



STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

Deveci Kocakoç I., Köymen Keser I., Outlier detection based on the functional coefficient of variation

Adeboye N. O., Folorunso S. O., Abimbola O. V., Akinbo R. Y., Modelling the volatility of African capital markets in the presence of the COVID-19 pandemic: evidence from five emerging economies in Africa

Verma V., Nath D. C., Dwivedi S. N., Bayesian estimation of fertility rates under imperfect age reporting

Chwila A., The prediction of new COVID-19 cases in Poland with machine learning models

Nadeem M., Noreen K., Kashif Rasheed H. M., Ahmed R., Ul Hassan M., New generators for minimal circular generalised neighbour designs in blocks of two different sizes

Eideh A., On representativeness, informative sampling, nonignorable nonresponse, semiparametric prediction and calibration

Zielińska-Kolasińska Z., Zieliński W., A new confidence interval for the odds ratio

Kefelegn E., Determinants of livestock products export in Ethiopia

Bhushan S., Kumar A., On some efficient classes of estimators using auxiliary attribute

El Moury I., Hadini M., Chebir A., Mohamed B. A., Adil E., Proposal of a causal model measuring the impact of an ISO 9001 certified Quality Management System on financial performance of Moroccan service-based companies

Tomczyk E., Dynamics of survey responses before and during the pandemic: entropy and dissimilarity measures applied to business tendency survey data

Campanelli L., Breaking Benford's law: a statistical analysis of COVID-19 data using the Euclidean distance statistic

EDITOR

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

EDITORIAL BOARD

Dominik Rozkrut (Co-Chairman)	<i>Statistics Poland, Warsaw, Poland</i>
Waldemar Tarczyński (Co-Chairman)	<i>University of Szczecin, Szczecin, Poland</i>
Czesław Domański	<i>University of Łódź, Łódź, Poland</i>
Malay Ghosh	<i>University of Florida, Gainesville, USA</i>
Graham Kalton	<i>University of Maryland, College Park, USA</i>
Mirosław Krzysko	<i>Adam Mickiewicz University in Poznań, Poznań, Poland</i>
Partha Lahiri	<i>University of Maryland, College Park, USA</i>
Danny Pfeffermann	<i>Professor Emeritus, Hebrew University of Jerusalem, Jerusalem, Israel</i>
Carl-Erik Särndal	<i>Statistics Sweden, Stockholm, Sweden</i>
Jacek Wesołowski	<i>Statistics Poland, and Warsaw University of Technology, Warsaw, Poland</i>
Janusz L. Wywił	<i>University of Economics in Katowice, Katowice, Poland</i>

ASSOCIATE EDITORS

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

EDITORIAL OFFICE

ISSN 1234-7655

Scientific Secretary

Marek Cierpiał-Wolan, *Statistical Office in Rzeszów, Rzeszów, Poland*, e-mail: m.cierpial-wolan@stat.gov.pl

Managing Editor

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

Secretary

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

Technical Assistant

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

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



Address for correspondence

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

CONTENTS

Submission information for authors	III
From the Editor	VII

Research articles

Deveci Kocakoç I., Köymen Keser I., Outlier detection based on the functional coefficient of variation	1
Adeboye N. O., Folorunso S. O., Abimbola O. V., Akinbo R. Y., Modelling the volatility of African capital markets in the presence of the COVID-19 pandemic: evidence from five emerging economies in Africa	17
Verma V., Nath D. C., Dwivedi S. N., Bayesian estimation of fertility rates under imperfect age reporting	39
Chwila A., The prediction of new COVID-19 cases in Poland with machine learning models	59
Nadeem M., Noreen K., Kashif Rasheed H. M., Ahmed R., Ul Hassan M., New generators for minimal circular generalised neighbour designs in blocks of two different sizes	85
Eideh A., On representativeness, informative sampling, nonignorable nonresponse, semiparametric prediction and calibration	93
Zielińska-Kolasińska Z., Zieliński W., A new confidence interval for the odds ratio	113
Kefelegn E., Determinants of livestock products export in Ethiopia	129
Bhushan S., Kumar A., On some efficient classes of estimators using auxiliary attribute	141
El Moury I., Hadini M., Chebir A., Mohamed B. A., Adil E., Proposal of a causal model measuring the impact of an ISO 9001 certified Quality Management System on financial performance of Moroccan service-based companies	159
Tomczyk E., Dynamics of survey responses before and during the pandemic: entropy and dissimilarity measures applied to business tendency survey data	185

Research Communicates and Letters

Campanelli L., Breaking Benford's law: a statistical analysis of COVID-19 data using the Euclidean distance statistic	201
About the Authors	217

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 SiTns 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>.

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	JournalGuide
CEEOL – Central and Eastern European Online Library	JournalTOCs
CEJSH (The Central European Journal of Social Sciences and Humanities)	Keepers Registry
CNKI Scholar (China National Knowledge Infrastructure)	MIAR
CNPIEC – cnpLINKer	Microsoft Academic
CORE	OpenAIRE
Current Index to Statistics	ProQuest – Summon
Dimensions	Publons
DOAJ (Directory of Open Access Journals)	QOAM (Quality Open Access Market)
EconPapers	ReadCube
EconStore	RePec
Electronic Journals Library	SCImago Journal & Country Rank
Elsevier – Scopus	TDNet
ERIH PLUS (European Reference Index for the Humanities and Social Sciences)	Technische Informationsbibliothek (TIB) – German National Library of Science and Technology
Genamics JournalSeek	Ulrichsweb & Ulrich's Periodicals Directory
Google Scholar	WanFang Data
Index Copernicus	WorldCat (OCLC)
J-Gate	Zenodo

From the Editor

The March issue of Statistics in Transition new series presents readers with a set of twelve articles written by twenty-seven authors from ten countries (in order of appearance): Turkey, Nigeria, India, Poland, Pakistan, Sweden, Palestine, Ethiopia, Morocco, and Canada. This diversity is also reflected in the thematic variety of the real-life issues raised in these texts.

An attentive reader may notice that despite being the first issue published this year, it has the number 2. This is due to the fact that it was previously issued (as the first) of a special issue, prepared jointly with the Statistics of Ukraine, entitled *A New Role for Statistics*, which is devoted to the challenges – and ways of overcoming them – in the functioning of the official statistics system in the conditions of the war in Ukraine, along with often original solutions (methodological and organizational) in order to maintain the continuity of the production of necessary data and important indicators.

Research articles

In the first paper, **Ipek Deveci Kocakoç** and **Istem Köymen Keser** discuss ***Outlier detection based on the functional coefficient of variation***. The aim of the article is twofold: to show that the functional coefficient of variation is more sensitive to abrupt changes than the functional standard deviation and to propose the utilisation of the functional coefficient of variation as an outlier detection tool. Several simulation trials have shown that the coefficient of the variation function allows the effects of outliers to be seen explicitly. The coefficient of variation function is proposed as an outlier identification method in this study. The CV function is a better descriptive statistics for determining abrupt changes than standard deviation. The availability of the first and second derivatives of the CV function also strengthens its utilization. In the case of outliers in the data set, it is also proven to be a useful statistic. By using the one-out method, outlier curves can easily be detected among others. Therefore, the CV function may be utilized in outlier detection as a confirmatory and complementary method to different outlier detection methods such as outliergram and functional boxplot.

Nureni Olawale Adeboye, **Sakinat Oluwabukonla Folorunso**, **Olawale Victor Abimbola**, and **Rasaki Yinka Akinbo** present ***Modelling the volatility of African capital markets in the presence of the COVID-19 pandemic: evidence from five emerging economies in Africa***. The study employed Exponential Generalised Autoregressive Conditional Heteroscedasticity (EGARCH) procedures to develop

stock volatility models for the pre- and COVID-19 era. The Fixed-Effects Two Stage Least Square (TSLS) technique was utilised to establish an empirical relationship between capital market volatility and the COVID-19 occurrence based on equity market indices and COVID-19 reported cases of five emerging African economies: Nigeria, Egypt, South Africa, Gabon and Tanzania. The stock series was made stationary at the first order differencing and the model results indicated that the stock volatility of all the countries responded sharply to the outbreak of COVID-19 with the average stock returns of Nigeria and Gabon suffering the most shocks. Through empirical analysis, this article has exemplified and emphasized the impact of COVID-19 on the volatility of stock markets within the African continent.

The next article, by **Vivek Verma, Dilip C. Nath, and S. N. Dwivedi** entitled *Bayesian estimation of fertility rates under imperfect age reporting*, outlines the application of the Bayesian method of parameter estimation to situations where the probability of age misreporting is high, leading to transfers of an individual from one age group to another. An essential requirement for Bayesian estimation is prior distribution, derived for both perfect and imperfect age reporting. As an alternative to the Bayesian methodology, a classical estimator based on the maximum likelihood principle has also been discussed. As evident from the obtained results, even with inaccuracy in age reporting, the Bayesian technique has been found most promising for estimating TMFR, and obtained Bayes' estimates are more precise and reliable than those obtained using the maximum likelihood procedure. Apart from the estimation of transition probabilities, the Bayesian technique has been found to be more useful in estimating the pattern of fertility rates even in situations where there is inaccuracy in age reporting.

Adam Chwila's paper *The prediction of new COVID-19 cases in Poland with machine learning models* proposes several possible machine learning approaches to forecasting new confirmed COVID-19 cases, including the LASSO regression, Gradient Boosted (GB) regression trees, Support Vector Regression (SVR), and Long-Short Term Memory (LSTM) neural network. The above methods are applied in two variants: to the data prepared for the whole Poland and to the data prepared separately for each of the 16 voivodeships (NUTS 2 regions). The learning of all the models has been performed in two variants: with the 5-fold time-series cross-validation as well as with the split into the single train and test subsets. The computations in the study used official statistics from government reports from the period of April 2020 to March 2022. The machine learning models can help not only successfully predict different COVID-19 characteristics in the short-term periods, but also explain the factors that have the highest impact on the predictions for considered datasets.

Muhammad Nadeem, Khadija Noreen, H. M. Kashif Rasheed, Rashid Ahmed, and Mahmood Ul Hassan discuss *New generators for minimal circular generalised neighbour designs in blocks of two different sizes*. The authors described that the minimal neighbour designs (NDs) are used when a response of a treatment (direct effect) is affected by the treatment(s) applied in the neighbouring units. Minimal generalised NDs are preferred when minimal NDs cannot be constructed. Through the method of cyclic shifts (Rule I), the conditions for the existence of minimal circular generalised NDs are discussed, in which $v/2$ unordered pairs do not appear as neighbours. Certain generators are also developed to obtain minimal circular generalised NDs in blocks of two different sizes, where $k_2 = 3, 4$ and 5 . All these designs are constructed using i sets of shifts for k_1 and two for k_2 .

Abdulhakeem Eideh's article *On representativeness, informative sampling, nonignorable nonresponse, semiparametric prediction and calibration* focuses on modelling framework of the semi-parametric prediction of a finite population total while specifying the probability distribution of the response units under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. Furthermore, based on the relationship between response distribution and population distribution, the authors introduce a new measure of the representativeness of a response set and a new test of nonignorable nonresponse and informative sampling, jointly. Finally, a calibration estimator is obtained when the sampling design is informative and the nonresponse mechanism is nonignorable. The paper is purely mathematical and focuses on the role of informativeness of sampling design and informativeness of nonresponse in adjusting various predictors for bias reduction.

Zofia Zielińska-Kolasińska and Wojciech Zieliński in their paper *A new confidence interval for the odds ratio* deal with the problem of interval estimation of the odds ratio. The authors propose a new confidence interval for the odds ratio which is based on the exact distribution of the sample odds ratio, hence it works for large as well as for small samples. The coverage probability of that confidence interval is at least the nominal confidence level, in contrast to the asymptotic confidence intervals known in the literature. The information on the sample sizes and the sample odds ratio is sufficient for constructing the new confidence interval. Unfortunately, no closed formulae for the ends of the confidence interval are available. However, for given n_A , n_B and observed cOR the ends may be easily numerically computed with the aid of the standard software such as R, Mathematica, etc.

In the paper *Determinants of livestock products export in Ethiopia*, **Ermyas Kefelegn** proposes a procedure for identifying the determinants of the export of

Ethiopian livestock products using vector autoregressive and vector error correction models. Multivariate time series is used to model the association between the products of the Ethiopian livestock export included in the study. Vector autoregressive and vector error correction models are used for modelling and inference. The results indicated the existence of a long term correlation between the volume of live animals, meat and leather exports. The volume of meat export is significantly affected by a lag occurring in the export of live animals in the short-run. The quarterly data from 2002 to 2017 were tested for seasonality and results revealed that all of the series were not affected by periodicity. Moreover, unit root tests show that all four series were non-stationary in level, but stationary after first differencing. The long-run equation shows that the volume of live animals export has a positive long-run relationship with the volume of meat export.

The paper *On some efficient classes of estimators using auxiliary attribute* by **Shashi, Bhushan and Anoop Kumar** considers some efficient classes of estimators for the estimation of population mean using known population proportion. The classical ratio, and regression estimators suggested by Naik and Gupta (1996) and Abd-Elfattah et al. (2010) estimators are identified as the members of the suggested class of estimators. The expressions of bias and mean square errors are derived up to first-order approximation. The authors have introduced various novel classes of estimators by combining difference estimator with Srivastava, Walsh and Log type estimators and Srivastava type estimator with Walsh type estimator for estimating the population mean of study variable utilizing the information on an auxiliary attribute and compared them with the relevant contemporary estimators till date. The proposed classes of estimators are recommended for the estimation of population mean when information is available in the form of auxiliary attribute.

El Moury Ibtissam, Mohamed Hadini, Adil Chebir, Ben Ali Mohamed and Echchelh Adil present *Proposal of a causal model measuring the impact of an ISO 9001 certified Quality Management System on financial performance of Moroccan service-based companies*. The paper's goal is to test and validate a causal model designed to measure the performance of an ISO 9001 certified Quality Management System (QMS) and its impact on a company's financial performance. By means of this causal analysis (model), the study examines the relationship between: QMS and the financial performance of 41 companies based in Morocco, the management responsibility process and all the QMS processes, the management resources process and all the QMS processes, and the organisational and financial performance of the studied companies. All of the considered firms are part of the service industry and range from medium-sized to large companies. The data gathered in this study have been instrumental in devising actionable insights.

In the next paper, **Emilia Tomczyk** discusses *Dynamics of survey responses before and during the pandemic: entropy and dissimilarity measures applied to business tendency survey data*. The author starts with the question how to verify whether the pandemic of 2020–2022 can be seen as just another contraction phase. Entropy and dissimilarity measures are employed to study the characteristics of the expectations and assessments expressed in the business tendency survey of Polish manufacturing companies. The empirical results show that the dynamics of the manufacturing sector data, particularly as far as general economic conditions are concerned, set the pandemic period apart. The economic consequences of the COVID-19 pandemic expressed in business tendency surveys tend to be unfavourable, but the statistical properties or the degree of the concentration of respondents' answers do not correspond closely either to the expansion or contraction phases of the business cycle.

Research Communicates and Letters

In the *Research Communicates and Letters* section an article by **Leonardo Campanelli**, entitled *Breaking Benford's law: a statistical analysis of COVID-19 data using the Euclidean distance statistic* analyses the COVID-19 weekly case counts by country provided by the World Health Organization, (updated to December 20, 2021) using the Euclidean distance statistical test of Benford's law. The null hypothesis that the first-digit distribution of those counts follows Benford's distribution is tested using weekly confirmed cases instead of daily ones following the requirement of having counts that extended over many orders of magnitudes so to improve the compliance of the data sets with Benford's law. For the same reason daily and weekly death counts were not considered. Also, cumulative cases were not considered as their numbers flatten (especially at the end of a 'wave'), thus distorting relative digit frequencies.

Włodzimierz Okrasa

Editor

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



Outlier detection based on the functional coefficient of variation

Ipek Deveci Kocakoç¹, Istem Köymen Keser²

Abstract

The coefficient of the variation function is a useful descriptive statistic, especially when comparing the variability of more than two curve groups, even when they have significantly different mean curves. Since the coefficient of variation function is the ratio of the mean and standard deviation functions, its particular property is that it shows the acceleration more explicitly than the standard deviation function. The aim of the study is twofold: to show that the functional coefficient of variation is more sensitive to abrupt changes than the functional standard deviation and to propose the utilisation of the functional coefficient of variation as an outlier detection tool. Several simulation trials have shown that the coefficient of the variation function allows the effects of outliers to be seen explicitly.

Key words: coefficient of variation function, outlier detection, functional data analysis.

1. Introduction

The desire to interpret many curvilinear data together has increased the requirements for functional data analysis methods, and with the development of these methods, it is aimed to reveal the underlying structures of functional data called curves or surfaces. With the development of functional equivalents of multivariate statistical analysis techniques, functional data analysis (FDA) techniques have found a wide range of applications such as financial data (Wang et al., 2021), medical data (Ullah and Finch, 2013), climate change (Ghumman et al., 2020), air quality (Martinez Torres et al., 2020), management science (Dass and Shropshire, 2012), pandemics (Tang et al., 2020), etc. A detailed review of applications of functional data analysis can be found in Ullah and Finch (2013).

In functional data analysis, firstly, Ramsay (1982), Ramsay and Dalzell (1991), Rice and Silverman (1991), and Ramsay and Silverman (1997) introduced basic descriptive

¹ Corresponding Author. Econometrics Dept., Dokuz Eylül University, Izmir, Turkey. E-mail: ipek.deveci@deu.edu.tr. ORCID: <https://orcid.org/0000-0001-9155-8269>.

² Econometrics Dept., Dokuz Eylül University, Izmir, Turkey. E-mail: istem.koymen@deu.edu.tr. ORCID: <https://orcid.org/0000-0003-2123-188X>.



statistics such as the mean function, the standard deviation function, and variance-covariance surfaces, which form the basis of functional multivariate methods. Detailed information on descriptive statistics of functional data can be found in Shang (2015). Besides, Keser et al. (2016) have discussed the coefficient of variation (CV) function, which is the functional equivalent of the coefficient of variation. Krzysko and Smaga (2019) examined the multivariate coefficient of variation for functional data as a value, not as a function.

In this paper, our aim is to use the coefficient of variation function for detection of the effect of outliers by utilizing its sensitivity to abrupt changes in the data. The CV function is necessary to compare the variation between curve groups especially when the mean curves are different between curve groups or when we want to emphasize the main variation in time points. Suppose the height of boys and girls have different mean curves. If we want to compare the variation of height, we need the CV function. This study aims to use this feature of the CV function to detect abrupt changes caused by outliers in the data.

FDA has some methods for outlier detection which are extensions of classic statistical methods. Integrated squared error (Hyndman and Ullah, 2007), depth-based weighting and trimming (Febrero et al., 2007 and 2008), functional bagplot and functional highest density region (HDR) boxplot (Hyndman and Shang, 2010), trimmed estimators (Gervini, 2012), functional boxplot, and adjusted functional boxplot (Sun and Genton, 2011, 2012), robust functional principle component analysis procedure (Sawant et al., 2012), projection-based trimming (Fraiman and Svarc, 2013), outliergram (Arribas-Gil and Romo, 2014), and probabilistic modelling (de Pinedo et al., 2020) are prominent. Hubert et al. (2015) studied multivariate functional outlier detection. In this study, we propose the utilization of the coefficient of variation function for detection of the effect of outliers.

The paper is organized as follows: Section 2 presents the coefficient of variation function via the basis function approach. In Section 3, the results of simulation studies conducted to compare the standard deviation function and the coefficient of variation function in terms of sensitivity to abrupt changes are given. Outlier detection utilization of the CV function is explained in detail. Finally, Section 4 deals with some conclusions and suggestions.

2. Coefficient of variation function

Classical descriptive statistics for univariate data can similarly be applied to functional data with minor modifications. If the variability of multiple functional data groups needs to be compared, especially when mean functions of these data groups are

different, the coefficient of variation function can be used instead of the standard deviation function (Keser et al., 2016).

$$\text{CV function} = \frac{\text{Standard deviation function}}{\text{mean function}} \quad (1)$$

In comparison with other approaches, the basis function approach is mainly used in functional data analysis when estimating the curves. According to the basis function method, the estimates of mean, standard deviation, and the CV function are as follows.

Here, \mathbf{B} is the $(n \times K)$ basis function matrix which consists of $B_i(t_j)$, $i=1,2,\dots, K$, $j=1,2,\dots, n$ values, where t_j denotes j -th time point, K denotes the number of basis functions, and n denotes the number of curves. \mathbf{C} is the variance-covariance matrix of coefficients which are obtained by the roughness penalty method or the least sum of squares method. According to the basis function approach, the variance-covariance matrix for n curves is $\mathbf{V} = \mathbf{B}^* \mathbf{C} \mathbf{B}^T$.

i) Calculation of the coefficient vector of the standard deviation function (std):

As $s = \sqrt{\text{diag}(\mathbf{V})}$, $s = \mathbf{B}^* std$. In order to find std , the coefficient vector of the standard deviation function, the following calculations are carried out.

