include the housing rental market, the economic condition of enterprises and the feeling of economic security of households. Author and co-author of scientific publications and market analyses and reports.

Komara Silvia is an Assistant Professor at the Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava. She specializes in machine learning, time series analysis, and forecasting, with a strong focus on achieving high accuracy in short-term predictions. Her research combines advanced statistical modelling and computational techniques to address challenges in data-driven decision-making across various domains.

Košíková Martina is an Assistant Professor at the Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava. Her research focuses especially on the statistical analysis of poverty and social exclusion. She deals with the application of advanced statistical methodologies, particularly general and generalized linear models, complemented by contrast analysis and marginal means analysis, to provide deeper insights into complex relationships within socio-economic data.

Kroszka Jan obtained a master's degree in Quantitative Methods in Economics at SGH Warsaw School of Economics and currently works as a research assistant. Previously, he also studied Sociology at the University of Warsaw and the University of Vienna. His research interests focus on political economy and applications of statistical methods in the social sciences.

Malviya Priyanka is a visiting faculty of School of Data Science and Forecasting, Devi Ahilya Vishwavidyalaya, Indore, M.P., India. She has published 8 research papers in international and national journals.

Nowak Piotr Bolesław received his PhD degree in Mathematics from University of Wrocław and now he is an Assistant Professor at the Institute of Economic Sciences of the Faculty of Law, Administration and Economics of the University of Wrocław. His research concerns mathematical statistics and its applications, in particular reliability theory, survival analysis, applied mathematics in economy and medicine.

Panek Tomasz is a Full Professor at the Institute of Statistics and Demography at the Warsaw School of Economics. His research interests focus on household living conditions, including poverty, social exclusion, inequality, quality of life, income and consumption, as well as electronic media consumption, multivariate comparative analysis methods, survey methodology, and data analysis. He is a co-initiator and co-author of the panel study on living conditions and quality of life "Social Diagnosis", conducted between 2000 and 2015. Professor Panek has coordinated numerous projects and led SGH teams' participation in national and international consortium projects. He has published over 100 research papers in international and national journals and several books and monographs. Professor Panek is a member of the Scientific Statistical Council of Statistics Poland and an elected member of the International Statistical Institute.

Sacchidanand Majumder is a dedicated PhD Fellow in Statistics at the University of Chittagong, Chattogram, Bangladesh, focusing on his dissertation titled "Influence of Household and Demographic Characteristics in Achieving Sustainable Development Goals (SDGs) in Bangladesh." He has a solid academic background, having earned a BSc (Hons) and MSc in Statistics, and an MPhil in Statistics from the same institution. Professionally, Sacchidanand is an expert in development and research, addressing critical issues in Bangladesh. His research interests include SDGs, poverty, child nutrition and mortality, education, water and sanitation, and access to energy, as well as statistical methodologies like multivariate analysis, modeling, forecasting, and Bayesian networks, etc. He actively contributes to the field through numerous publications in international and national journals and conference proceedings.

Salih Ahmed Mahdi is an Assistant Professor at the Department of Statistics, College of Administration and Economics, Wasit University. At the same time, he is an expert in Big Data Analysis, and he is leading studies and programs for the Iraqi Central Organization of Statistics & Information Technology (COSIT). Moreover, he is a professional trainer in MATLAB programming in the Iraqi program for developing higher studies students. His main areas of interest include multidimensional poverty data analysis, big data analysis, child labor and social deprivation data analysis. He has

more than 25 published papers and studies, and serves as a member of editorial boards: Al Kut Journal of Economics and Administrative Sciences (KJEAS), Wasit university press.

Šoltés Erik is a Professor at the Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava. His research focuses on statistical modelling through regression models, general and generalized linear models, and statistical analysis based on statistical inference and multivariate statistical methods. He also deals with the theory of credibility. In his scientific work, he focuses mainly on the area of poverty and social exclusion, business demography, and the application of credibility theory and statistical modelling in non-life insurance. He is the author of more than 180 professional and scientific works (25 in WoS and SCOPUS), for which there are more than 420 citations and responses (about 150 in WoS and SCOPUS).

Šoltésová Tatiana is an Associate Professor at the Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava. Her scientific activity is focused on the application of stochastic methods and models in life insurance. She also deals with the application of mathematical and statistical methods and procedures in the field of poverty and social exclusion. The results of her research work in this area are published in several conference proceedings and scientific journals (several are in WoS and SCOPUS).