Since the \mathbf{B} matrix may not be a square matrix, it is converted into a square matrix by multiplying both sides by \mathbf{B}^T because of the inverse problem.

$$\mathbf{B}^T s = \mathbf{B}^T \mathbf{B}^* std$$

$\mathbf{B}^T \mathbf{B}$ matrix may not be invertible by the singular value decomposition. In this case, the Cholesky decomposition or Ridge regression may be used.

$$\mathbf{B}^T s = \mathbf{D}$$

$$\mathbf{B}^T \mathbf{B} = \mathbf{E}$$

While \mathbf{R} is an upper triangular matrix, according to the Cholesky decomposition:

$$\mathbf{E} = \mathbf{R}^T \mathbf{R}.$$

$$\mathbf{D} = \mathbf{R}^T \mathbf{R}^* std$$

$$(\mathbf{R}^{-1})^T \mathbf{D} = (\mathbf{R}^{-1})^T \mathbf{R}^T \mathbf{R}^* std$$

$$(\mathbf{R}^{-1})^* (\mathbf{R}^{-1})^T \mathbf{D} = (\mathbf{R}^{-1})^* (\mathbf{R}^{-1})^T \mathbf{R}^T \mathbf{R}^* std$$

So, the coefficient vector of the standard deviation function is

$$std = (\mathbf{R}^{-1})^* (\mathbf{R}^{-1})^T \mathbf{D}.$$

ii) Calculation of the coefficient vector of the mean function \bar{c} :

While \bar{y} is the mean coordinate vector, \bar{c} may be obtained in a similar way as the std :

$$\bar{c} = (\mathbf{B}^T \mathbf{B})^* \mathbf{B}^T \bar{y}.$$

iii) Calculation of the coefficient vector of the CV function:

We propose that the coefficients of the CV function be calculated as:

$$CV = \frac{\mathbf{B}^* std}{\mathbf{B}^* \bar{c}}$$

As denoted by Keser et al. (2016), especially when the curves have a mean function very close to zero, the CV function gets affected. This should be considered as a drawback of the statistic itself and exists in the original CV concept.

3. Simulation study

Since this CV function is a ratio of two functions, it is affected more by consecutive abrupt changes and can easily detect time points where essential changes of data occur. It is also proven by simulation studies in the next section that the CV function and its derivatives are as efficient as the standard deviation for detecting significant changes between time points for data with and without outliers.

In this study, we propose that CV functions may also be promoted as an outlier detection tool. Inspecting CV functions by using one-curve-out method may especially be used for this purpose. In order to show this usage of CV functions, outliers are investigated by CV functions and also by adjusted functional boxplots and adjusted outliergrams simultaneously.

3.1. The behaviour of CV function in abrupt changes

In this section simulation studies are conducted in order to compare the CV function with the standard deviation function for the detection of abrupt changes in functional data sets. For this purpose, n curves with 100 discrete points are generated by the main model $X(t) = \sin 4\pi t + \xi(t)$, $t \in [0, 1]$, where $\xi(t)$ is normally distributed with 0 mean and some covariance structure between $[0.2, 0.8]$. All analyses are done for $n=50$, 100 and 150 data sets. Since the size of the data set did not change the results, for the sake of easy interpretation and graphical readability, only $n=50$ case is reported here.

Computed descriptive statistics for the data set are given in Figure 1a and Figure 1b. The green lines show 50 curves, the red line shows the mean curve, the dotted line shows the standard deviation curve and the blue line shows the CV curve. Since the scales are not very compatible, presenting two different figures is preferred. The peaks on the CV function show the points where abrupt changes occur, which cannot be easily determined by the standard deviation function.

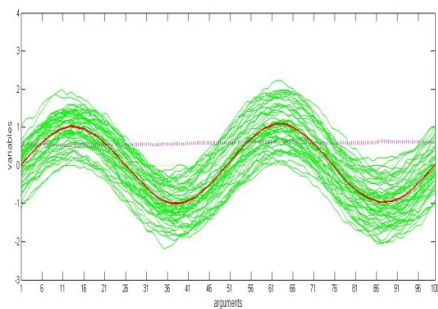


Figure 1a: Sample Data, mean and standard deviation function

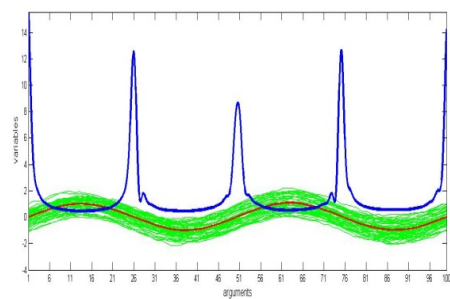


Figure 1b: Sample Data, mean and CV function

The time points in which the variability increases or decreases are not easily detectable in Figure 1a. Since CV is the proportion of two values, abrupt changes can be more easily detected especially when the range of standard deviation and mean functions are small. Here, the focus is not the value of the function at that point, but the abruptness of the situation. A partial view of CV values for one of the data sets is given in Table 1. It can be seen that there are abrupt changes for knots 1, 25, 26, 75,76, 99, and 100, which all have small changes in either their means or standard deviations, but distinctive CV values, as also can be seen from Figure 1b.

Table 1: A partial view of CV values for one of the data sets

Knots (Time points)	Std. Deviation coefficient	Mean coefficient	CV coefficient
1	0.4712	-0.0303	15.5499
2	0.4744	0.0951	4.9908
3	0.4735	0.2173	2.1794
.			
.			
.			
24	0.4606	0.1987	2.3177
25	0.4652	0.0748	6.2150
26	0.4602	-0.0369	12.4740
27	0.4568	-0.1611	2.8350
28	0.4675	-0.2741	1.7055
29	0.4826	-0.3855	1.2520
.			
.			
.			
73	0.5155	0.3025	1.7042
74	0.5128	0.1784	2.8746
75	0.5036	0.0397	12.6714
76	0.5064	-0.1005	5.0384
77	0.5115	-0.2339	2.1867
78	0.5221	-0.3560	1.4668
79	0.5145	-0.4744	1.0846
.			
.			
.			
97	0.5537	-0.3924	1.4109
98	0.5501	-0.2702	2.0363
99	0.5542	-0.1529	3.6237
100	0.5463	-0.0383	14.2749

Here, the CV function is apparently helpful for detecting abrupt changes in data. The same results are valid for data sets with outliers as can be seen in the next section.

3.2. The behaviour of CV function for data with outliers

There are two types of outliers in functional data analysis: magnitude outliers and shape outliers. In general, magnitude outliers appear far apart from other curves, and shape outliers have a distinct pattern from other curves. In this study, contamination models for outliers are chosen from Arribas-Gil and Romo's (2014) R file from their supplementary materials. In order to examine the behaviour of CV functions for data sets with and without outliers, outlier curves are generated with contamination models given below. Outlier curves are entered as the 51st curve after 50 non-outlier curves (generated from the main model) in the data set.

The contamination model for shape outlier (assuming that $\xi(t)$ is standard normally distributed) is:

$$X(t) = \cos(4\pi t - 0.25) + 0.05 \xi(t) .$$

The contamination model for magnitude outlier (assuming that $\xi(t)$ is standard normally distributed) is:

$$X(t) = \sin(4\pi t) + 2 + 0.05 \xi(t)$$

A representation of one sample data set from one of the trials is given in Figure 2 with 50 curves of non-outlier data, one shape, and one magnitude outlier.

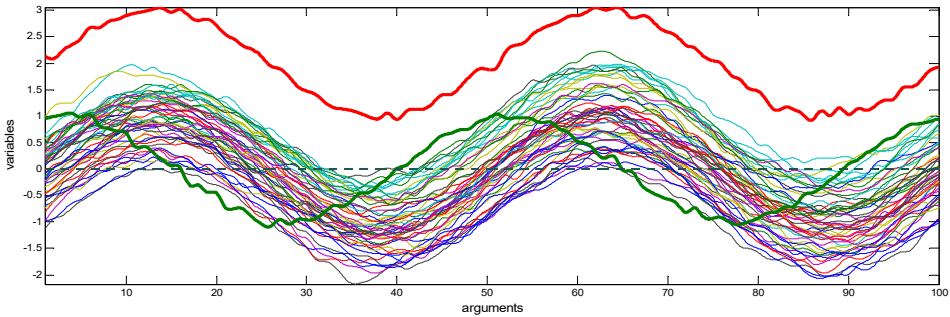


Figure 2: Data curves with shape and magnitude outliers
(Green bold line: shape outlier, Red bold line: magnitude outlier)

In our study, along with CV, Sun and Genton's (2011) adjusted functional boxplot and Arribas_Gill and Romo's (2014) adjusted outliergram are utilized simultaneously to detect outliers for practical purposes and reconfirmation. Graphs of those detection methods for the sample data set in Figure 2 are given in Figure 3.

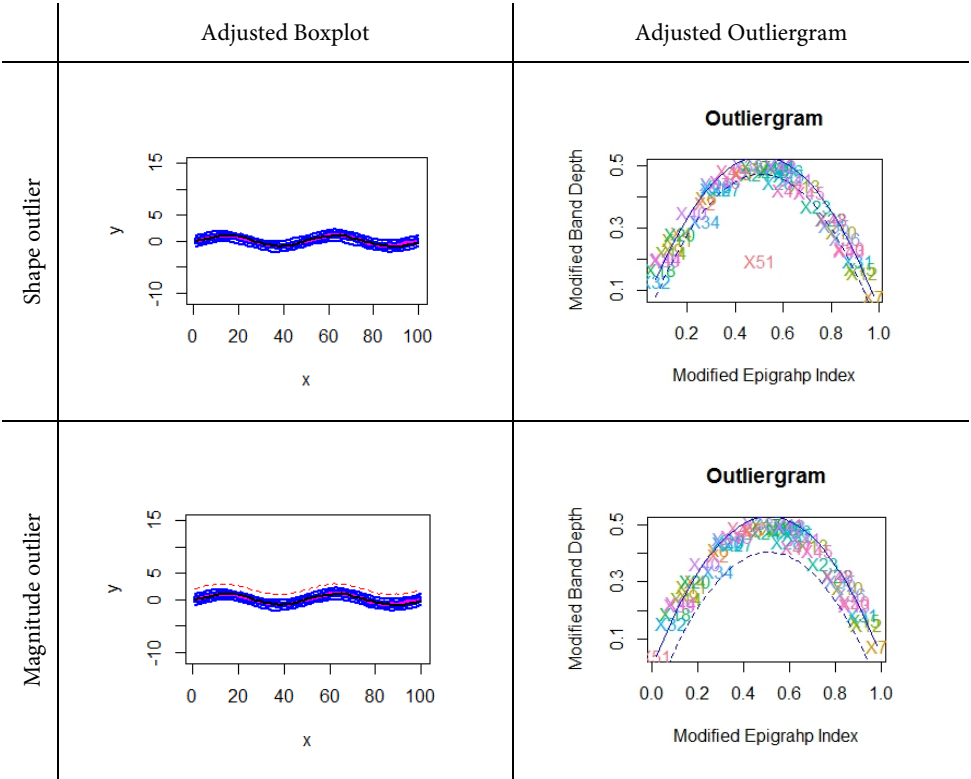


Figure 3: Outlier Detection by Adjusted Boxplot and Adjusted Outliergram

The functional adjusted boxplot is designed to detect outliers of magnitude, while the adjusted outliergram is designed to detect outliers of shape that are more difficult to find.

The descriptive statistics are the envelope of the 50% central region, the median curve, and the maximum non-outlying envelope for the adjusted functional boxplot (Sun and Genton, 2011), based on the centre outward ordering induced by band depth for functional data. By inflating the internal region (the envelope), the outer region (the fence) is obtained. Any curve that crosses the fences is depicted as possible outliers. For our sample data set, as expected, the adjusted functional boxplot easily detects magnitude outliers (dotted curve in the bottom left corner in Figure 3), however, shape outlier is not detected (top left corner in Figure 3).

Adjusted outliergram gives the boundaries for dashed parabola for the detection of shape outliers. Any value outside these boundaries is determined as an outlier. The further the curves in the sample are identical and straight, the nearer the points in the outliergram to the dashed parabola. On the other hand, the noisier the curves and the large number of crossing points between them, the more dispersed the points in the

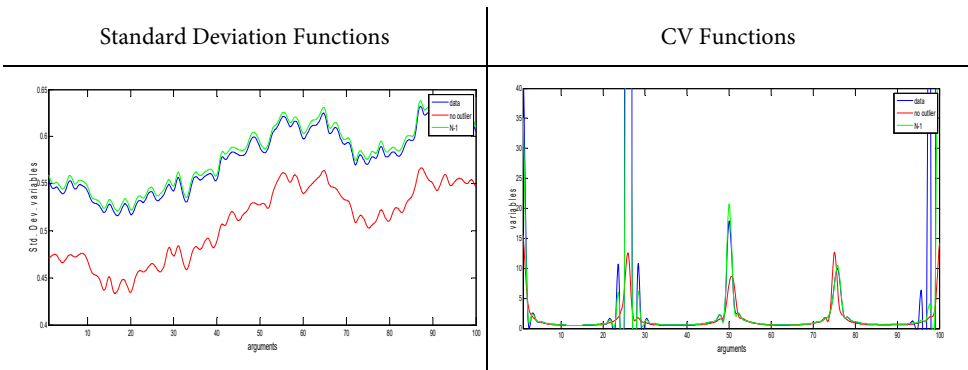
outliergram. The points with the largest distances to the parabola represent the most outlying curves, in terms of shape, of the sample (Arrabas Gil, Romo, 2014). For our sample data set, the 51st curve entered as the shape outlier lies out and far from the boundaries (top right corner in Figure 3) whereas the 51st curve entered as the magnitude outlier lies inside the parabola but far away from the other curves (bottom right corner in Figure 3).

Since the FDA is a visual method, many random data sets are generated from the models but, for clarity and conciseness, only three are randomly selected from the models in the study. Yet, it should be noted that all other data sets have similar results and may be provided by the authors. Changes in both standard deviation functions and CV functions are examined in every generated data set for the following three cases for both shape and magnitude outliers and the results are presented in Figure 4 and Figure 5 for those three selected data sets. Outliers are included separately in the data sets in order to avoid masking effect over each other.

- Case 1: Data with one outlier – (total: $N = 51$ curves, 51st curve is the outlier)
- Case 2: Data with one outlier when one of the non-outlier curves is excluded – (total: $N-1 = 50$ curves)
- Case 3: Data with no outliers – (total: $N-1 = 50$ curves)

In Case 2, every non-outlier curve is excluded one at a time. Since there are 50 curves, the computations are made 50 times for each data set and similar conclusions are made, so only one of the results is reported here.

Figure 4 shows results for data sets with one magnitude outlier and Figure 5 shows results for data sets with one shape outlier. In both figures, the left panel of each row shows the standard deviation function for three randomly selected cases while the right panel shows the proposed the coefficient of variation function. The same analysis is conducted also with normalized data but no difference is found and therefore not reported here. Absolute values of CV are used for reflecting the variability better and easier comparison with standard deviations.



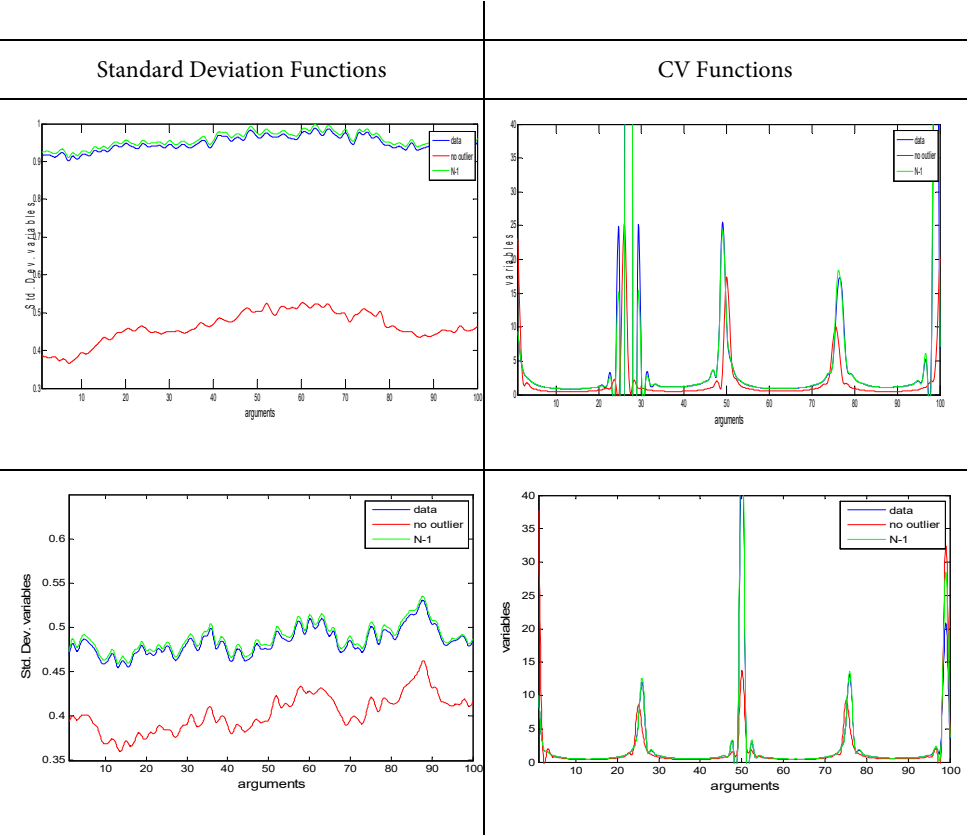
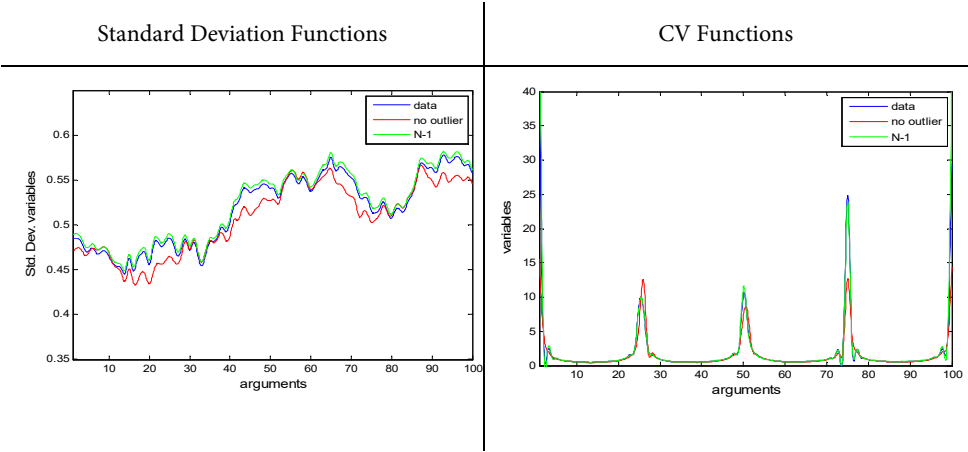


Figure 4: Standard Deviation and CV Functions for Magnitude Outlier for Randomly Selected Cases



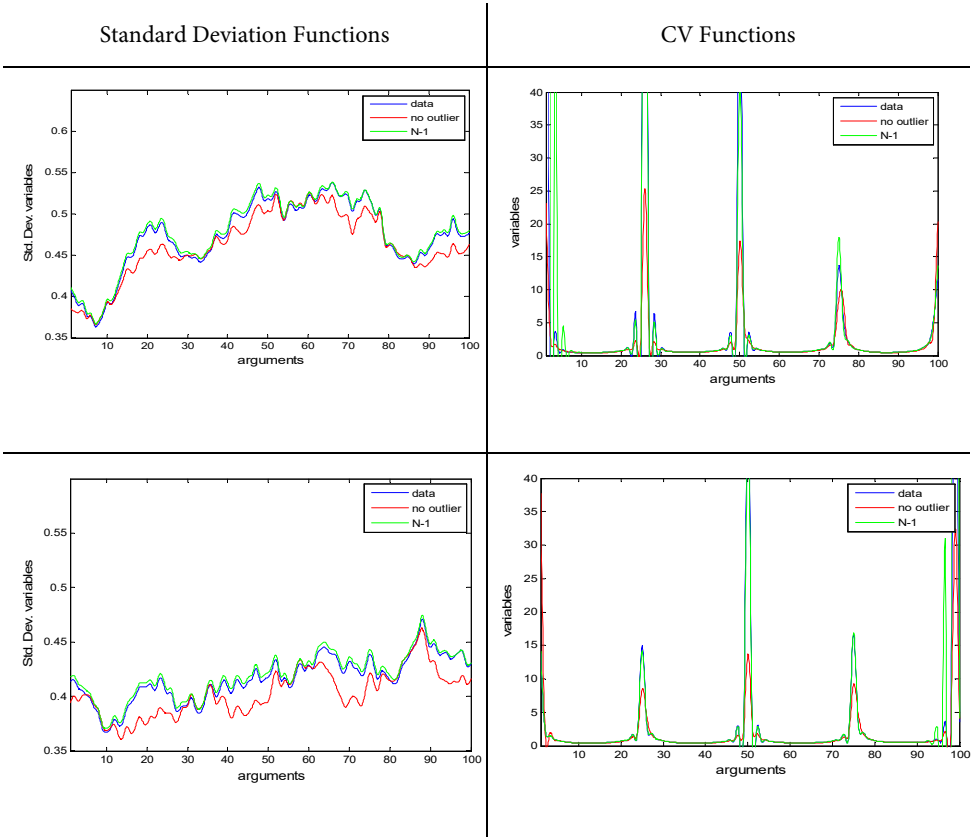


Figure 5: Standard Deviation and CV Functions for Shape Outlier

When examining standard deviation functions for all data sets, it can be seen that the standard deviation function of data with outlier lies distant from the standard deviation function of non-outlier data, as expected.

When examining CV functions for all data sets, it can mostly be seen that the CV function of data with outliers lies distant from the CV function of data without outliers similar to the results for the standard deviation function. When any one of the non-outlier curves are excluded (Case 2), both standard deviation and CV functions have very close results to data with outliers. This may be a supporting result that outlier curves affect standard deviation and CV functions.

Since cubic B-splines are used to obtain the standard deviation and CV functions, the first and second derivative functions of these functions are also continuous. Therefore, the movements of the curves can easily be examined. Interpretations of derivative functions rather than the functions themselves may provide a stronger inference. Besides, utilizing derivative functions when comparing the curves makes them more comparable with respect to the origin. The first and second derivative

functions of standard deviation and CV functions for magnitude outliers are given in Figures 6 and 7 respectively. The first derivative function shows the velocity while the second one shows the acceleration.

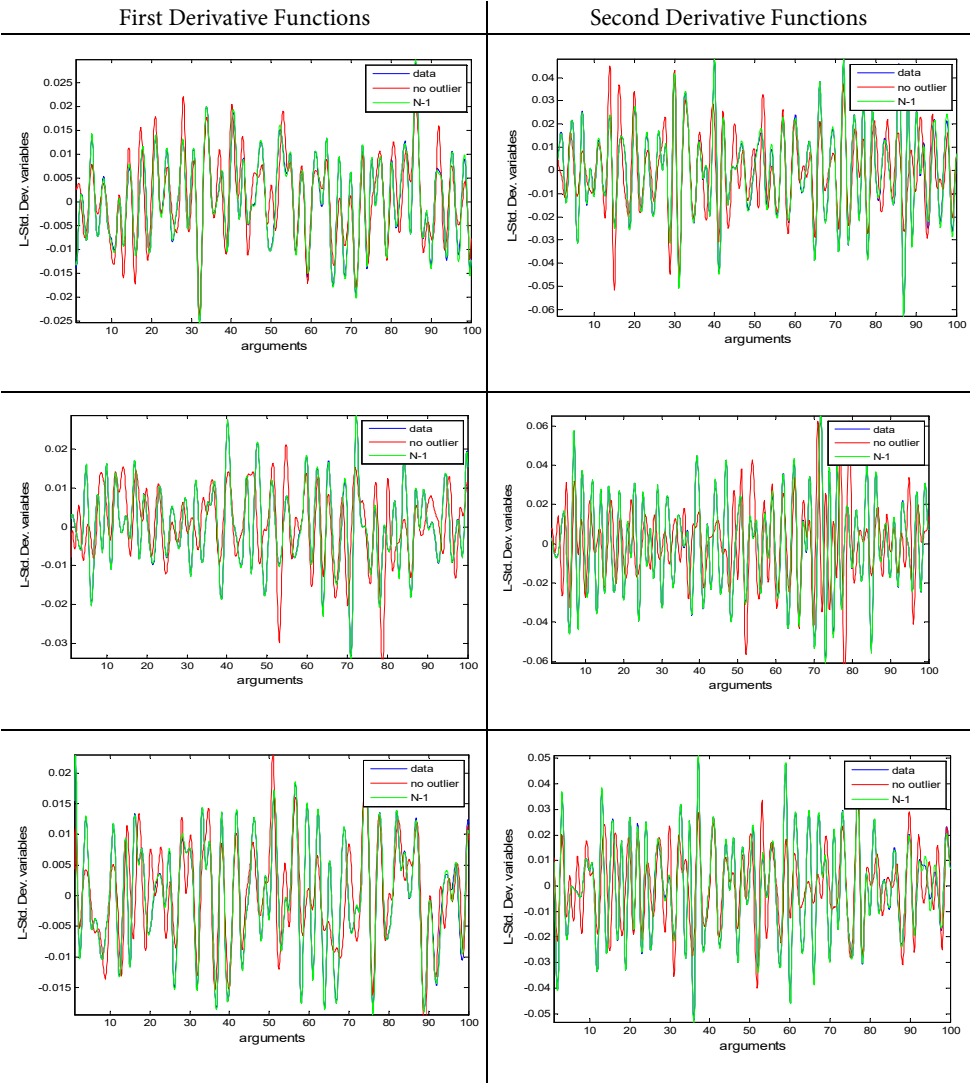


Figure 6: Derivative functions of standard deviation function for magnitude outlier

When Figure 6 is examined, data with outliers and data with (N-1) curves show a very similar, even overlapping, behaviour for both the first and second derivative functions while non-outlier data lies distantly and with shifts. By utilizing derivative functions, the dimensions of ups and downs have become more comparable.

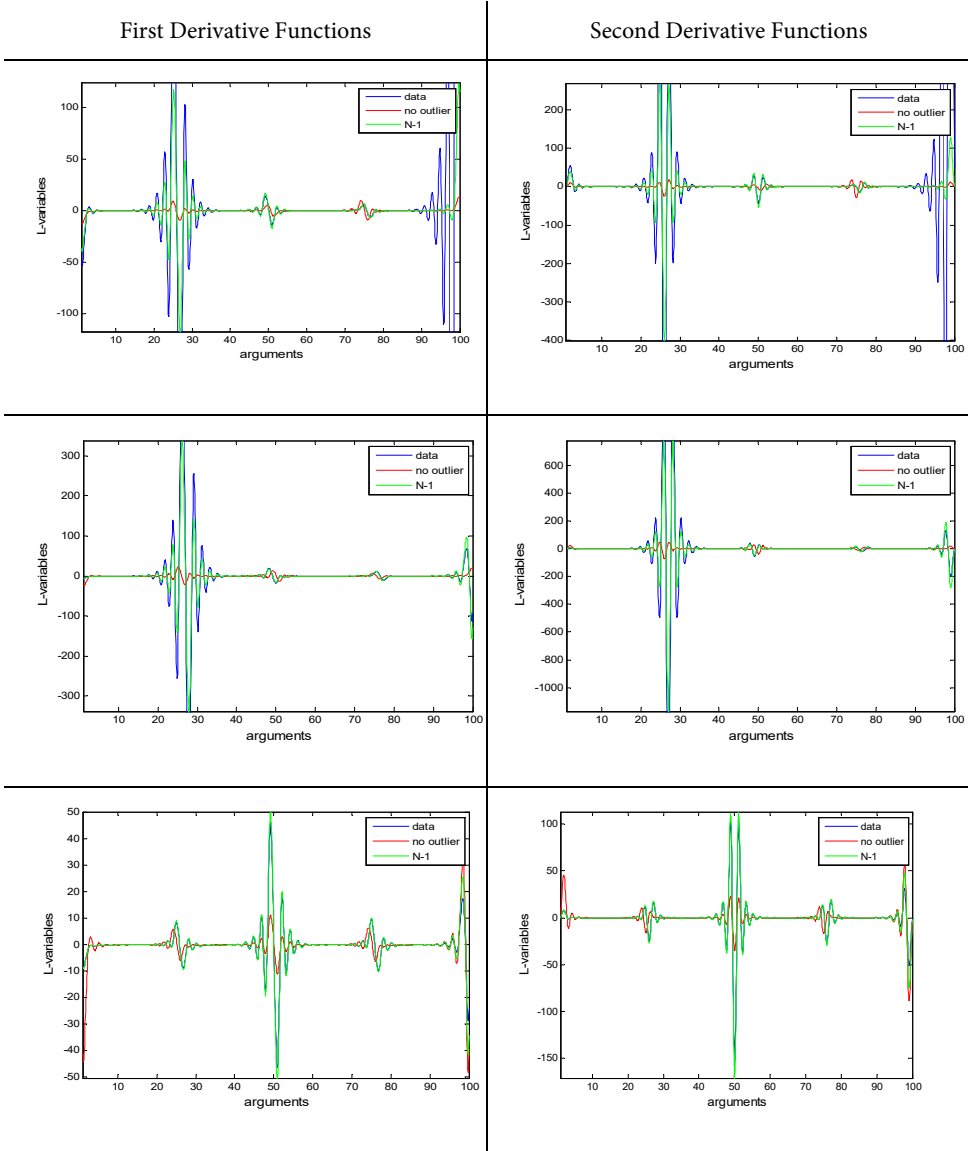


Figure 7: Derivative functions of CV function for magnitude outlier

When Figure 7 is examined, it can be seen that CV especially emphasizes time points in which abrupt changes appear. As in Figure 6, both derivative functions for non-outlier data lie distant from the other two data sets with outliers. Thus, it can be said that derivative functions have similar behaviour as the original curve functions. However, the derivative functions of the CV function do not get lost in small changes and can focus on abrupt changes better than that of the standard deviation function. The standard deviation function and its derivatives are strongly affected by the smallest

changes due to the effect of the mean. Examination of derivative functions enables better comparisons for ups and downs in all curves and even small changes can be determined by their help.

The behaviour of the CV function for data sets with and without outliers leads way to the one-out method as an outlier detection tool by itself. Since now we know that outliers affect the size of peaks in CV functions, any outlier curve can be detected by its different size of the CV function peaks. In order to show that, we examined 51 CV functions obtained by excluding one curve at a time and saw that the 51st curve has a very different peak structure than the others (especially for magnitude outlier), concluding that this curve may be an outlier.

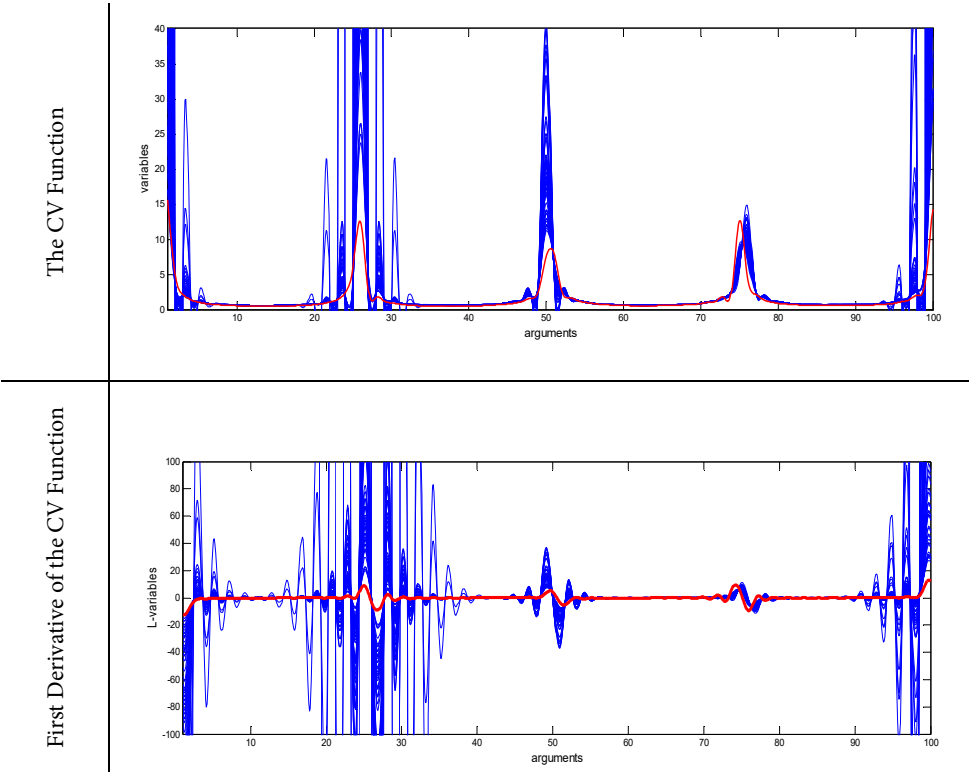


Figure 8: CV function and its first derivative for the one-out method

In Figure 8, the blue lines show the 50 curves which include the outlier, and the red curve shows the function when the outlier curve is excluded. Here, we also validate and confirm the outliers that we found by adjusted outliergram and adjusted functional boxplot. Therefore, the CV function can be utilized as a visual outlier curve detection tool.

4. Conclusions

The coefficient of variation function is proposed as an outlier identification method in this study. The CV function is a better descriptive statistics for determining abrupt changes than standard deviation. The availability of the first and second derivatives of the CV function also strengthens its utilization. In the case of outliers in the data set, it is also proven to be a useful statistic. By using the one-out method, outlier curves can easily be detected among others. Therefore, the CV function may be utilized in outlier detection as a confirmatory and complementary method to different outlier detection methods such as outliergram and functional boxplot.

More automated and easy detection of points with abrupt changes and curves which are outliers may be developed as a further study. Confidence intervals or probabilities for this detection may also be investigated.

References

- Arribas-Gil, A., Romo, J., (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, 15, pp. 603–619.
- de Pinedo, Á. R., Couplet, M., Marie, N., Marrel, A., Merle-Lucotte, E., and Sueur, R., (2020). Functional outlier detection through probabilistic modelling. In: Aneiros G., Horová I., Hušková M., Vieu P. (eds) *Functional and high-dimensional statistics and related fields*. IWFOS 2020. Contributions to Statistics. Springer, Cham. https://doi.org/10.1007/978-3-030-47756-1_30.
- Dass, M., Shropshire, C., (2012). Introducing functional data analysis to managerial science. *Organizational Research Methods*, 15(4), pp. 693–721.
- Febrero, M., Galeano, P., and Gonzalez-Manteiga, W., (2007). A functional analysis of NO_x levels: Location and scale estimation and outlier detection. *Computational Statistics*, 22(3), pp. 411–427.
- Febrero, M., Galeano, P., and Gonzalez-Manteiga, W., (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NO_x levels. *Environmetrics*, 19(4), pp. 331–345.
- Fraiman, R., Svarc, M., (2013). Resistant estimates for high dimensional and functional databased on random projections. *Computational Statistics & Data Analysis*, 58, pp. 326–338.
- Gervini, D., (2012). Outlier detection and trimmed estimation for general functional data. *Statistica Sinica*, 22(4), pp. 1639–1660.

- Ghumman, A. R., Alodah, A., Haider, H., and Shafiquzzaman, M., (2020). Evaluating the impact of climate change on stream flow: integrating GCM, hydraulic modelling and functional data analysis. *Arabian Journal of Geosciences*, 13(17), pp. 1–15.
- Hubert, M., Rousseeuw, P. J., and Segaert, P., (2015). Multivariate functional outlier detection. *Statistical Methods & Applications*, 24(2), pp. 177–202.
- Hyndman, R. J., Ullah, M. S., (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10), pp. 4942–4956.
- Hyndman, R. J., Shang, H. L., (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, 19, pp. 29–49.
- Krzysko, M., Smaga L., (2019). A multivariate coefficient of variation for functional data, *Statistics and Its Interface*, 12(4), pp. 647–658.
- Keser, İ. K., Kocakoç, İ. D., and Şehirlioğlu, A. K., (2016). A new descriptive statistic for functional data: functional coefficient of variation. *Alphanumeric Journal*, 4(2), pp. 1–10.
- Martínez Torres, J., Pastor Pérez, J., Sancho Val, J., McNabola, A., Martínez Comesaña, M., and Gallagher, J., (2020). A functional data analysis approach for the detection of air pollution episodes and outliers: A case study in Dublin, Ireland. *Mathematics*, 8(2), p. 225.
- Ramsay J. O., (1982). When the data are functions. *Psychometrika*, 47, pp. 379–396.
- Ramsay, J. O., Dalzell, C. J., (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3), pp. 539–561.
- Ramsay J. O, Silverman B. W., (1997). *Functional data analysis*. Springer-Verlag, New-York.
- Rice John A., Silverman B. W., (1991). Estimating the mean and covariance structure when the data are curves. *Journal of the Royal Statistical Society, Series B.*, Vol. 53, No.1, pp. 233–243.
- Ullah S., Finch C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(43), pp. 539–572.
- Sawant, P., Billor, N., and Shin H., (2012). Functional outlier detection with robust functional principal component analysis. *Computational Statistics*, 27, pp. 83–102.

- Shang, H. L., (2015). Resampling techniques for estimating the distribution of descriptive statistics of functional data. *Communications in Statistics – Simulation and Computation*, 44(3), pp. 614–635, doi: 10.1080/03610918.2013.788703.
- Sun Y., Genton M. G., (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20, pp. 316–334.
- Sun, Y., Genton M. G., (2012). Adjusted functional boxplots for spatio-temporal data visualization and outlier detection. *Environmetrics*, 23(1), pp. 54–64.
- Tang, C., Wang, T., and Zhang, P., (2020). Functional data analysis: An application to COVID-19 data in the United States, *arXiv preprint arXiv:2009.08363*.
- Wang, D., Li, X., Tian, S., He, L., Xu, Y., and Wang, X., (2021). Quantifying the dynamics between environmental information disclosure and firms' financial performance using functional data analysis. *Sustainable Production and Consumption*, 28, pp. 192–205.

Modelling the volatility of African capital markets in the presence of the Covid-19 pandemic: evidence from five emerging economies in Africa

Nureni Olawale Adeboye¹, Sakinat Oluwabukonla Folorunso²,
Olawale Victor Abimbola³, Rasaki Yinka Akinbo⁴

Abstract

The growing concern over the global effects of the COVID-19 pandemic on every aspect of human endeavour has necessitated a continuous modelling of its impact on socio-economic phenomena, allowing the formulation of policies aimed at sustaining future economic growth and mitigating the looming recession. The study employed Exponential Generalised Autoregressive Conditional Heteroscedasticity (EGARCH) procedures to develop stock volatility models for the pre- and COVID-19 era. The Fixed-Effects Two Stage Least Square (TSLS) technique was utilised to establish an empirical relationship between capital market volatility and the COVID-19 occurrence based on equity market indices and COVID-19 reported cases of five emerging African economies: Nigeria, Egypt, South Africa, Gabon and Tanzania. The stock series was made stationary at the first order differencing and the model results indicated that the stock volatility of all the countries responded sharply to the outbreak of COVID-19 with the average stock returns of Nigeria and Gabon suffering the most shocks. The forecast values indicated a constant trend of volatility shocks for all the countries in the continuous presence of the COVID-19 pandemic. Additionally, the confirmed and death cases of COVID-19 were found to increase stock prices while recovered cases bring about a reduction in the stock prices in the studied periods.

Key words: African countries, capital market, COVID-19, volatility, GARCH model.

1. Introduction

Capital market is one of the major pilots of fiscal growth and economic development, of which its activities have been a daily occurrence save for non-working

¹ Department of Statistics, Osun State University, Osogbo, Nigeria. E-mail: nureni.adeboye@uniosun.edu.ng.
<https://orcid.org/0000-0002-8023-221X>

² Department of Mathematical Sciences, Olabisi Onabanjo University, Ago Iwoye, Nigeria.
E-mail: sakinat.folorunso@oouagoiwoye.edu.ng.

³ Data Science Nigeria, Nigeria. E-mail: olawale@datasciencenigeria.ai.