Singh Housila P. is a retired Professor of Statistics, Vikram University, Ujjain, M.P., India. His research area of interests includes sample surveys and statistical inference. Professor Housila P. Singh has guided 23 students for their PhD degrees and 17 students for their MPhil degrees. Housila P. Singh has published more than 530 research papers in international and national journals of repute. As per Google Scholar, he has citations: 7529, h-index: 42 and i-10-index: 172.

Tailor Rajesh is a Professor of Statistics at the School of Studies in Statistics, Vikram University, Ujjain, M.P., India. He has authored over 120 research papers in esteemed international and national journals. Under his mentorship, 13 students have earned their PhD and 8 students have completed their MPhil degrees.

Wójcik Sebastian is an Assistant Professor at the Institute of Mathematics at the University of Rzeszów and Head of the Mathematical Statistics Division at the Statistical Office in Rzeszów. His main areas of interest include probability theory, functional equations, utility theory, as well as data analysis and visualization in R.

Zwierzchowski Jan is a faculty member at the Institute of Statistics and Demography at the Warsaw School of Economics (SGH). His research interests encompass the measurement of quality of life, poverty, and social exclusion. His professional expertise focuses on the methodology of conducting social research.

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 207–212

Acknowledgements to Reviewers

The Editor and Editorial Board of Statistics in Transition new series wish to thank the following persons who served as peer-reviewers of manuscripts for the *Statistics in Transition new series* – Volume 25, Numbers 1–5. The authors` work has benefited from their feedback.

Adesina, Sunday Olumide, Department of Mathematics, Redeemers university, Ede Osun State, Nigeria

AlakuŞ, Kamil, Department of Statistics, Mayus University, Turkey

Al-Aqti, Ahmed Baqer Jaafar, Department of Mathematics, College of Education for Pure Science, University of Thi-Qar, Iraq

Al-Nasser, Amjad D., Department of Statistics, Yarmouk University, Jordan

Ariffin Masron, Tajul, Department of Management, University Sains Malaysia (USM), Malesia

Asgharzadeh, Akbar, Department of Statistics, University of Mazandaran, Iran

Bagheri Khoolenjani Nayereh, Department of Statistics, University of Isfahan, Iran

Bayoud, Husam Awni, College of Sciences and Humanities, Fahad Bin Sultan University, Saudi Arabia

Białek, Jacek, Department of Statistical Methods, University of Lodz & Statistics Poland, Poland

Brzeziński, Michał, Department of Political Economy, Warsaw Centre for Ecological Economics, University of Warsaw, Poland

Cerqueti, Roy, Department of Social and Economic Sciences, Sapienza University of Rome, Italy

Chandra, Girish, Department of Statistics, University of Allahabad, India

Cheng, Yang, Data Mining and Machine Learning Group (DMGroup), School of Computer Science, Beijing University of Posts and Telecommunications, China

Chutiman, Nipaporn, Department of Mathematics, Faculty of Science, Mahasarakham University, Mahasarakham, Thailand

Clarke, Brenton R., Department of Mathematics and Statistics, Murdoch University, Australia

Cohen, Achraf, Department of Mathematics and Statistics, University of West Florida, USA

Dehnel, Grażyna, Department of Statistics, Poznan University of Economics and Business, Poland

Delia, David, Department of Economics, Vasile Goldis Western University of Arad, Romania

Derid, Iryna, Department of International Business and Economic Theory, V. N. Karazin Kharkiv National University Ukraine, Ukraine

Dhaiban, Ali Khaleel, Department of Statistics, Al-Mustansiriyah University, Iraq

Dhar, Soma, Department of Statistics, Handique Girls' College, India

Dihidar, Kajal, Indian Statistical Institute, Kolkata, India

Dittmann, Iwona, Department of Finance, Wroclaw University of Economics and Business, Poland

Domański, Czesław, Department of Statistical Methods, University of Lodz, Poland

Dorugade, A.V., Department of Statistics, Y C Mahavidyalaya Halkarni University of India, India

Dutta, Subhankar, Department of Mathematics, National Institute of Technology, India

Dziechciarz, Józef, Department of Econometrics and Operations Research, Wroclaw University of Economics and Business, Poland