⁴ Department of Mathematical Sciences Federal Polytechnic Ilaro, Nigeria.
E-mail: rasaki.akinbo@federalpolyilaro.edu.ng.



days (Adeboye and Fagoyinbo, 2017; Olowe, 2009). Globally, there is an increasing reliance on stock trading data as a fundamental tool for making reliable investment decisions and the Africa case is not an exception. Therefore, it is a fundamental reality that the capital market constitutes one of the major pitfalls of the Covid-19 outbreak. The capital market can be an extremely volatile place with broad day-to-day swings that present a significant investment risk. In modelling stock market time series data, the presence of long memory is very obvious. The behaviours of the investors are influenced, which can make their decisions based on different investment horizons. Stock market data were found to exhibit characteristics that are more consistent with long memory (Baillie, 1996). However, national and regional economic factors like interest rate policies, inflation trends, tax etc. have been found to substantially contribute to the directional change of the market, thus portend greater potentials to influence volatility. Capital market volatility is a statistical evaluation of the variation in returns for a given stock or market index. According to Chiang and Dong (2001), a higher level of volatility appears to be associated with higher average returns in most cases, however, unexpected volatility that is adverse could spell doom for would-be investors. This has been a contentious issue in wealth creation across the globe with a surfeit of literature on how to mitigate its effect and its impact on investment drives and economic activities of every nation (Ser-Huang and Taylor, 1992; Pramod and Puja, 2015). However, according to Ayinde et al. (2020) and Oyelola et al. (2020), the global occurrence of coronavirus with its antecedent daily upsurge in the count of confirmed cases around the globe is becoming alarming, and according to the NSE (2020) report, one can say it has nearly shut down the capital market with almost all the known bullish stocks experiencing a worsening rate of volatility. Weltman (2020) opined that risk experts in financial matters have made significant efforts to rearrange their market appraisal in light of the unprecedented economic challenges posed by the Covid-19 crisis that has put nearly all the globe in virtual lockdown. In the same Euromoney report, M. Nicolas Firzli refers to the Covid-19 effect on the financial market as the peak of all financial crisis and opined that it is bringing to the fore many repressed financial and geopolitical disorders. As highlighted by the Financial Times Stock Exchange and Dow Jones Industrial Average reports, 100 companies listed on the London stock exchange with the highest market capitalization dropped beyond three percent as COVID-19 outbreak worsened and spread beyond China. On February 27th 2020, due to persistent concerns posed by the coronavirus outbreak, most United States capital market indices indicated the sharpest declines since 2008. On the overall, capital markets declined beyond 30% as at March 2020 implied volatilities of equities have spiked to crisis levels; credit spreads on non-investment grade debt have widened sharply as investors attempt to reduce risks (Barron's, 2020). This uncertainty in global financial markets is occurring despite the elaborate financial reforms conceptualized by

G-20 financial authorities in the post-crisis era. According to OECD (2020) interim economic outlook highlighted in early March 2020, Covid-19 had already worsened China economic growth, and subsequent outbreaks in other continents were eroding prospects for economic growth. Hitherto, governments of countries have introduced unprecedented measures to contain the epidemic. While it is of necessity to contain the virus, it is of note that measures involved have led to both socio and economic quagmire in the countries mostly affected. Thus, the shutdowns could lead to high declination in the level of economic development, thereby causing most consumers' expenditure to be adversely affected. The magnitude of these occurrences would far outweigh the economic recession experienced during the global financial meltdown if the situations persist for too long.

Presently, COVID-19 remains the most traumatic pandemic threatening the entire globe. According to Adebayo et al., 2020, the first COVID-19 confirmed case in Africa was reported on 14th February 2020 in Egypt, which has been chosen to represent the North Africa region in this research. Since then, the number of reported cases has experienced geometric increases. The United States Embassy in Egypt gave the statistics as 90,413 confirmed cases with 4,480 deaths as of July 23rd, 2020. Nigeria's first case was reported on 27th February 2020, when an Italian citizen in Lagos tested positive for the virus (NCDC, 2020; MacLean et al. 2020). Ayinde et al. (2020) gave the statistics of Nigeria coronavirus as 1,932 confirmed cases, 319 discharged cases, and 58 deaths as of 30th April 2020. These records have experienced daily increase with the total number of confirmed cases in Nigeria now stood at 41,180, of which 860 deaths have been recorded as of 28th July 2020 according to WHO (2020) coronavirus global updates. The WHO reports for African regional office also confirmed South Africa to be the epicentre of the COVID-19 outbreak in Africa region with 45,9761 positive cases identified and 7,257 recorded deaths as of 28th July 2020. South Africa is presently ranked fifth in the whole world. Gabon and Tanzania were confirmed to have recorded their first COVID-19 pandemic in March 2020. Up to 5,087 COVID-19 cases were reported in Gabon as of June 24 with a death toll of 40. As for Tanzania, government authorities ceased all attempts of reporting COVID cases in May 2020 after President John Magufuli alleged that laboratories were giving out fake results of confirmed cases.

WHO records on Covid-19 as of 30th July 2020 indicated 17,039,160 confirmed cases and 667,084 deaths globally. The spread of the disease is so alarming to the extent that the United States of America have recorded 4,427,493 confirmed cases and 150,716 deaths as of date, despite the existence of a well-structured medical system in place. These statistics is closely followed by South America with 2,983,227 confirmed cases, 108,432 deaths, and 2,002,553 recoveries; Europe was reported to have 1,596,917 confirmed cases and 578,319 deaths according to ECDC and CDC (2020) report. Although African continent still has the least records of 625,562 confirmed

cases, 13,763 deaths, and 193,481 recoveries according to NBS (2020) report accessed on July 17th, 2020, the socio-economic effect of the pandemic had been so alarming and has further worsened the living standard of the citizenry in Africa (Adhikari et al., 2020). Though the Covid-19 pandemic has been widely acclaimed to have originated from Wuhan city, Hubei China (Giordano et al., 2020; Nadeem, 2020; Huang et al., 2020; Adegboye et al., 2020; Guo et al., 2019). However, in their study, Adegboye et al. (2020) emphasized that the risk of importing the pandemic from Europe to Africa exceeded that of importation from China. Martinez-Alvarez et al. (2020) compared early transmission of the pandemic in selected countries and observed a more rapid spread of the virus in some West African countries than in Europe. Gilbert et al., (2020); Vladimir and Vasily (2020) opined that situation in African countries could be more fatal than what is being reported, as most of African countries are unprepared and not sufficiently capable in the management of disease outbreak. Thus, the need to model capital market volatility as occasioned by the pandemic, to serve as the impetus to project new normal for capital market businesses in the African region and the world at large.

This study investigates the capital market volatility for the pre-COVID-19 era and compared with the activities during the Covid-19 pandemic using equity market indices and Covid-19 reported cases of five (5) emerging economies in Africa. One prominent and economically endowed country was chosen from each of the regions in the African continent. These countries are Nigeria (West Africa), South Africa, Egypt (North Africa), Tanzania (East Africa), and Gabon (Central Africa). We use Exponential Generalized Autoregressive Conditional Heteroscedasticity (EGARCH) to model the stock volatility for each of the African countries in the two periods. Fixed-Effects Two Stage Least Square (TSLS) is then employed to model the coronavirus impact on the market activities. EGARCH and TSLS models would be capable of providing forecasts to predict the market volatility in the continuous presence of the pandemic. Currently, most literature on stock market volatility adopted regional and national economic factors approaches for its modelling. For instance, see Ser-Huang & Taylor (1992); Theodore & Lewis (1992); Blinder & Merges (2001); Chiang & Diong (2001); Grundy & Kim (2002); Pramod & Puja (2015). The contribution of this paper is the modelling of the Covid-19 trajectory in the day-to-day volatility of stock market affairs in the African continent.

2. Materials and Methods

The data set used for this research is for the selected countries from the major sub regions of African continent. Stock updates for the five countries under consideration were sourced from Yahoo Finance while COVID-19 data were sourced from different

legitimate sources such as World Health Organization (WHO), European Centre for Disease Prevention and Control (ECDC), Nigeria Center for Disease Control (NCDC), Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE), UNICEF and US embassy in African countries. The time series observations of daily returns of stock exchange closing prices between 2019–10–01 to 2020–02–28 (pre-COVID era) and 2020–03–02 to 2021–04–30 (during COVID-19) for the five selected African countries under study, utilized for this research are available as supplementary data shared on GitHub repository link <https://bit.ly/37LqPkw>. This supplementary data also includes information of the reported Covid-19 cases for the regions. However, some of the observations were no longer available after being subjected to differencing in order to attain stationarity. As a result of these omitted lagged data points, data imputation techniques as suggested by Olalekan et al. (2020) were employed to obtain the missing observation so as to be able to fit the EGARCH model on the data set.

2.1. Volatility Model (Exponential Generalized Autoregressive Conditional Heteroscedasticity)

E-GARCH is a family of the GARCH model. The E-GARCH model was proposed by Nelson (1991) to overcome the challenges of volatility clustering in the handling of GARCH for modelling financial time series.

Let ε_t denote the error term of a time series $\{X_t\}$. If the typical size of ε_t is characterized by stochastic piece u_t and time-dependent standard deviation σ_t , then

$$\varepsilon_t = u_t \sigma_t \quad (1)$$

where the stochastic piece u_t is a strong white noise process and the time-dependent variance can be expressed as

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2, \alpha_0 > 0 \quad (2)$$

$$\sigma_t^2 = \alpha_i + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 \quad \alpha_i > 0, i > 0 \quad (3)$$

where p is the length of ARCH lags.

Considering the Autoregressive Moving Average Model $[ARMA_{(pq)}]$ given as

$$X_t - \alpha_1 X_{t-1} - \cdots - \alpha_p X_{t-p} = \varepsilon_t - \beta_1 \varepsilon_{t-1} - \cdots - \beta_q \varepsilon_{t-q} \quad (4)$$

If equation (4) is assumed for the error variance, then we have

$$X_t = X'_{t-p} b + \varepsilon_t \quad (5)$$

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \cdots + \alpha_p \varepsilon_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2 \quad (6)$$

Thus, equation (3) can be written for $ARMA_{(pq)}$ as

$$\sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^q \beta_j \sigma_{t-j}^2 \quad (7)$$

$$\ln \sigma_t^2 = \omega + \sum_{i=1}^p \alpha_i (|\varepsilon_{t-i}| - E|\varepsilon_{t-i}|) + \gamma_i \varepsilon_{t-1} \sum_{j=1}^q \beta_j \ln \sigma_{t-j}^2 \quad (8)$$

Equation (7) is the generalized autoregressive conditional heteroskedasticity [$GARCH_{(pq)}$] model while equation (8) is the $EGARCH_{(pq)}$ where p is the order of the ARCH component; q is the order of the GARCH component; σ_t is the volatility at time t ; ω is the intercept, $\alpha_1, \dots, \alpha_p$ are the parameters of the ARCH component; $\beta_1, \beta_2, \dots, \beta_q$ are the parameters of the GARCH component model; γ is the magnitude of the shock and ϵ_t is a zero mean white noise as. It is pertinent to note that there are no sign restrictions for the EGARCH parameters since $\ln \sigma_t^2$ can be negative. All the parameters ($\mu, \omega, \alpha, \gamma, \beta$) are estimated simultaneously by maximizing the log likelihood, where μ is the expected shares return.

2.2. Model Specification

The time-series econometric model specified for this research is given as

$$Stock_{it} = f(Confirmed_{1,it}, Recovered_{2,it}, Death_{3,it}) + \epsilon_{it} \quad (9)$$

When this model is written explicitly, it becomes

$$Stock_{it} = \beta_0 + \beta_1(Confirmed)_{1,it} + \beta_2(Recovered)_{2,it} - \beta_3(Death)_{3,it} \quad (10)$$

The instrumental variables specified for this model are confirmed, death and recovered variables. The lagged cases variables, i.e. d_confirmed, d_death and d_recovered were excluded from the list of instruments since they are endogenous variables and thus correlated with the residuals.

2.3. Fixed-Effects Two Stage Least Square (TSLS)

TSLS is a fixed effect model which is a special case of instrumental variables regression. In this model, there are two distinct stages of which the first stage involves finding the portions of the endogenous and exogenous variables that can be attributed to the instruments. The stage involves estimating an OLS regression of each variable in the model on the set of instruments while the second stage is a regression of the original equation, with all of the variables replaced by the fitted values from the first-stage regressions. The coefficients of this regression are the TSLS estimates.

By denoting Z as the matrix of instruments, y and X as the dependent and explanatory variables respectively, the computed coefficients and its covariance matrices are given by the equations

$$b_{TSLS} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \quad (11)$$

$$\hat{\Sigma}_{TSLS} = s^2(X'Z(Z'Z)^{-1}Z'X)^{-1} \quad (12)$$

where S^2 is the estimated residual variance.

The strategy for estimation involved taking deviations of the group means to have

$$y_{it} - \bar{y}_i = (x_{1it} - \bar{x}_{1i})\beta_1 + (x_{2it} - \bar{x}_{2i})\beta_2 + \varepsilon_{it} - \bar{\varepsilon}_i. \tag{13}$$

3. Results and Discussion

All the indices used for the E-Garch modelling were pulled from the share returns of five (5) African countries considered in this research, for pre and post COVID-19 era. Furthermore, the share returns of all the countries were merged with the daily cases, recoveries and deaths due to COVID-19.

3.1. Descriptive Statistics

Table 1: Descriptive Statistics of Stock Returns of the Considered Countries before Covid-19

Index	Nigeria	South Africa	Tanzania	Gabon	Egypt
Mean	-0.00120	-0.00049	-0.0039	-0.00112	0.00040
Median	-0.00159	0.00117	0.0000	0.00000	0.00000
Minimum	-0.06580	-0.05855	-0.1923	-0.03396	-0.08898
Maximum	0.04702	0.03027	0.1569	0.02290	0.09483
Standard Dev	0.01448	0.01467	0.0375	0.00896	0.02690
Skewness	-0.44	-1.1	-1.2	-0.62	0.41
Kurtosis	7.3	5.5	14	4.8	5.1

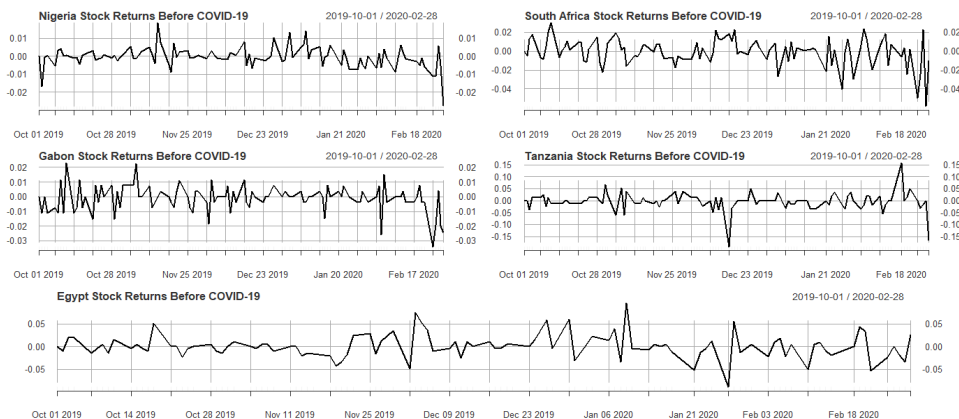
Table 1 indicates the descriptive statistics of stock returns before COVID-19. The result shows that Tanzania has the highest volatility in the daily returns of the series, for the periods of average negative share returns for the countries except Egypt. In addition to this, the daily stock returns of Nigeria, South Africa, Tanzania, and Gabon show traces of series that are negatively skewed while the daily stock returns of Egypt are positively skewed. Furthermore, the kurtosis values for the five countries are greater than three (3), which explains that the data set has abnormal peaked than a normal distribution.

According to Table 2 below, Nigeria has the highest mean stock returns of 0.0024 with Egypt having the highest deviation out of all the other four countries. Looking closely, there exist negative skewness for Nigeria and South Africa, while Tanzania, Gabon and Egypt are positively skewed. The series are equally peaked (leptokurtic) for the countries save for that of Egypt which is platykurtic.

Table 2: Descriptive Statistics of Stock Returns of the Considered Countries During Covid-19

Index	Nigeria	South Africa	Tanzania	Gabon	Egypt
Mean	0.00022	0.00013	0.0015	0.0019	0.0024
Median	0.00051	0.00269	0.0000	0.0000	0.000
Minimum	-0.05063	-0.14823	-0.1525	-0.14407	-0.1965
Maximum	0.04251	0.10090	0.3111	0.13333	0.3235
Standard Dev	0.01013	0.02868	0.0513	0.02527	0.0585
Skewness	-0.859	-1.066	1.35	0.382	1.222
Kurtosis	8.58	9.38	9.86	13.92	10.00

The trend of the stock returns time series before and during the COVID-19 periods is presented in Figures 1 and 2 below. The figures captured the different types of shocks observed by daily stock returns in the two periods.

**Figure 1:** Stock Returns Plot of the Countries Before Covid-19

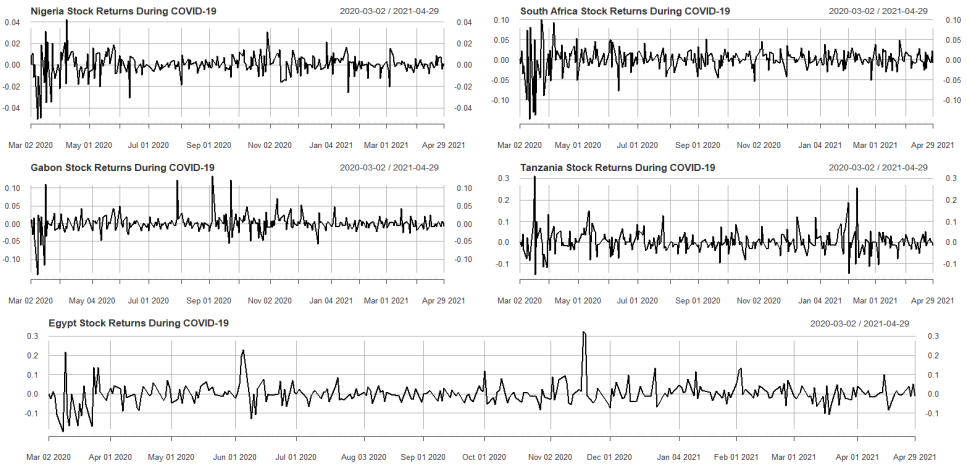


Figure 2: Stock Returns Plot of the Countries During Covid-19

The above plots revealed that each of the stock returns were not stationary at the initial stage and this leads to differencing and each of the data was made stationary

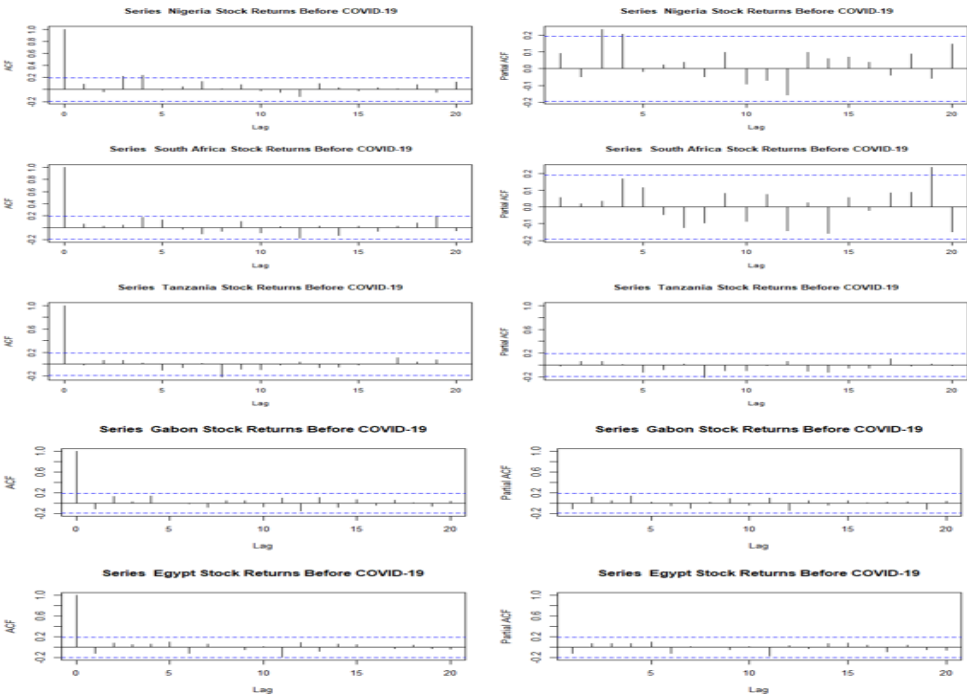


Figure 3: ACF and PACF plots Before COVID-19 for Countries

at the first-order difference, in compliance with the results of the Augmented Dickey Fuller (ADF) test which shows the presence of unit root. The results of ADF are as presented in Table 3 below. The table contains the summary of the results at the level and at the first differencing, which is also a confirmation that the series is stationary at the first order differencing. The small p-values $< \alpha = 0.05$ significance level showed that the series is stationary at the first difference.

Table 3: ADF Results for Unit root Test During COVID-19 for Countries

Country	Test	Lag order	P-value 5%	Test Statistic
Nigeria	@Level	4	0.2	-3
	@1 st Difference	5	0.01	-6
South Africa	@Level	4	0.2	-3
	@1 st Difference	5	0.01	-6
Tanzania	@Level	4	0.2	-4
	@1 st Difference	5	0.01	-5
Gabon	@Level	4	0.2	-3
	@1 st Difference	4	0.04	-4
Egypt	@Level	4	0.06	-3
	@1 st Difference	5	0.01	-5

A precise order of differencing was determined by the plots of ACF and PACF. A cross-examination of the ACF showed the presence of long memory structure, and the two plots provided significant spikes at lag 1 in both cases as shown in Figure 3 and Figure 4 for Nigeria returns. Achieving this stationarity condition is highly essential for modelling the adopted volatility technique.

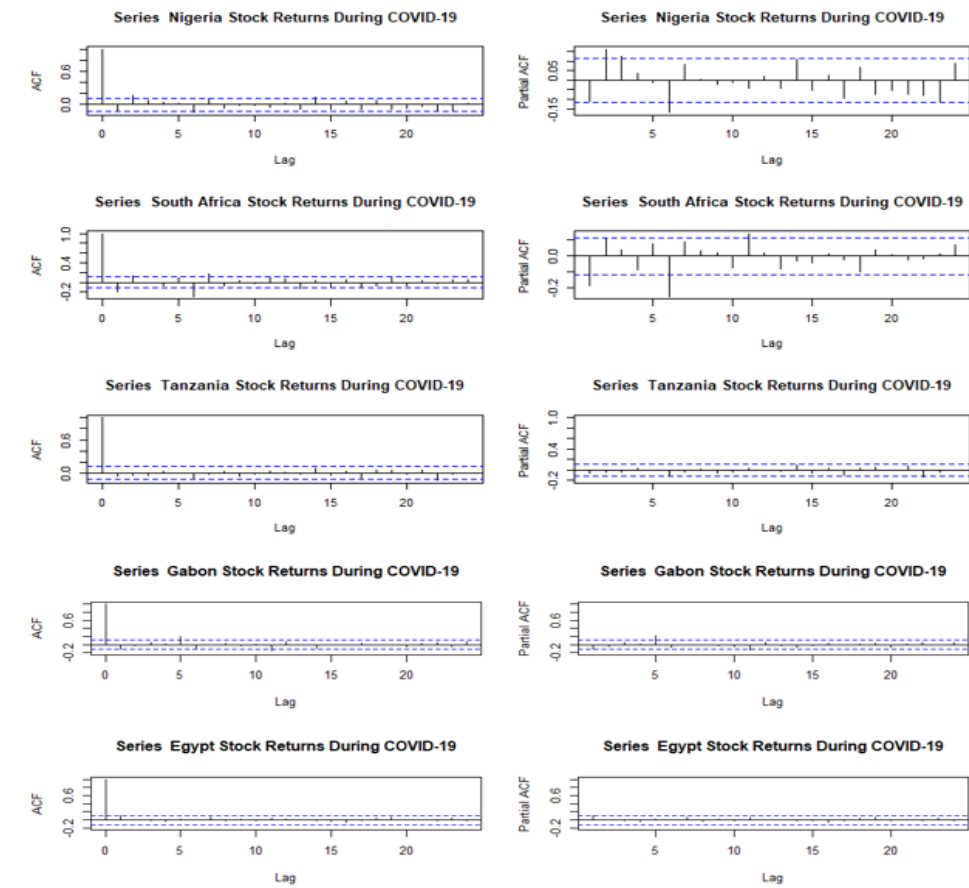


Figure 4: ACF and PACF plots During COVID-19 for Countries

3.2. Summary of Findings: Country by Country Volatility Rates for Pre and Covid-19 Periods

This research adopted a model order of $p = 1$ and $q = 1$ because it has the lowest values of Akaike, Bayes, Shibata and Hanan-Quinn information criteria. More so, it has been established as the best order that fits financial time series excellently (Tsay, 2005).

Table 4: Results of Egarch (1,1) Modelling of Stock Returns Volatility Before COVID-19

Parameters	Nigeria	South Africa	Tanzania	Gabon	Egypt
μ	-0.00605 (0.1198)	-0.00378 (0.1185)	0.00370 (0.4231)	0.00001 (0.9464)	-0.00250 (0.4688)
ω	-5.13609 (0.0011)**	-1.24177 (0.1357)	-2.70244 (0.0007)**	-4.56987 (0.0019)**	-2.21648 (0.0037)**

Table 4: Results of Egarch (1,1) Modelling of Stock Returns Volatility Before COVID-19 (cont.)

Parameters	Nigeria	South Africa	Tanzania	Gabon	Egypt
α	-0.04821 (0.7585)	-0.32487 (0.0720)	0.11536 (0.4359)	0.06709 (0.5249)	-0.11993 (0.4566)
β	0.15536 (0.5428)	0.81935 (0.0000)**	0.47592 (0.0012)**	0.41317 (0.0272)*	0.62058 (0.0000)**
γ	0.96837 (0.0000)**	0.66643 (0.0017)**	1.17323 (0.0000)**	1.03121 (0.0000)**	1.26898 (0.0000)**

Note: P-Values are in parenthesis and *, ** statistically significant at the 5% and 1% significant level.

Table 4 shows the results of EGARCH (1,1) models for the considered countries before the occurrence of Covid-19 in Africa with the following models specified respectively.

$$\log(\sigma_t^2) = -5.136 + 0.155\log(\sigma_{t-1}^2) - 0.048\frac{e_{t-1}}{\sigma_{t-1}} + 0.968\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (14)$$

$$\log(\sigma_t^2) = -1.241 + 0.819\log(\sigma_{t-1}^2) - 0.324\frac{e_{t-1}}{\sigma_{t-1}} + 0.666\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (15)$$

$$\log(\sigma_t^2) = -2.702 + 0.465\log(\sigma_{t-1}^2) - 0.115\frac{e_{t-1}}{\sigma_{t-1}} + 1.173\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (16)$$

$$\log(\sigma_t^2) = -4.569 + 0.413\log(\sigma_{t-1}^2) + 0.067\frac{e_{t-1}}{\sigma_{t-1}} + 1.031\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (17)$$

$$\log(\sigma_t^2) = -2.216 + 0.620\log(\sigma_{t-1}^2) - 0.119\frac{e_{t-1}}{\sigma_{t-1}} + 1.268\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (18)$$

Examining the models (14) – (18), the leverage effect α was negative and at the same time not significant for Nigeria, South Africa, and Egypt's daily stocks. This shows the presence of high risk in stock returns due to the increased leverage induced by negative shocks in the aforementioned countries. On the other hand, there exist no leverage effect on the stock returns of Tanzania and Gabon due to the positive values of their α values. This implies that volatility responds better to favorable news than it does to unsavory news of equal magnitude, as a pointer to the fact that there exists little or no risk in the stock returns of Tanzania and Gabon before the pandemic, especially with the positive average returns of 0.00370 and 0.0001 estimated respectively for the countries compared to that of Nigeria, South Africa and Egypt.

The results for β show that only the coefficient of South Africa is close to one (1) and this implies high persistence of volatility shocks for the country while countries like Nigeria, Tanzania, Gabon, and South Africa have experienced a low persistence of volatility shock.

The γ parameter shows the extent at which the magnitude of the shock to the variance affects the future volatility in the daily returns of each country's stock and also the spillover. The estimated γ results are positive estimates for all the countries, and this

implies that the magnitude of the spillover effect of the volatility is positively related and significant at both 1% and 5% levels. This shows that the changes in the behaviour of the daily stock prices of each country will influence changes in subsequent behaviours of the prices.

Table 5: Results of Egarch (1,1) Modelling of Stock Returns During COVID-19

Parameters	Nigeria	South Africa	Tanzania	Gabon	Egypt
μ	-0.00044 (0.35501)	-0.00154 (0.34770)**	0.0017 (0.54)	-0.0008 (0.4827)	0.0040 (0.0569)
ω	-3.03575 (0.000442)**	-1.40256 (0.02342)*	-2.7806 (0.0000)**	-4.4548 (0.0000)**	-2.0429 (0.0000)**
α	-0.016552 (0.848044)	-0.12288 (0.2391)	0.0430 (0.5940)	-0.0007 (0.9926)	0.0851 (0.2318)
β	0.652116 (0.0000)**	0.79215 (0.0000)**	0.4897 (0.0000)**	0.36050 (0.0002)**	0.60816 (0.0000)**
γ	1.06462 (0.0000)**	0.69089 (0.0093)**	0.91140 (0.0000)**	0.99653 (0.0000)**	0.78343 (0.0000)**

Note: P-Values are in parenthesis and *, ** statistically significant at the 5% and 1% significant level.

Table 5 shows the results of the EGARCH (1,1) models for the considered countries during Covid-19 in Africa with the following models specified in the equations below.

$$\log(\sigma_t^2) = -3.035 + 1.064\log(\sigma_{t-1}^2) - 0.016\frac{e_{t-1}}{\sigma_{t-1}} + 0.652\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (19)$$

$$\log(\sigma_t^2) = -1.402 + 0.690\log(\sigma_{t-1}^2) - 0.122\frac{e_{t-1}}{\sigma_{t-1}} + 0.792\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (20)$$

$$\log(\sigma_t^2) = -2.780 + 0.911\log(\sigma_{t-1}^2) + 0.043\frac{e_{t-1}}{\sigma_{t-1}} + 0.489\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (21)$$

$$\log(\sigma_t^2) = -4.454 + 0.996\log(\sigma_{t-1}^2) - 0.0007\frac{e_{t-1}}{\sigma_{t-1}} + 0.360\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (22)$$

$$\log(\sigma_t^2) = -2.042 + 0.783\log(\sigma_{t-1}^2) + 0.0851\frac{e_{t-1}}{\sigma_{t-1}} + 0.608\left|\frac{e_{t-1}}{\sigma_{t-1}}\right| \quad (23)$$

Equations (19), (20), (21), (22), and (23) are the respective Egarch (1,1) models' specification for daily stock prices in Nigeria, South Africa, Tanzania, Gabon, and Egypt during the COVID19 period.

For the Covid-19 period, the average stock returns for the countries are -0.00044, -0.00154, 0.0017, -0.0008 and 0.0040 respectively for Nigeria, South Africa, Tanzania, Gabon and Egypt. These results clearly shown that only Nigeria, South Africa and Gabon average stock returns have suffered significantly from the pandemic effects, compared to the pre-Covid-19 periods. We noticed that there exists no leverage effect on the stock returns of Nigeria, Tanzania, Gabon and Egypt and this implies that

volatility in these countries responded well to the Pandemic more than they did in the previous era. These countries, however, experienced no leverage effect due to the fact that α is either positive or negative and which at the same time was not significant. The results for β show that the coefficient values for Nigeria, South Africa and Egypt are close to one. These values imply high persistence of volatility shocks for these countries during the Covid-19 pandemic and, on the other hand, we noticed that the β coefficient for Tanzania and Gabon is relatively low. The γ parameter shows the extent at which the magnitude of the shock to the variance affects the future volatility in the daily returns of each country's stock and also the spillover. These estimated γ shows that there is a positive estimate for Nigeria, South Africa, Tanzania, Gabon and Egypt. This means that the magnitude or the spillover of the volatility is positively related and significant at 1% level. This shows that the changes in the behaviour of the daily stock prices of each of the countries will influence changes in subsequent behaviours of the prices.

The graphs in Figure 5 below presented the impact of Covid-19 on the volatility of stock returns for the five countries. Based on the graphs, the stock volatility of all the countries responded sharply to the outbreak of COVID-19. With the exception of Gabon, the returns for the countries nosedived and remained constant for most of the periods ε_{t-1} .

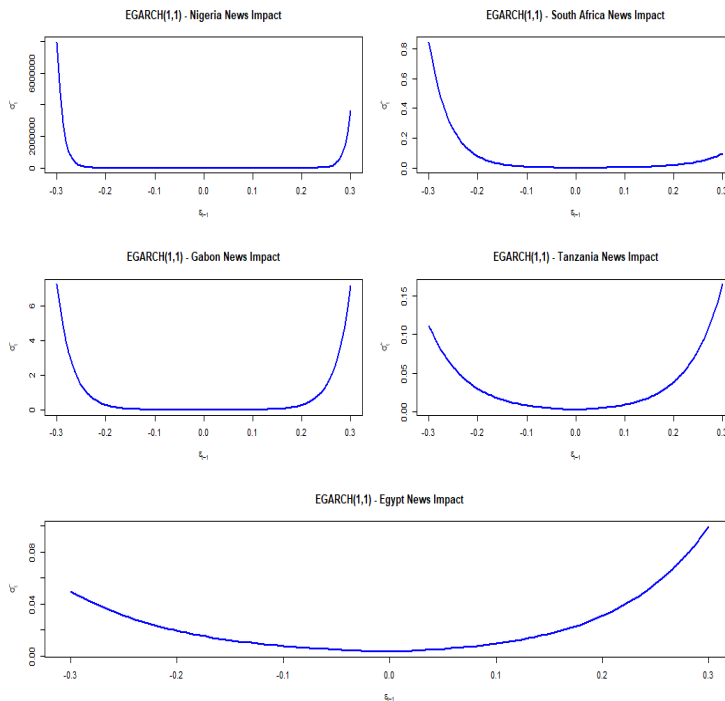


Figure 5: COVID-19 Impact on the Stock Returns of Nigeria, South Africa, Tanzania, Gabon and Egypt

3.3. Consolidated Effects of Covid-19 Trajectory on Stock Volatility

The results of 2SLS estimated to model the effect of Covid-19 on the countries' stock returns are presented in Table 6 while its validity statistic are given in Table 7 below.

Table 6: Results of Instrumental Variable (2sls)

Specification	coefficient	std. error	t-ratio	p-value
Const	-47.01110362	16.01185123	-2.936019262	0.0033**
Confirmed	0.006351376	0.001401122	4.533065161	6.14E-06**
Deaths	0.036818759	0.008865751	4.152920586	3.41E-05**
Recovered	-0.008172232	0.001820635	-4.488672282	7.55E-06**

Note: P-Values are in *, ** are statistically significant at the 5% and 1% significant level.

Table 7: Model Diagnostic of IV (2sls)

Mean dependent var	22.502	S.D. dependent var	38.0
Sum squared residual	86709876.35	S.E. of regression	202.6
R-squared	0.019968308	Adjusted R-squared	0.018
F(3, 2111)	7.157	P-value(F)	8.82E-05
Log-likelihood	-46703.4761	Akaike criterion	93414.9
Hausman test	911.00	p-value	3.97E-200

The Instrumental variable model which was specified from Table 6 above is shown below:

$$Stock_{it} = -47.011 + 0.0063(Confirmed)_{1,it} - 0.008(Recovered)_{2,it} + 0.036(Death)_{3,it} \quad (27)$$

Equation (27) specified the direct impact of Covid-19 cases (i.e. confirmed, deaths and recoveries) on the stock returns of all the countries pooled together and the individual parameters are significant based on their p-values provided in Table 7. The model provided overall goodness of fit as reflected in its F-value of 7.157 with P-value of 0.0000882. More so, the results of the estimated coefficients aligned with the a priori. That is, confirmed and death cases increase price volatility while recovered cases will bring about reduction in the stock prices. Based on the results shown in Table 7, this instrument can be considered as exogenous given that the null hypothesis is not rejected at both 1% and 5% levels as measured by the Hausman test statistic.

3.4. Forecast

Predicted values of the fitted EGARCH (1,1) were studied using the test data after adequacy check of the models was done. An unconditional sigma forecast was made for the days in the month of October, November and December 2021. The forecast values exhibited a constant trend of volatility shocks for all the countries in the continuous presence of the Covid-19 pandemic. However, Nigeria volatility experienced a significant spike during the few days of October before maintaining a constant trend. The extreme coloured ends of the graphs presented in Figure 6 (a - e) depicted the forecasts.

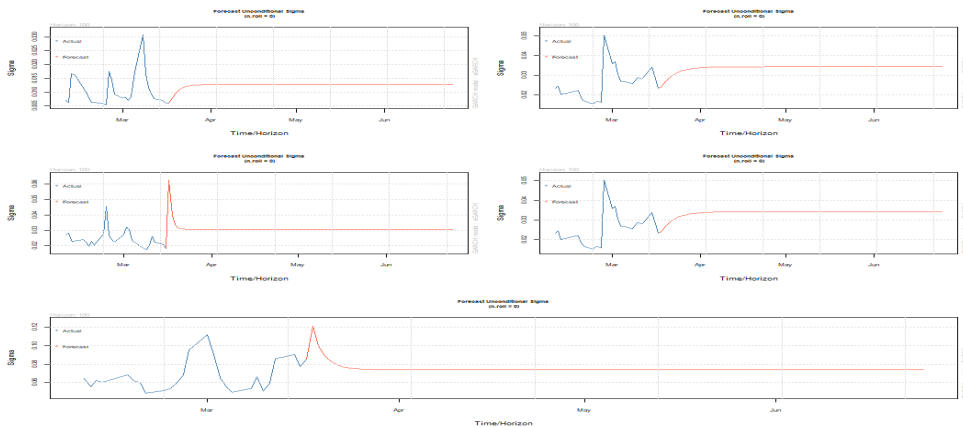


Figure 6: Volatility Forecasts in the Continuous Presence of COVID-19

3. Conclusions

This article has exemplified and emphasized through empirical analysis, the impact of Covid-19 on the volatility of stock markets within the African continent. Stock volatility during COVID-19 compared well with that of the pre-COVID-19 period and it has been well established that the stock volatility of all the countries responded sharply to the outbreak of COVID-19 with the average stock returns of Nigeria and Gabon suffering the most shocks from the pandemic effects. The stock returns of the five countries equally exhibited a long memory as the autocorrelation function of the series showed persistence characteristic with exponential decay towards zero, which is one of the features of a long memory process. The results thus implied high persistence of volatility shocks for all the countries during the Covid-19 pandemic and that the magnitude or the spillover effect of the volatility is positively related and highly significant. The positive feat achieved in terms of average returns by South Africa, Tanzania and Egypt during the pandemic may be attributed to smart investors' bargain,

their bullish attitudes gingered by the release of positive year-end financial results of several quoted companies, coupled with improved dividend declaration. The study has also established that confirmed and death cases increase stock price volatility while recovered cases will bring about reduction in stock prices for all the countries considered in this research. Also, the forecast values exhibited a constant trend of volatility shocks for all the countries in the continuous presence of the Covid-19 pandemic. The implication of this trend is such that many investors will not be willing to stake their funds in the capital markets as long as the Covid-19 pandemic persists. Thus, stock prices might remain unchanged for a long period due to the inactive capital market.

The above deduction is in line with the submission of Jeremy Schneider, a personal financial expert at Personal Finance Club, on the effect of the ongoing war on the stock market of Ukraine. He posited that the war has introduced new uncertainty to a stock market that has already had a shaky start to the year, and that the S&P 500 saw its most dramatic one-day drop since May 2020, amid a war with no end in sight. With hundreds of civilians dead, including children, and more than half a million refugees having fled Ukraine, the most important consequence is clearly the human cost, rather than anything having to do with people's investments. As the war continues, so does the unpredictability of the consequences beyond the borders of Ukraine, according to Vaughn (2022). This conclusion is also in tune with the earlier work of Scott *et al.* (2016), where the authors were of the opinion that uncertainty caused by irregular variation such as the ongoing war between Russia and Ukraine is associated with greater stock price volatility and reduced investment.

Acknowledgement

The authors wish to acknowledge the online data producers through which the data for this research were sourced.

References

- Adebayo A. O., Esther O. O., Kafilat A. B., Mayowa J. F., (2020). Preliminary Evaluation of COVID-19 Disease Outcomes, Test Capacities and Management Approaches among African Countries. *Cold Spring Harbor Laboratory*.
- Adeboye, N. O., Fagoyinbo, I. S., (2017). Fitting of Seasonal Autoregressive Integrated Moving Average to the Nigerian Stock Exchange Trading Activities. *Edited Proceedings of 1st International Conference, Professional Statistical Society of Nigeria*. Vol. 1, pp. 12–16.

- Adegboye, O., Adekunle, A., Pak, A., Gayawan, E., Leung, D., Rojas, D., Elfaki, F., McBryd, E., Eisen, D., (2020). Change in outbreak epicenter and its impact on the importation risks of COVID-19 progression: A modelling study. *medRxiv* 2020.
- Adhikari S. P., Shameng S, Wu Y., Mao Y., Ye R., Wang Q., Sun C., Sylvia S., Rozelle S., Raat H, Zhou H., (2020). Epidemiology, causes, clinical manifestation and diagnosis, prevention and control of coronavirus disease (COVID-19) during the early outbreak period: a scoping review. *Infectious Diseases of Poverty* 2020, Vol. 9, issue 1, pp. 29. doi:10.1186/s40249-020-00646-x.
- Arellano, M., Bond. S., (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, Vol.58, issue 2, pp. 277–297.
- Ayinde, K., Adewale, F. L., Rauf, I. R., Alabi, O. O., Okon, C. E., Ayinde, E. O., (2020). Modeling Nigerian Covid-19 cases: A comparative analysis of models and estimators. *Chaos, Solitons and Fractals*, Vol. 138 (2020) 109911. <https://doi.org/10.1016/j.chaos.2020.109911>, 0960-0779/© Elsevier.
- Baillie, R., (1996). Long Memory Processes and Fractional Integration in Econometrics. Vol. 73, issue 1, pp. 5–59.
- Baker, S. R., Bloom, N., Davis, S. J., (2016). Measuring economic policy uncertainty. *The quarterly journal of economics*, 131(4), pp. 1593–1636.
- Barron's (2020). The Dow Is Down 700 Points as the Coronavirus Strikes in Europe. <https://www.barrons.com/articles>. Accessed July 28.
- Blinder, J. J., Merges, M. J., (2001). Stock Market Volatility and Economic Factors. *Review of Quantitative Finance and Accounting*, Vol. 17, pp. 5–26.
- CDC, (2020). Centre for Disease Control. Available at <https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html>. Accessed July 20.
- Chiang, T. C., Doong, S., (2001). Empirical Analysis of Stock Returns and Volatility: Evidence from Seven Asian Stock Markets Based on TAR-GARCH Model. *Review of Quantitative Finance and Accounting*, Vol. 17, pp. 301–318 <https://doi.org/10.1023/A:1012296727217>.
- ECDC, (2020) European Centre for Disease Prevention and Control. Available at <https://www.ecdc.europa.eu/en/covid-19-pandemic>. Accessed July 20.
- Giordano G, Blanchini F, Bruno R, Colaneri P, Filippo A. D., Matteo Ad, Colaneri M., (2020). *Nat Med* 2020: 1–32. doi:10.1038/s41591-020-0883-7.