Eilers, Paul, Department of Biostatistics, Erasmus University, The Netherland

Ettahir, Aziz, Department of Mathematics, Mohammed V University, Rabat, Morocco

Fajriyah, Rohmatul, Department of Statistics, Univeristas Islam Indonesia, Indonesia

Gaire, Arjun Kumar Gaire, Department of Population Studies, Khwopa Engineering College, Nepal

Gemeay, Ahmed M., Department of Mathematics, Tanta University, Egypt

Genest, Christian, Department of Mathematics and Statistics, McGill University, Canada

Glaser-Opitzová, Helena, Statistical Office of the Slovak Republic, Slovakia

Kokoszka, Piotr, Department of Statistics, Colorado State University, USA

Kosztowniak, Aneta, Department of Applied Economics, Warsaw School of Economics, Poland

Kot, Stanisław Maciej, Department of Statistics and Econometrics, Gdańsk University of Technology, Poland

Krzyśko, Mirosław, Professor Emeritus, Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznan, Poland

Kuźmiński, Łukasz, Department of Process Management, Wroclaw University of Economics and Business, Poland

Lahiri, Partha, Department of Mathematics, University of Maryland, USA

Leśkow, Jacek, Department of Informatics, Cracow University of Technology, Poland & American University Kyiv, Ukraine

Lula, Paweł, Department of Statistics, Krakow University of Economics, Poland

Mahmood, Ehab. A., Department of Banking and Finance, University of Babylon, Iraq

Makhdoom, Iman, Department of Statistics, Payame Noor University, Iran

Marek, Luboš, Department of Statistics and Probability, Prague University of Economics and Business, Czech Republic

Maślankowski, Jacek, Department of Economic Informatics, University of Gdansk, Poland

Michalková, Mária, Department of Applied Mathematics, University of Žilina, Slovakia

Młodak, Andrzej, Statistical Office in Poznań, Poland & Interfaculty Department of Mathematics and Statistics, University of Kalisz, Poland,

Münnich, Ralf, Department of Economics, University of Trier, Germany

Myck, Michał, Centre for Economic Analysis, CenEA, Szczecin, Poland

Thillaigovindan, Natesan N., College of Natural Sciences Arba Minch University, Arba Minch, Ethiopia

Nestić, Danijel, The Institute of Economics, Zageb, Croatia

Okaba, Yoshihiko, National Academy of Agrarian Sciences of Ukraine, Japan

Okrasa, Włodzimierz, Cardinal Stefan Wyszyński University in Warsaw & Statistics Poland, Poland

Singh, Sarjinder, Department of Statistics, St. Cloud State University, USA

Sokołowski, Andrzej, Department of Statistics, Cracow University of Economics, Poland

Suntornchost, Jiraphan, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Thailand

Švábová, Lucia, Department of Economics, University of Žilina, Slovakia

Szmytkie, Robert, Institute of Geography and Regional Development, University of Wroclaw, Wroclaw, Poland

Śliwka, Piotr, Institute of Computer Science, Cardinal Stefan Wyszyński University in Warsaw, Poland

Tiensuwan, Montip, Department of Mathematics, Mahidol University, Thailand

Tongur, Can, Statistical Office of Sweden, Sweden

Trzęsiok, Joanna, Department of Economic and Financial Analysis, University of Economics in Katowice, Poland

Van Hoa, Tran, Centre for Strategic Economic Studies (CSES), Victoria University, Australia

Verma, Med Ram, Division of Livestock Economics, Statistics and Information Technology, Indian Veterinary Research Institute, India

Wagalla, Alplhonse, Department of Mathematics, Bomett University College, Kenia

Wołyński, Waldemar, Department of Mathematical Statistics and Data Analysis, Collegium Mathematicum, Adam Mickiewicz University in Poznan, Poland

Wywiał, Janusz, Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland

Zeghdoudi, Halim, LaPS laboratory, Badji-Mokhtar University, Algeria

Zielińska, Anetta, Department of Advanced Research in Management, Wroclaw University of Economics and Business, Poland

Żądło, Tomasz, Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Poland

STATISTICS IN TRANSITION new series, December 2024
Vol. 25, No. 4, pp. 213–218

Index of Authors, Volume 25, 2024

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 CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).