- Gilbert, M., Pullano, G., Pinotti, F., Valdano, E., Poletto, C. Boëlle, P. Y., D'Ortenzio, E., Yazdanpanah, Y., Eholie, S. P., Altmann, M., (2020). Preparedness and vulnerability of African countries against importations of COVID-19: a modelling study, *Lancet*, Vol. 395, pp. 871–877.
- Grundy, B. D., Kim, Y., (2002). Stock Market Volatility in a Heterogeneous Information Economy. *Journal of Financial and Quantitative Analysis*, Vol. 37, issue 1, pp.1–27. <https://doi.org/10.2307/3594993>.
- Guo Y, Cao Q, Hong Zhong-Si, Yan Y., (2019). The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Military Medical Research*, Vol. 7, issue 1, doi:10.1186/s40779-020-00240-0.
- GitHub, (2020). *olawale0254/Dataset-of-daily-stock-prices-and-Covid-19-Reported-Cases-in-Selected-African-Countries*. A Data Repository for an Econometric Modelling for Stock and Covid-19, <https://bit.ly/37LqPkw>.
- Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y., (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet 2020*, Vol. 395, pp. 497–506, doi:10.1016/S0140-6736(20)30183-5.
- JHU CSSE, (2020). An Interactive Web-based dashboard to track COVID-19 in real time, <https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6>.
- Maclean, O.A., Orton, R., Singer, J. B., Robertson, D. L., (2020). Response to *On the origin and continuing evolution of SARS-CoV-2*, <http://virological.org/t/response-to-on-the-origin-and-continuing-evolution-of-sars-cov-2/418>.
- Olowe, R. A., (2009). Modelling Naira/Dollar Exchange Rate Volatility: Application of Garch and Asymmetric Models. *International Review of Business Research Papers*, Vol. 5, issue 3, pp. 377–398.
- NCDC, (2020). <http://covid19.ncdc.gov.ng>. Accessed July 16.
- Nigerian Stock Exchange Weekly Market Report, (2020). <https://www.nse.com.ng>. Accessed July 18.
- Martinez-Alvarez, M., Jarde, A.; Usuf, E., Brotherton, H., Bittaye, M., Samateh, A. L., Antonio, M., Vives-Tomas, J., D'alessandro, U., Roca, A., (2020). COVID-19 Pandemic in west Africa. *Lancet Glob. Health*.
- NBS, (2020). Covid-19 Impact Monitoring, Round 2, Nigeria COVID-19, National Longitudinal Phone survey with support from the World Bank, <https://www.nigerianstat.gov.ng>. Accessed July 17.

- Nelson, D. B., (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach, *Econometrica*, 59(2), pp. 347–370, doi: 10.2307/2938260, <https://www.jstor.org/stable/2938260>.
- Nadeem S., (2020). Coronavirus COVID-19: Available Free Literature Provided by Various Companies. Journals and Organizations around the World. *J Ong Chem Res* 2020, Vol. 5, issue 1. pp. 7–13. Document ID: 2020JOCR37, doi: 10.5281/zenodo.3722904.
- OECD, (2020). Global financial markets policy responses to COVID-19. <http://www.oecd.org/coronavirus/policy-responses/global>. Accessed July 28.
- Olalekan J. Akintande., Olusanya E. Olubusoye., Adeola F. Adenikinju., Busayo T. Olanrewaju, (2020). Modeling the determinants of renewable energy consumption: Evidence from the five most populous nations in Africa. *Energy*, 206(2020) 117992, <https://doi.org/10.1016/j.energy.2020.117992>, 0360-5442/© 2020 Elsevier.
- Oyelola A. A., Adeshina I. A., Ezra G., (2020). Early Transmission Dynamics of Novel Coronavirus (COVID-19) in Nigeria. *International Journal of Environmental Research and Public Health*, Vol. 17, pp. 3054, doi: 10.3390/ijerph17093054.
- Pramod, K. N., Puja, P., (2015). Stock Market Volatility and Equity Trading Volume: Empirical Examination from Brazil, Russia, India and China (BRIC). *Global Business Review*, 16(5), <https://doi.org/10.1177/0972150915601235>.
- Ser-Huang, P., Taylor, S. J., (1992). Stock returns and volatility: An empirical study of the UK stock market. *Journal of Banking and Finance*, 16(1), pp. 37–59, [https://doi.org/10.1016/0378-4266\(92\)90077-D](https://doi.org/10.1016/0378-4266(92)90077-D)Get rights and content.
- Scott R. B., Nicholas B., Steven J. D., (2016). Measuring Economic Policy Uncertainty. *The Quarterly Journal of Economics*, 131(4), pp. 1593–1636, <https://doi.org/10.1093/qje/qjw024>.
- Tsay, R. S., (2005). Analysis of Financial Time Series – 2nd Ed. Wiley-Interscience.
- Theodore, F. D., Lewis, M., (1992). Stock market volatility and the information content of stock index options. *Journal of Econometrics*, Vol. 52, issue 1–2, pp. 267–287.
- UNICEF, (2020). UNICEF Gabon Covid-19 Report, <https://www.africanews.com/2020/06/28/coronavirus>, Accessed July 28, 2020.
- U.S. Embassy in Egypt, (2020). Covid-19 Information. <https://eg.usembassy.gov/u-s-citizen-services/covid-19-informatio>, Accessed July 28.

- Vladimir R. C., Vasily V. L., (2020). Ribonucleocapsid Assembly/packaging Signals in the Genomics of the Coronaviruses SARS-CoV and SARS-CoV-2; Detection, Comparison and implications for Therapeutic Targeting. *Journal of Biomolecular Structure and Dynamics*.
- Vaughn H., (2020). War in Ukraine Means Fresh Volatility for the Stock Market. That Shouldn't Change Your Investing Strategy, <https://time.com/nextadvisor/investing/stock-market-rattled-by-russia-ukraine/> Accessed May 10, 2022.
- Weltman, J., (2020). ECR risk experts contemplate another financial crisis. *Euromoney*, <https://www.euromoney.com/article>. Accessed July 28.
- WHO, (2020). Africa report COVID-19 case, <https://www.afro.who.int/health-topics/coronavirus-covid-19>. Accessed July 28.
- WHO, (2020). Global Covid-19 updates, <https://covid19.who.int>. Accessed July 28.
- Yahoo Finance, (2020). Stock Market Live Quotes, Business & Finance News. <https://finance.yahoo.com/>.
- Yi Hu, Dongmei G., Ying D., and Shouyang W., (2014). *Estimation of Nonlinear Dynamic Panel Data Models with Individual Effects*. Mathematical Problems in Engineering, Article ID 672610, <http://dx.doi.org/10.1155/2014/672610>.

Bayesian estimation of fertility rates under imperfect age reporting

Vivek Verma¹, Dilip C. Nath², S. N. Dwivedi³

ABSTRACT

This article outlines the application of the Bayesian method of parameter estimation to situations where the probability of age misreporting is high, leading to transfers of an individual from one age group to another. An essential requirement for Bayesian estimation is prior distribution, derived for both perfect and imperfect age reporting. As an alternative to the Bayesian methodology, a classical estimator based on the maximum likelihood principle has also been discussed. Here, the age misreporting probability matrix has been constructed using a performance indicator, which incorporates the relative performance of estimators based on age when reported correctly instead of misreporting. The initial guess of performance indicators can either be empirically or theoretically derived. The method has been illustrated by using data on Empowered Action Group (EAG) states of India from National Family Health Survey-3 (2005–2006) to estimate the total marital fertility rates. The present study reveals through both a simulation and real-life set-up that the Bayesian estimation method has been more promising and reliable in estimating fertility rates, even in situations where age misreporting is higher than in case of classical maximum likelihood estimates.

Key words: Fisher information, square error loss function, age-specific marital fertility rate, Bayes estimator, maximum likelihood principle.

1. Introduction

The purpose of any demographic or health sample survey is to provide information on the demographic parameters of the concerned population. In demographic studies, the age of an individual plays an important role, and misreporting leads to transfers of an individual from one age to another. Misreporting causes subjective biases due to random and systematic errors in data that influence the estimate of the population parameters. Earlier studies by Hussey and Elo (1997), Narasimhan et al. (1997) and Denic et al. (2004), Yi (2008), and Neal et al. (2012) show that age misreporting is still highly prevalent in many countries including India. As a result of misreporting, various measures and vital indicators that are age-dependent get influenced (Coale and Li (1991), Szoitzysek et al. (2017)). To overcome this problem many alternative methods have been discussed by Bhat (1990), Dechter and Preston (1991), Bhat (1995), and Nwogu and Okoro (2017), which are based on the

¹Corresponding Author. Department of Statistics, Assam University, Silchar, Assam, India. E-mail: viv_verma456@yahoo.com. ORCID: <https://orcid.org/0000-0003-2537-4431>.

²School of Applied and Pure Sciences, Royal Global University, Guwahati, Assam, India. E-mail: dilipc.nath@gmail.com.

³All India Institute of Medical Sciences, New Delhi, India. E-mail: dwivedi7@gmail.com. ORCID: <https://orcid.org/0000-0003-4262-6143>.

requirement and availability of the other related information to detect and measure these errors.

The total marital fertility rate (TMFR) is considered as one of the important measures of the overall summary of marital fertility. The measure of TMFR is basically a linear function of the number of live births to the women in each group. Therefore, the distribution for the total marital fertility rate of a population is difficult to get in an explicit form. Hence, the total fertility rate is estimated by using this linear function. The procedure for estimation and prediction of TMFR using various alternative methods was already explored in studies by Garenne et al. (2001), Yadava and Kumar (2002), Martin et al. (2011), and Pathak and Verma (2013).

Under the assumption that TMFR is an unknown but fixed quantity and there is no age misreporting, many studies have been derived and investigated this. But, in practice, TMFR is a random quantity, and can quantify the randomness specifying suitable prior distribution for it. As such, the Bayesian approach could be successfully applied for making statistical inference on TMFR.

Fertility is regarded as one of the essential demographic measures and is influenced by age misreporting. Imperfect or wrong age reporting has been remained a methodological problem (Murray et al., 2018; Singh et al., 2020; Schoumaker, 2020), and for the sake of analysis, sophisticated methodological techniques are needed to address this situation during estimation. For situations like the estimation of age-specific mortality (Bhatta and Nandram, 2013), projecting populations (Daponte et al., 1995), school completion (Barakat et al., 2021), where age-misreported, the Bayesian methodology has been found very effective to estimate the population characteristics.

Under the assumption that age was correctly reported in recent years, various Bayesian methodology-based estimates of fertility rates have been also introduced by Oh (2018), Liu and Raftery (2018), Borges (2019), and Schmertmann and Hauer (2019). But the problem of age misreporting remains unexplored. The Bayesian inference on TMFR, based on the linear function of birth in married women in each age group, has not been considered much in the literature. The study is different from the existing one as it considered limited assumptions on the structure of data and choice of prior distribution in terms of hyper-parameters values. The present study attempts to progress in the same direction of utilizing the Bayesian paradigm to estimate TMFR considering that the age has been misreported. The present study aimed to derive a prior TMFR using the same linear function following Fishers' information. As an alternative to the Bayesian methodology, a classical estimator using the maximum likelihood principle has also been discussed. The performance of the derived posterior distributions is also generalized and investigated for both perfect and imperfect age-reporting situations. Here, we hypothesized that the Bayesian estimation method might provide a more promising and reliable estimation of fertility rates, even in cases where age misreporting is higher than classical maximum likelihood estimates.

This article is organized in the following way. Section 2 provides classical and Bayesian estimates of TMFR, based on the maximum likelihood principle and the Bayesian method, respectively, under perfect age reporting. In Section 3, the procedure is generalized for imperfect estimates of TMFR when age is misreported. Section 4 illustrates the performance of the derived prior and its associated posterior distribution through numerical simulation.

Section 5 illustrates the proposed estimate through real-life data of women belonging to the childbearing age-group, *i.e.* 15-44 years, from third rounds of the National Family Health Survey (NFHS-3) of 2005-2006 in India. Section 6 provides the results and discussion. Lastly, Section 6 gives a summary and conclusion.

2. TMFR Estimation under Perfect Age Reporting

Let us consider a population of married women who are in the childbearing age group (*i.e.*, 15-44 years) at a particular period. Let X_{ai} denote a binary form of the event of ever occurrence of birth to the i^{th} women during the study period within the a^{th} childbearing age-interval, where $i = 1, 2, \dots, n_a$ and $a = 1, 2, \dots, c$. Here, c denotes the number of non-overlapping age-groups and n_a is the number of women in the a^{th} age group. The cases of twin births in a particular interval are not considered a serious issue in reality as these events are rare and found to be one out of 240 births in the database. The probability mass function (p.m.f.) of age-specific birth occurrence to a woman is given by

$$f(x_{ai}|p_a) = p_a^{x_{ai}}(1 - p_a)^{1-x_{ai}}, \quad x_{ai} = 0, 1, \quad 0 < p_a < 1, \quad (2.1)$$

where p_a denotes the probability that a child was born to a married woman belonging to a^{th} childbearing age-group, referred to as the age-specific married fertility rate (ASMFR) of mothers belonging to a^{th} age-group, for all $a = 1, 2, \dots, c$. For any age-group, say a , let $Y_a \left[= \sum_{i=1}^{n_a} X_{ai} \right]$ denote the total number of children born to n_a women belonging to that age-group, then Y_a is assumed to follow the Binomial (n_a, p_a) distribution. The estimate of probability that a child was born to a married woman in a^{th} age-group, p_a , is obtained using the observed sample, say Y_a .

2.1. Estimator of TMFR based on Maximum Likelihood Principle

Let $f(y_a|p_a)$ denote the p.m.f. of Y_a and by applying the standard maximum likelihood (ML) principle, the ML estimate of p_a , for all $a = 1, 2, \dots, c$, is obtained as

$$\hat{p}_a = \arg \max_{p_a} f(y_a|p_a) = \frac{y_a}{n_a} \quad (2.2)$$

and if the condition

$$\sum_{y_a=0}^{n_a} \frac{\delta}{\delta p_a} f(y_a|p_a) = 0 \quad (2.3)$$

is satisfied, then the variability explained by the estimator of p_a is given by

$$V(\hat{p}_a) \geq \{\mathcal{J}(\hat{p}_a)\}^{-1} = \frac{\hat{p}_a(1 - \hat{p}_a)}{n_a}. \quad (2.4)$$

Classically, the estimate of TMFR has been obtained using the estimates of the probabilities, p_1, p_2, \dots, p_c , using the linear function:

$$\psi(p) = \sum_{a=1}^c \alpha_a p_a, \quad \alpha_a \geq 0. \quad (2.5)$$

Fisher's information of the probability that a child born to a married woman belonging to a^{th} childbearing age-group, p_a , using standard notation, has been obtained as

$$\mathcal{J}(p_a) = \sum_{y=0}^{n_a} \left(\frac{\delta}{\delta p_a} \log f(y_a | p_a) \right)^2 f(y_a | p_a) \quad (2.6)$$

and the inverse of Fisher's information matrix of age classified probabilities vector, say $p = (p_1, \dots, p_c)$, has been given by

$$\mathcal{J}^{-1}(p) = \mathcal{J}^{-1}(p_1, p_2, \dots, p_c) = \begin{bmatrix} \frac{p_1(1-p_1)}{n_1} & 0 & \dots & 0 \\ 0 & \frac{p_2(1-p_2)}{n_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{p_c(1-p_c)}{n_c} \end{bmatrix}$$

For the given linear function, $\psi(p)$, of TMFR in equation (2.5), the gradient of p has been obtained as

$$\begin{aligned} D_{\psi}^T(p) &= \left[\frac{\partial \psi(p)}{\partial p_1} \quad \frac{\partial \psi(p)}{\partial p_2} \quad \dots \quad \frac{\partial \psi(p)}{\partial p_a} \quad \dots \quad \frac{\partial \psi(p)}{\partial p_c} \right] \\ &= [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_a \quad \dots \quad \alpha_c]. \end{aligned}$$

Let $v_a = (\mathcal{J}^{-1}(p))_{aa} = (a^{th} \text{ diagonal element of } \mathcal{J}^{-1}(p)) = \frac{p_a(1-p_a)}{n_a}$, then

$$\begin{aligned} D_{\psi}^T(p) \mathcal{J}^{-1}(p) &= [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_c] \begin{bmatrix} v_1 & 0 & \dots & 0 \\ 0 & v_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & v_c \end{bmatrix} \\ &= [\alpha_1 v_1 \quad \alpha_2 v_2 \quad \dots \quad \alpha_c v_c] \end{aligned} \quad (2.7)$$

$$D_{\psi}^T(p) \mathcal{J}^{-1}(p) D_{\psi}(p) = \sum_{a=1}^c \alpha_c^2 v_a. \quad (2.8)$$

The mean and variance of $\psi(p)$ based on the ML estimates are of the form

$$\hat{\psi}(p)^M = \sum_{a=1}^c \alpha_a \hat{p}_a \quad (2.9)$$

$$\begin{aligned} V(\hat{\psi}(p)^M) &= D_{\psi}^T(p) \mathcal{J}^{-1}(p) D_{\psi}(p) \\ &= \sum_{a=1}^c \alpha_a^2 \frac{\hat{p}_a(1-\hat{p}_a)}{n_a}. \end{aligned} \quad (2.10)$$

As $Y_a \sim \text{Binomial}(n_a, p_a)$, for all $a = 1(1)c$, and $\psi(p)$ is estimated as $\hat{\psi}(p) = \sum_{a=1}^c \alpha_a \hat{p}_a$, from the central limit theorem, we have

$$\frac{\psi(p) - \hat{\psi}(p)}{\sqrt{D_{\psi}^T(p) \mathcal{J}^{-1}(p) D_{\psi}(p)}} \sim N(0, 1). \quad (2.11)$$

The confidence interval for TMFR, $\psi(p)$, has been obtained using the above equation as

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\psi(p) - \hat{\psi}(p)}{\sqrt{D_{\psi}^T(p) \mathcal{J}^{-1}(p) D_{\psi}(p)}} \leq z_{\frac{\alpha}{2}}\right) \approx 1 - \alpha, \quad (2.12)$$

where $z_{\frac{\alpha}{2}}$ is the $(\frac{\alpha}{2})^{th}$ quantile from the top of the standard normal distribution.

2.2. Bayes Estimators of TMFR

In the previous sub-section, it has been assumed that the numbers of live births that occurred to women belong to a^{th} age-interval, say Y_a , follow the distribution denoted as $f(y_a|p_a)$. TMFR, $\psi(p)$, is defined as a linear function of unknown but fixed probabilities of having a live birth to a married woman in a^{th} age-group. But in practical situations, TMFR might be a random quantity and can model that randomness through the Bayesian approach by specifying suitable prior distribution for $\psi(p)$. To suggest a prior distribution for $\psi(p)$, a linear function of p_a 's is difficult to be obtained directly. Here, an attempt has been made to derive a prior distribution for $\psi(p)$, based on the linear functions of p_a 's as

Theorem-1: Suppose $\tau(\cdot)$ defines the prior distribution for $\psi(p) = \sum_{a=1}^c \alpha_a p_a$, a linear function of probabilities that a child born to a married woman in a^{th} age-group, say $p_1, \dots, p_a, \dots, p_c$, is given by

$$\tau(p) = \tau(p_1, \dots, p_c) \propto \left\{ \sum_{a=1}^c \alpha_a^2 p_a(1-p_a) \right\}^{1/2} \prod_{a=1}^c p_a^{-1}(1-p_a)^{-1} \quad (2.13)$$

Proof: The proof of Theorem 1 is given in the Appendix.

The posterior distribution based on the matching prior of TMFR, $\psi(p)$, for the given sample has been obtained as

$$q(p|data) \propto L(p|data) \tau(p)$$

$$q(p|y) \propto \left\{ \sum_{a=1}^c \alpha_a^2 p_a (1-p_a) \right\}^{1/2} \prod_{a=1}^c p_a^{y_a-1} (1-p_a)^{n_a-y_a-1}, \quad (2.14)$$

where $y_a = \sum_{i=1}^{n_a} x_{ai}$. Here, the posterior distribution, $q(p|y)$, does not have any explicit form. For this reason one has to get samples from $q(p|x)$ to get the posterior distribution of $\psi(p)$. This is done by simulating $N (= 100000)$ (with burning period 10000) values from the posterior distribution as $\{p_1^{(l)}, p_2^{(l)}, \dots, p_c^{(l)}; l = 1, 2, \dots, N\}$ for fixed values of α_a 's, then these samples have been used for the computation of TMFR as $\psi^{(1)}(p), \psi^{(2)}(p), \dots, \psi^{(N)}(p)$, where $\psi^{(l)}(p) = \sum_{a=1}^c \alpha_a p_a^{(l)}$.

The procedure of the Monte Carlo simulation technique was adopted to estimate an empirical HPD interval of $\psi(p|x)$ using the posterior samples by the following procedure:

1. $\psi^{(1)}(p), \psi^{(2)}(p), \dots, \psi^{(N)}(p)$ are sorted $\psi_{(1)}(p) \leq \psi_{(2)}(p) \leq \dots \leq \psi_{(N)}(p)$
2. Computation of the credibility interval of $100(1 - \alpha)\%$ is done as

$$\Delta_l = (\psi_{(l)}(p), \psi_{(l+[1-\alpha]N)}(p)); \forall l = 1, 2, \dots, N - [1 - \alpha]N$$

3. The $100(1 - \alpha)\%$ credible interval is denoted as Δ_l^* , and is the one which has the smallest interval width among all credible intervals.

Note : The posterior mean and variance of $\psi(p)$ can be approximated as

$$\hat{\psi}(p)^B = E(\psi(p)|x) \simeq \frac{1}{N} \sum_{l=1}^N \psi^{(l)}(p) \quad (2.15)$$

and

$$V(\hat{\psi}(p)^B) = V(\psi(p)|x) \simeq \frac{1}{N} \sum_{l=1}^N [\psi^{(l)}(p)]^2 - \left[\frac{1}{N} \sum_{l=1}^N \psi^{(l)}(p) \right]^2. \quad (2.16)$$

3. Effect of Age Misreporting

The obtained estimates and their related discussions are enough to infer TMFR if each woman correctly reported ages. But, the works of Narasimhan et al. (1997) and Denic et al. (2004), Yi (2008), and Neal et al. (2012) have suggested that age misreporting is still highly prevalent in many countries, including India and hence the error in age reporting is inevitable. As a result, the fertility measures, including TMFR, might get highly underestimated or overestimated, which is likely to inappropriately influence related policy planning

leading to poor health care and/or undue economic burden. So, it is necessary to study how robust the proposed estimates are when the prevalence of misreporting of age is high.

Let 'a' symbolize the age-interval reported by a woman and a^* denote true age-interval, where $a, a^* = 1, 2, \dots, c$. The probability that a child born to a married woman in a^{th} reported age-group may be formulated as

$$p_{a^*}^* = \sum_a \pi_{a,a^*} p_a, \quad a, a^* = 1, 2, \dots, c \quad (3.1)$$

where $p_{a^*}^*$ denotes the probability that a child born to a married woman in her true age-interval a^* , and π_{a,a^*} is the probability of shifting from the true age class a^* to a due to misreporting. Equation (3.1) can be represented in a matrix form as follows:

$$p^* = \pi p, \quad (3.2)$$

where $p = (p_1, p_2, \dots, p_c)'$ is a column vector representing probabilities of birth to a woman based on their reported ages, $p^* = (p_1^*, p_2^*, \dots, p_c^*)'$ is a column vector of probabilities of birth to a women as per their true ages, and π is assumed to be a stochastic transition probabilities matrix was (π_{a,a^*}) of order $c \times c$. The π_{a,a^*} 's are such that

$$0 \leq \pi_{a,a^*} \leq 1, \sum_{a^*} \pi_{a,a^*} = \sum_a \pi_{a,a^*} = 1, \quad \forall a, a^*$$

Based on the above probabilistic model for misreporting of age, we have the following observations:

Theorem 2: If $\alpha_a = \alpha_a^*$, then the estimate of TMFR based on the classical procedure of estimation does not take into account the age misreporting mechanism, i.e.i.e.,

$$\sum_{a^*=1}^c p_{a^*}^* = \sum_{a=1}^c p_a \Rightarrow \psi(p^*) = \psi(p) \quad (3.3)$$

Proof: Let the coefficients of the linear function of TMFR for both prefect age reporting and misreporting are the same i.e. $\alpha_a^* = \alpha_a$, for all $a, a^* = 1, 2, \dots, c$, then

$$\begin{aligned} \sum_{a^*} p_{a^*}^* &= \sum_{a^*} \sum_a \pi_{a,a^*} p_a = \sum_a \left(\sum_{a^*} \pi_{a,a^*} \right) p_a = \sum_{a=1}^c p_a \\ \Rightarrow \psi(p^*) &= \sum_{a^*} \alpha_{a^*}^* p_{a^*}^* = \sum_a \alpha_a p_a = \psi(p). \end{aligned} \quad (3.4)$$

Under imperfect age reporting scenario, the variance of maximum likelihood estimate has been obtained after replacing p_a by p_a^* and $n_a = n_{a^*}$ as

$$V(\hat{\psi}^*(p^{*M})) = \sum_{a^*=1}^c \alpha_{a^*}^2 \frac{\hat{p}_{a^*}^* (1 - \hat{p}_{a^*}^*)}{n_{a^*}}. \quad (3.5)$$

and the posterior distribution of equation (2.14) will be of the form

$$h(p^*|x) \propto \left\{ \sum_{a=1}^c \alpha_a^2 p_a^* (1-p_a^*) \right\}^{1/2} \prod_{a=1}^c (p_a^*)^{y_a-1} (1-p_a^*)^{n_a-y_a-1}. \quad (3.6)$$

In the context of the Bayesian framework, the distribution of p_i^* , for each i , is a mixture of the distributions of c independent variables p_a , $a = 1, 2, \dots, c$ with mixing proportions $\pi_{i1}, \pi_{i2}, \dots, \pi_{ic}$ respectively. Here again, under age misreporting scenario, the posterior, $h(p^*|x)$, does not have any explicit form and hence it is evaluated by the following Monte Carlo simulation technique.

4. Numerical Study

In this section the proposed procedures have been illustrated numerically through a simulation study. For demonstration purpose we first draw a random observation from Uniform(0,1) of size c , say p_a , for all $a = 1, 2, \dots, c$, where ' c ' denotes the numbers of groups. By using the same p_a a random number has been generated from $Binomial(n, p_a)$, where assumed $n_1 = n_2 = \dots = n_c = n$, i.e., number of individuals corresponding to each group is the same and $\alpha_c = 1$ for all $a = 1, 2, \dots, c$. The suggested prior and posterior distribution of $\psi(p) = \sum_{a=1}^c \alpha_a p_a$, $\alpha_a \geq 0$, defined in equations (2.13) and (2.14) not have any explicit forms, therefore, the simulation procedure discussed in Section (2.2) will be followed to characterize of $\psi(p)$.

Here we have been computed both ML and Bayesian estimators of $\hat{\psi}(p)^M$ and $\hat{\psi}(p)^B$, respectively, for both perfect and imperfect classification frameworks. Under the assumption of perfect classification of individuals into groups, the comparison among the ML and Bayes' estimators of $\psi(p)$ can be made based using their MSEs under the square risk function as

$$R_{\hat{\psi}(p)^B}(\psi(p)) = E(\hat{\psi}(p)^B - \psi(p))^2, \quad (4.1)$$

$$R_{\hat{\psi}(p)^M}(\psi(p)) = E(\hat{\psi}(p)^M - \psi(p))^2. \quad (4.2)$$

As the posterior mean is obtained by minimizing the Bayes risk under the squared error loss function, the procured Bayes estimator of an unknown parameter has often been found superior to the corresponding ML estimator concerning MSE. It is to be emphasized that the estimator based on ML principal neither depends on any prior distribution for the parameter nor it requires any particular loss function. Thus, in such a situation, the comparison among the ML and Bayes estimator ought to be made so that the criteria do not depend on the nature of prior information regarding unknown parameters. As the MSE of an estimator is also considered risk under squared error loss, it has been treated as a risk function for comparison purposes. The comparison is done by calculating the estimated relative risk of

Bayes estimators concerning $\hat{\psi}(p)^M$ and is defined by

$$\hat{\theta}_{\hat{\psi}(p)^B} = \frac{\hat{R}(\hat{\psi}(p)^M)}{\hat{R}(\hat{\psi}(p)^B)} \quad (4.3)$$

For generalization of the suggested methodology in the imperfect classification of group situation, and comparison of ML and Bayesian technique, the misclassification matrix, π , is known. For the demonstration purpose we have considered the particular form of π , misclassification transition probabilities matrix *i.e.* (π_{a,a^*}) of order $c \times c$ as,

$$\pi = \begin{pmatrix} \rho & \delta & \delta & \dots & \delta \\ & \rho & \delta & \dots & \delta \\ & & \rho & \dots & \delta \\ & & & \ddots & \vdots \\ & & & & \rho \end{pmatrix}, \quad 0 < \rho < 1, \quad \delta = \frac{1-\rho}{s-1}, \quad (4.4)$$

where ρ denotes the probability of an accurately classified group and δ denotes the inaccuracy, which has been assumed as equally distributed across the remaining groups. Here ' $\rho = 1$ ' corresponds to the case of perfect classification. To illustrate the performance of both ML and Bayesian estimators under perfect and imperfect classification frameworks, for different choices of group size $c \in \{3, 5, 7\}$, the number of observation in each group $n = \{50, 100\}$ and $\rho = 0.8, 0.9, 1.0$, estimates $(\hat{\psi}(p)^M, \hat{\psi}(p)^B)$, 95 % confidence and credible intervals and relative risk of Bayes estimators $\hat{\theta}_{\hat{\psi}(p)^B}$ have been obtained. Based on the simulation of 100000 times the obtained results have been depicted in Table 2.

5. Application to Real life data

In this section, an illustration of the proposed procedure using real-life data on Indian married women has been discussed. For this study, we took the data set for the third round of the National Family Health Survey-3 (NFHS-3) for the years 2005-06 from the Measure DHS Demographic and Health Surveys (DHS). DHS provides a nationally representative state survey that helps estimate various key indicators of fertility, infant mortality, family planning practice, maternal and child care, and access to mother and child services (NFHS-III(2005-2006)). NFHS-3 is conducted by the Ministry of Health and Family Welfare (Mo-HFW), Government of India, and managed by the International Institute of Population Sciences (IIPS), Mumbai, covering 29 states and 7 Union Territories of India (NFHS-III(2005-2006)). Here, the samples of NFHS-3 are treated as our population of interest and the study population comprised of the women residing at the Empowered Action Groups (EAG) states of India *viz.* (a) Bihar ($n = 3818$) (b) Uttaranchal ($n = 2953$) (c) Chhatisgarh ($n = 3810$) (d) Jharkhand ($n = 2983$) (e) Orissa ($n = 4540$) (f) Rajasthan ($n = 3892$) (g) Madhya Pradesh ($n = 6427$) and (g) Uttar Pradesh ($n = 12183$), which are considered as socio-economically

backward and have high fertility rates compare to other states.

In order to estimate TMFR, defined in equation (2.5), corresponding to each of the selected Indian states under both ML and Bayesian methods, here married women belonging to childbearing age interval (15-44 years) have been grouped into six ($c = 6$) non-overlapping equal subgroups of 5 year interval viz. 15-19, ..., 40-44. Age interval 45-49 could not be considered due to the lack of a sufficient number of women. Further, the information on the birth status in the last year of the survey has been considered a study period. Corresponding to each selected woman, information regarding their age and whether any birth occurred or not during the study period has been collected.

The problem of estimation of TMFR for the situation where age has been misreported as in equation (3.1), through ML and Bayesian technique, is possible only when the π matrix is known. The present study suggested two different methods to obtain π matrix.

Firstly, we considered the particular form of π , misclassification transition probabilities matrix presuming that the correct age reporting was done at five different levels viz. $\rho = 100\%, 90\%$, and 80% in equation (4.4), where ' $\rho = 100\%$ ' corresponds to the case of perfect or correct age reporting. The impact of perfect or imperfect age reporting has been presented as Table 2 and change in pattern of p_a has been depicted in Figure 1.

Alternatively: The π matrix can be simulated empirically by using independent observation from the same underlying population corresponding to each c age class. The π_{a,a^*} has been estimated as the proportion of women out of total women whose reported age belonging to the age-interval a belongs to the true age-interval a^* in the set of c class. Here, the true age of the mother is determined using the other additional reported information viz. age at first marriage(A_M), duration of Gauna (return marriage) if performed(A_G), marriage to first birth duration(A_{FB}) and age of the first child(A_C). The difference among the reported age(A_R) and age calculated using above information i.e. $A_R - (A_M + A_G + A_{FB} + A_C)$, has been considered as error in reporting. The empirical estimates of π_{a,a^*} , for all a^*, a , have been obtained by repeatedly observing the set of values for sufficiently large number of times, and, finally computed the proportion of cases where the age a^* has been reported as a . The approximate π^E matrix following this procedure based on the available information can be estimated empirically using the whole population. Obtained estimates of TMFR ($\psi(p)$) under different model assumptions are presented as Table 3 and 4. All computations are carried out using Statistical Analysis System (SAS) package, University edition and R package (version-3.4.0).

6. Findings and Discussion

Table 2 depicts the results under both perfect and imperfect age misreporting situations, where Bayesian estimates $\psi(p)^B$ are not only found to be more reliable but also always provide compact and efficient credible interval as compared to $\psi(p)^M$. It also shows that the Bayesian methodology is capable enough to capture the change in estimates due to misclassification in terms of estimation with better accuracy.

The performance of the estimation procedures, as far as TMFR is concerned, has been presented in Tables 3 and 4. Overall, the results indicate that the Bayes estimates of TMFR

Table 1: Empirical estimate of the age misreporting error probability matrix for India

Reported Age-interval (a)	True Age-interval (a^*)					
	15-19	20-24	25-29	30-34	35-39	40-44
15-19	0.992	0.008	0.000	0.000	0.000	0.000
20-24	0.067	0.923	0.010	0.000	0.000	0.000
25-29	0.002	0.113	0.868	0.017	0.000	0.000
30-34	0.000	0.003	0.145	0.832	0.021	0.000
35-39	0.000	0.000	0.005	0.182	0.808	0.004
40-44	0.000	0.000	0.001	0.006	0.173	0.819

for all EAG states have shown a decreasing trend until correct age reporting decreases to 90% and starts increasing after that. As theoretically shown, the ML estimates have shown no impact due to misreporting. Table 3 shows that the Bayes estimates of TMFR under both perfect and imperfect age reporting have been found more precise *i.e.* with lesser risk than those of the ML estimates. The 95% credible intervals based on the Bayes estimators have been found narrower than those obtained using the ML estimates. It implies that the proposed Bayes estimators based on the suggested prior provide estimates more precisely and accurately address the issues of misreporting while estimating TMFR. Among the ML and Bayes estimates generated by using empirically estimated transition probabilities matrix, π^E , in Table 4, results also reveal that Bayes’ estimates of TMFR of selected Indian states are comparatively more precise (with narrower credible intervals). It is also to be emphasized that Bayes’ estimates of TMFR (Table 3) under the presumption that the age has been perfectly reported ($\rho = 1$), corresponding to each Indian state, have been found close to the values of TMFR obtained during 2005-06 *viz.* Uttaranchal (4.0), Uttar Pradesh (5.7), Bihar (5.2), Jharkhand (4.9), Orissa (4.4), Chhattisgarh (4.9), Madhya Pradesh (4.9) and Rajasthan (4.6).

Since the probabilities that a child born to a married woman belonging to a^{th} childbearing age-group, p_a , are sensitive towards age reporting, they are affected immensely due to misreporting. Figure 1 depicts the estimates of p_a based on the Bayesian principle, which shows a significant variation in the pattern in Bayes’ estimates of p_a with a change in levels of the inaccuracy of age reporting corresponding to each Indian state. No systematic pattern has been observed in the obtained estimates, as all states are demographically distinct. Still, variation in Bayes’ estimates of TMFR is expected with a change in levels of misreporting. The degree of distortions in the Bayes estimates of p_a at age a has been noticed comparatively higher than those obtained using the principle of maximum likelihood. In particular, as the proportion of misreporting increases from 5% to 20%, the Bayes estimates of p_a are getting more distorted.

The primary reason for accepting the suggested Bayesian estimates of fertility rates is that the derived prior distribution is subjective and empirical. Here, we have also discussed its formalization and update for imperfect age reporting situations, which demographers or policy-makers routinely experience. Further, we compared the proposed Bayesian estimates

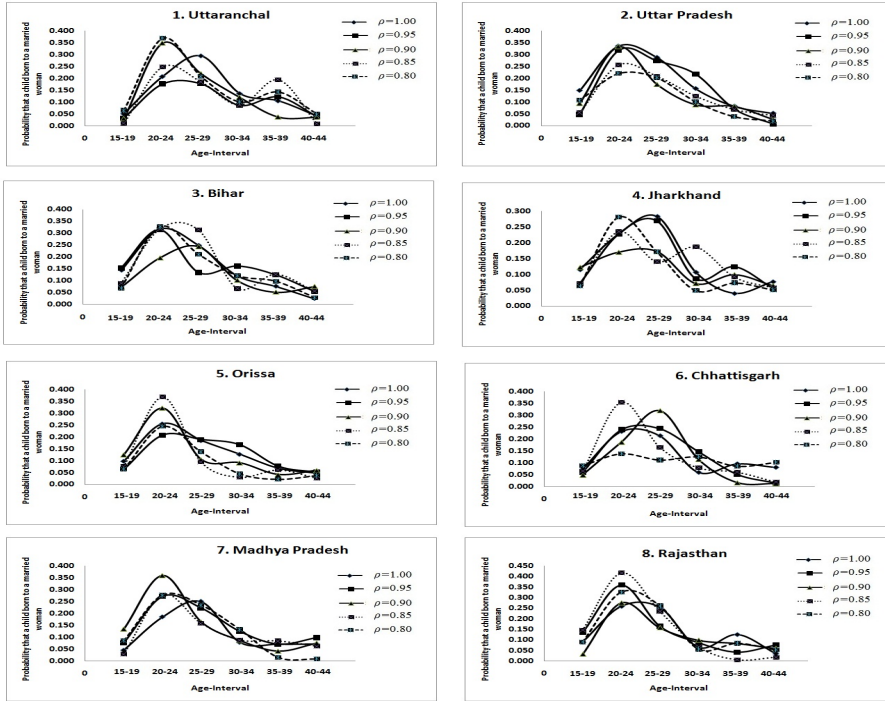


Figure 1: Bayes estimates of the probability that a child born to a married woman belonging to a^{th} childbearing age-group (p_a) in EAG States of India, when there is perfect age reporting ($\rho = 1$) and misreporting lies in 5%- 20%.

with the classical through relative risk, and an attempt has been made to generalize this comparison for the imperfect age-reporting situations. As the likelihood function contains the observation and combining the Bayesian approach with the classical model, the Bayesian approach can incorporate more realistic conditions and data into the estimation.

Table 2: Simulation results of relative risks and their 95% confidence and credible intervals under perfect classification($\rho = 1$) and under misclassification of 0%, 10% and 20%

Class Size	n	ρ	ML			Bayes'			RR
			$\hat{\psi}(p^*)^M$	95% Interval		$\hat{\psi}(p^*)^B$	95% Interval		$(\hat{\theta}_{\hat{\psi}(p^*)^B})$
3	50	1.00	1.960	1.750	2.170	1.954	1.885	2.023	11
		0.90	1.960	1.745	2.175	1.957	1.880	2.035	6
3	150	0.80	1.960	1.740	2.180	1.940	1.856	2.024	6.5
		1.00	1.947	1.833	2.060	1.933	1.893	1.974	3
		0.90	1.947	1.828	2.066	1.949	1.906	1.993	4
		0.80	1.947	1.823	2.070	1.938	1.894	1.983	4
5	50	1.00	3.340	3.105	3.575	3.324	3.275	3.374	14
		0.90	3.340	3.090	3.590	3.261	3.208	3.313	16
5	150	0.80	3.340	3.079	3.601	3.351	3.289	3.413	18
		1.00	3.447	3.303	3.591	3.449	3.415	3.483	5
		0.90	3.447	3.297	3.596	3.440	3.412	3.469	6
		0.80	3.447	3.293	3.600	3.428	3.394	3.462	6
7	50	1.00	4.540	4.289	4.791	4.217	4.179	4.255	16
		0.90	4.540	4.264	4.816	4.565	4.549	4.581	20
7	150	0.80	4.540	4.244	4.836	4.392	4.357	4.427	23
		1.00	4.480	4.336	4.624	4.230	4.193	4.267	5
		0.95	4.480	4.328	4.632	4.346	4.316	4.375	6
		0.90	4.480	4.321	4.639	4.286	4.250	4.322	7
		0.80	4.480	4.309	4.651	4.484	4.467	4.501	8

$$\hat{\theta}^*_{\hat{\psi}(p)^B} = \frac{\hat{R}(\hat{\psi}(p^*)^M)}{\hat{R}(\hat{\psi}(p^*)^B)}$$

7. Conclusion

In the present article, we have derived a prior for total marital fertility rate using Fishers’ information and its related posterior distributions under perfect age reporting and generalized for misreporting scenarios. Since the posterior distributions of TMFR (in the Bayesian paradigm) are complicated, a direct comparison with the maximum likelihood principle (in connection with classical framework) is not straightforward. Thus, through simulation, a comparison among classical and Bayes’ estimates of TMFR is presented. Both the simulated and real-life based results show that the suggested Bayesian estimators of $\psi(p)$ and TMFR lead to population parameters more closely than classical ML estimators and are much more precise than maximum likelihood estimates, even in imperfect scenarios. As evident from the obtained results, even with inaccuracy in age reporting, the Bayesian technique has been found most promising for estimating TMFR, and obtained Bayes’ estimates are more precise and reliable than those obtained using the maximum likelihood procedure.

To conclude, apart from the estimation of transition probabilities, the Bayesian technique has been found to be more useful in estimating the pattern of fertility rates even in situations where there is inaccuracy in age reporting.

Table 3: Estimates of TMFR ($\psi(p)$) in EAG States of India, when there is perfect age reporting($\rho = 1$) and misreporting is 10%, and 15%

State	n	ML			Bayes'			$\hat{\theta}_{\hat{\psi}(p^*)^B}$
		$\hat{\psi}(p^*)^M$	95% Interval		$\hat{\psi}(p^*)^B$	95% Interval		
$\rho = 1.00$								
Uttaranchal	2953	3.08	2.76	3.39	4.17	4.00	4.34	3.7
Uttar Pr.	12183	3.54	3.37	3.71	5.30	5.24	5.37	7.0
Bihar	3818	3.77	3.45	4.08	4.66	4.57	4.75	6.5
Jharkhand	2983	3.26	2.93	3.58	4.24	4.14	4.34	6.8
Orissa	4540	2.38	2.15	2.60	3.90	3.82	3.99	6.5
Chhattisgarh	3810	2.69	2.43	2.96	3.79	3.71	3.86	3.6
Madhya Pr.	6427	2.92	2.71	3.14	3.50	3.36	3.63	2.4
Rajasthan	3892	3.40	3.11	3.69	4.12	3.98	4.26	4.4
$\rho = 0.90$								
Uttaranchal	2953	3.08	2.76	3.40	3.95	3.88	4.03	6.8
Uttar Pr.	12183	3.54	3.37	3.72	3.97	3.89	4.06	4.0
Bihar	3818	3.77	3.45	4.09	4.20	4.13	4.26	6.8
Jharkhand	2983	3.26	2.93	3.58	3.49	3.41	3.58	7.0
Orissa	4540	2.38	2.15	2.61	3.72	3.60	3.84	3.5
Chhattisgarh	3810	2.69	2.42	2.96	3.48	3.44	3.51	6.3
Madhya Pr.	6427	2.92	2.71	3.14	4.29	4.29	4.52	4.0
Rajasthan	3892	3.40	3.10	3.70	3.51	3.41	3.62	7.7
$\rho = 0.85$								
Uttaranchal	2953	3.08	2.75	3.40	3.68	3.59	3.76	9.3
Uttar Pr.	12183	3.54	3.37	3.72	3.78	3.63	3.93	1.3
Bihar	3818	3.77	3.44	4.09	4.86	4.80	4.92	5.4
Jharkhand	2983	3.26	2.92	3.59	3.91	3.80	4.03	9.7
Orissa	4540	2.38	2.14	2.61	3.28	3.17	3.38	4.7
Chhattisgarh	3810	2.69	2.42	2.96	3.72	3.64	3.79	9.5
Madhya Pr.	6427	2.92	2.71	3.14	3.49	3.39	3.59	4.0
Rajasthan	3892	3.40	3.10	3.70	4.44	4.44	4.75	4.6

Table 4: Estimates of TMFR ($\psi(p)$) in EAG States of India, under imperfect age-reporting using empirical π^E .

State	ML			Bayes'			$\hat{\theta}_{\hat{\psi}(p)^B}$
	$\hat{\psi}(p)^M$	95% Interval		$\hat{\psi}(p)^B$	95% Interval		
Uttaranchal	3.187	2.861	3.514	4.108	3.995	4.221	9.3
Uttar Pardesh	3.656	3.482	3.83	4.513	4.37	4.656	1.6
Bihar	3.875	3.552	4.198	3.758	3.649	3.868	9.0
Jharkhand	3.369	3.036	3.701	4.243	4.145	4.341	9.7
Orissa	2.464	2.232	2.697	3.697	3.587	3.808	4.7
Chhattisgarh	2.793	2.52	3.066	3.988	3.932	4.043	6.3
Madhya Pradesh	3.027	2.808	3.246	3.742	3.606	3.878	2.4
Rajasthan	3.513	3.214	3.812	4.022	3.858	4.187	3.3

Acknowledgements

The authors would like to thank all anonymous reviewers and the editorial team for their valuable comments and suggestions that helped us improve the article’s quality.

References

Barakat, B. F., Dharamshi, A., Alkema, L. and Antoninis, M., (2021). Adjusted Bayesian Completion Rates (ABC) Estimation (No. at368). Center for Open Science.

Bhat, P. M., (1990). Estimating transition probabilities of age misstatement. *Demography*, 27(1), pp. 149–163.

Bhat, P. N., (1995). Age misreporting and its impact on adult mortality estimates in South Asia. *Demography India*, 24(1), pp. 59–80.

Bhatta, D., Nandram, B., (2013). A Bayesian adjustment of the HP law via a switching nonlinear regression model. *Journal of Data Science*, 11(1), pp. 85–108.

Borges, G. M., (2019). A Bayesian framework for estimating fertility from multiple data sources. *Anais*, pp. 1–8.

Coale, A. J., Li, S., (1991). The effect of age misreporting in China on the calculation of mortality rates at very high ages. *Demography*, 28(2), pp. 293–301.

Coal Ansley J., Trussell T. James, (1974). Model Fertility Schedules: Variation in the Age Structure of Childbearing in Human Populations. *Population Index*. 40(2), pp. 191–228.

- Daponte, B. O., Kadane, J. B. and Wolfson, L. J., (1997). Bayesian demography: projecting the Iraqi Kurdish population, 1977-1990. *Journal of the American Statistical Association*, 92(440), pp. 1256–1267.
- Datta, G. S., Ghosh, J. K., (1995). On priors providing frequentist validity for Bayesian inference. *Biometrika*, 82(1), pp. 37–45.
- Dechter, A. R., Preston, S. H., (1991). Age misreporting and its effects on adult mortality estimates in Latin America. *Population Bulletin of the United Nations*, (31-32), pp. 1–16.
- Denic, S., Saadi, H. and Khatib, F., (2004). Quality of age data in patients from developing countries. *Journal of Public Health*, 26(2), 168–171.
- Garenne, M., Tollman, S., Kahn, K., Collins, T. and Ngwenya, S., (2001). Understanding marital and premarital fertility in rural South Africa. *Journal of Southern African Studies*, 27(2), pp. 277–290.
- Hussey, J. M., Elo, I. T., (1997). Cause specific mortality among older African Americans: Correlates and consequences of age misreporting. *Social Biology*, 44(3-4), pp. 227–246.
- Liu, P., Raftery, A. E., (2018). Accounting for Uncertainty About Past Values In Probabilistic Projections of the Total Fertility Rate for All Countries. *arXiv preprint arXiv:1806.01513*.
- Martin, J. A., Hamilton, B. E., Ventura, S. J., Osterman, M. J., Kirmeyer, S., Mathews, T. J. and Wilson, E. C., (2011). Births: final data for 2009. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System, 60(1), pp. 1–70.
- Murray, C. J., Callender, C. S., Kulikoff, X. R., Srinivasan, V., Abate, D., Abate, K. H., ... and Bililign, N., (2018). Population and fertility by age and sex for 195 countries and territories, 1950?2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159), pp. 1995–2051.
- NFHS-III. (2005-2006). National Family Health Survey, International Institute for Population Sciences, Bombay.
- Narasimhan, R. L., Retherford, R. D., Mishra, V. K., Arnold, F. and Roy, T. K., (1997). Comparison of Fertility Estimates from India's Sample Registration System and National Family Health Survey.

- Neal, S., Matthews, Z., Frost, M., Fogstad, H., Camacho, A. V. and Laski, L., (2012). Childbearing in adolescents aged 12–15 years in low resource countries: a neglected issue. New estimates from demographic and household surveys in 42 countries. *Acta obstetricia et gynecologica Scandinavica*, 91(9), pp. 1114–1118.
- Nwogu, E. C., Okoro, C., (2017). Adjustment of Nigeria population censuses using mathematical methods. *Canadian Studies in Population*, 44(3-4), pp. 149–64.
- Oh, J., (2018). A comparison and prediction of total fertility rate using parametric, non-parametric, and Bayesian model. *The Korean Journal of Applied Statistics*, 31(6), pp. 677–692.
- Pullum, T. W., (2006). An assessment of age and date reporting in the DHS Surveys 1985–2003.
- Pathak, P., Verma, V., (2013). Projection of Indian Population by Using Leslie Matrix with Changing Age Specific Mortality Rate, Age Specific Fertility Rate and Age Specific Marital Fertility Rate. In *Advances in Growth Curve Models*. Springer, New York, NY, pp. 227–240.
- Schoumaker, B., (2020). Fertility estimates from full birth histories and HDSS. In United Nations Expert Group Meeting on Methods for the World Population Prospects 2021 and Beyond.
- Schmertmann, C. P., Hauer, M. E., (2019). Bayesian estimation of total fertility from a population's age sex structure. *Statistical Modelling*, 19(3), pp. 225–247.
- Singh, B. P., Singh, N. and Singh, S., (2020). Estimation of total fertility rate: an indirect approach using auxiliary information. *Journal of the Social Sciences*, 48(3), pp. 789–798.
- Szoitysek, M., Poniat, R. and Gruber, S., (2017). Age heaping patterns in Mosaic data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, pp. 1–26.
- Yadava R. C., Kumar A., (2002). On an Indirect Estimation of Total Fertility Rate from Open Birth Interval Data, *Demography India*, 31(2), pp. 211–222.
- Yi, Z., (2008). Reliability of age reporting among the Chinese oldest-old in the CLHLS datasets. In *Healthy Longevity in China*. Springer, Dordrecht, pp. 61–78.

Appendix

Proof of Theorem 1: Let X_{ai} be a binary variable, denoting the birth status of i^{th} woman belonging to a^{th} age-group and (p_a) be the probability that a child birth occurred to a married woman in the same age-group, for all $i = 1, 2, \dots, n_a$ and $a = 1, 2, \dots, c$. Let $\psi(p) = \sum_{a=1}^c \alpha_a p_a$ be the linear function of probabilities, p_1, \dots, p_c , $\mathcal{I}^{-1}(p)$ be the inverse of Fisher's information matrix and $D_\psi^T(p)$ denote the gradient of $\psi(p)$, where $p = \{p_1, \dots, p_a, \dots, p_c\}$. Let us consider

$$\begin{aligned} \gamma^T(p) &= \frac{D_\psi^T(p) \mathcal{I}^{-1}(p)}{\sqrt{D_\psi^T(p) \mathcal{I}^{-1}(p) D_\psi(p)}} = \left[\frac{\frac{\alpha_1 p_1 (1-p_1)}{n_1}}{\sqrt{\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a}}} \quad \dots \quad \frac{\frac{\alpha_c p_c (1-p_c)}{n_c}}{\sqrt{\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a}}} \right] \\ &= [\gamma_1(p) \quad \gamma_2(p) \quad \dots \quad \gamma_c(p)], \end{aligned} \quad (7.1)$$

where $\gamma_a(p) = \frac{\alpha_a p_a (1-p_a)}{n_a} \left(\sqrt{\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a}} \right)^{-1}$. In the context of deriving a prior distribution of a parameter, Dutta and Ghose (1995) has suggested the criteria that must be satisfied to establish the posterior distribution for a parametric function under which $\sum_{a=1}^c \frac{\partial}{\partial p_a} \gamma_a(p) \tau(p) = 0$.

Let

$$\tau(p) = \left(\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a} \right)^{1/2} \prod_{a=1}^c p_a^{-1} (1-p_a)^{-1}$$

then

$$\begin{aligned} \gamma_1(p) \tau(p) &= \frac{\frac{\alpha_1 p_1 (1-p_1)}{n_1}}{\sqrt{\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a}}} \left(\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a} \right)^{1/2} \prod_{a=1}^c p_a^{-1} (1-p_a)^{-1} \\ &= \frac{\alpha_1}{n_1} \prod_{a \neq 1}^c p_a^{-1} (1-p_a)^{-1} \Rightarrow \frac{\partial}{\partial p_1} \gamma_1(p) \tau(p) = 0 \end{aligned} \quad (7.2)$$

and

$$\begin{aligned} \gamma_j(p) \tau(p) &= \frac{\frac{\alpha_j p_j (1-p_j)}{n_j}}{\sqrt{\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a}}} \left(\sum_{a=1}^c \frac{\alpha_a^2 p_a (1-p_a)}{n_a} \right)^{1/2} \prod_{a=1}^c p_a^{-1} (1-p_a)^{-1} \\ &= \frac{\alpha_j}{n_j} \prod_{a \neq j}^c p_a^{-1} (1-p_a)^{-1} \Rightarrow \frac{\partial}{\partial p_j} \gamma_j(p) \tau(p) = 0 \end{aligned} \quad (7.3)$$

From the above equations (7.2) and (7.3) we have

$$\sum_{a=1}^c \frac{\partial}{\partial p_a} \gamma_a(p) \tau(p) = 0,$$

which satisfied the condition required to be a prior distribution, $\tau(p)$, of a parameter. Therefore,

$$\tau(p) \propto \left\{ \sum_{a=1}^c \alpha_a^2 p_a (1-p_a) \right\}^{1/2} \prod_{a=1}^c p_a^{-1} (1-p_a)^{-1} ; 0 < p_a < 1$$

and hence we get the required proof.

The prediction of new Covid-19 cases in Poland with machine learning models

Adam Chwila¹

Abstract

The COVID-19 pandemic has had a huge impact both on the global economy and on everyday life in all countries all over the world. In this paper, we propose several possible machine learning approaches to forecasting new confirmed COVID-19 cases, including the LASSO regression, Gradient Boosted (GB) regression trees, Support Vector Regression (SVR), and Long-Short Term Memory (LSTM) neural network. The above methods are applied in two variants: to the data prepared for the whole Poland and to the data prepared separately for each of the 16 voivodeships (NUTS 2 regions). The learning of all the models has been performed in two variants: with the 5-fold time-series cross-validation as well as with the split into the single train and test subsets. The computations in the study used official statistics from government reports from the period of April 2020 to March 2022. We propose a setup of 16 scenarios of the model selection to detect the model characterized by the best ex-post prediction accuracy. The scenarios differ from each other by the following features: the machine learning model, the method for the hyperparameters selection and the data setup. The most accurate scenario for the LASSO and SVR machine learning approaches is the single train/test dataset split with data for the whole Poland, while in case of the LSTM and GB trees it is the cross validation with data for whole Poland. Among the best scenarios for each model, the most accurate ex-post RMSE is obtained for the SVR. For the model performing best in terms of the ex-post RMSE, the interpretation of the outcome is conducted with the Shapley values. The Shapley values make it possible to present the impact of auxiliary variables in the machine learning model on the actual predicted value. The knowledge regarding factors that have the strongest impact on the number of new infections can help companies to plan their economic activity during turbulent times of pandemics. We propose to identify and compare the most important variables that affect both the train and test datasets of the model.

Key words: machine learning, time series, COVID-19, forecasting, economic activity.

¹ University of Economics in Katowice, Katowice, Poland. E-mail: achwila@gmail.com.
ORCID: <https://orcid.org/0000-0003-4671-4298>.



1. Introduction and literature review

As of the start of the second quarter of 2022, the world is still struggling with the outbreak of the Covid-19 pandemic. The first official case of Covid-19 in Poland was registered on March 4, 2020, and as of 17 December 2020 the sum of all confirmed cases since March 2020 was equal to 1.17 million (Ministry of Health Republic of Poland, 2022). The predictions of the daily new infections can be very helpful in several different areas: the preparation of hospitals and medical services, the introduction of new restrictions that potentially can reduce the dynamic of the pandemic, and the plans regarding the future economic activity of the companies. They can also influence the development of vaccination programs.

Overall, the dynamic of the pandemic occurred to be a non-trivial issue due to many factors that can potentially influence the number of new infections. It also causes the forecasting of new confirmed cases much more challenging. Therefore, the application of the machine learning models that can potentially deliver accurate predictions based on non-linear dependencies is worth researching. This paper is structured as follows: Section 2 discusses the applied models, Section 3 describes the considered datasets as well as the concept of time series cross-validation, Section 4 discusses the results, Section 5 discusses limitations and the proposition of the future research, Section 6 summarizes the research.

Many researchers have successfully applied different forecasting approaches at different stages of pandemic development. There were several attempts to forecast the dynamic of the Covid-19 outbreak with the compartmental epidemiology models: in Italy (Giordano et al., 2020), China, Italy, and France (Fanelli, Piazza, 2020), Japan, Singapore, South Korea, and Italy (Chen, Lu, 2020), China, South Korea, Australia, USA and Italy (Cooper et al., 2020), Nigeria (Okuonghae, Oname, 2020). The different variants of compartmental models are mainly the modifications of the SIR susceptible-infected-removed model, which based on different parameters predicts the curves of pandemic dynamics (Kermack, McKendrick, 1927). However, these models focus mainly on the prediction of the whole pandemic dynamics, rather than on the daily changes based on the most recent data. Some studies regarding Covid-19 SIR models concluded that these models can be very sensitive to the assumed parameter describing the fraction of asymptomatic cases (Arino, Portet, 2020). The number of daily new confirmed cases was predicted with the ARIMA model with application to the data from January to February 2020 (Benvenuto et al., 2020).

Various classes of machine learning models have been applied to the Covid-19 data: Support Vector Machines regression, based on the lagged values of new daily confirmed cases (Peng, Nagata, 2020), logistic model (Wang et al., 2020), the long-short term neural network model for data in Canada (Chimmula, Zhang, 2020) and in India (Tomar, Gupta, 2020). The studies that involved finding the countries with similar

dynamics of Covid-19 outbreak with k-means and hierarchical analyses have been also conducted (Aydin, Yurdakul, 2020). The dependence on the mortality rate associated with the Covid-19 outbreak and the weather conditions with machine learning models have been studied for the data for Italy (Malki et al., 2020). Some studies regarding forecasting the new Covid-19 infections pointed out the following challenges associated with the complex machine learning models: the small datasets of historical data regarding the Covid-19 pandemic (less than 150 observations) and the inclusion of the variables connected to the government restrictions (Ahmad et al., 2020). In this paper, we try to refer to both issues, by the usage of the dataset with more than 250 daily records of data as well as the introduction of the variables associated with government restrictions. There were also studies regarding the performance of different machine learning methods (i.e. neural networks and Support Vector Machines) on small Covid-19 datasets (Fong et al., 2020). The cubic regression was also applied to the Covid-19 data from China (Gu et al., 2020). In the case of the new confirmed cases in Greece, there was a suggested network-defined splines model (Demertzis et al., 2020). The attempt to estimate the unobserved Covid-19 infections with an unbiased hierarchical Bayesian estimator with the auxiliary variable of current fatalities has been conducted for North American data (Vaid et al., 2020). Besides the application of machine learning methods in the case of pandemic forecasting, there has been a lot of research that compared the performance of machine learning methods with classic approaches. In the case of medical applications, there is a study comparing the performance of Support Vector Machines and neural networks with logistic regression for the problem of a number of oocytes retrieved, where the accuracy of machine learning models was higher than for the logistic regression (Barnett-Itzhaki et al., 2020). A study focused on ozone concentration prediction (Jumin et al., 2020) showed that Gradient Boosted trees outperformed the performance of linear regression and neural network models. In the case of air pollution concentration (Chen et al., 2019) the comparative study of different algorithms showed that generalized boosted model, random forest, and bagging outperformed backward stepwise linear regression, Support Vector Regression, and neural networks. There was also a study that analyzed results from 14 different articles based on the Covid-19 modelling with supervised and unsupervised methods (Kwekha-Rashid et al., 2021). The authors concluded that machine learning can produce an important role in COVID-19 investigations, prediction, and discrimination. Additionally, it can be involved in the health provider programs and plans to assess and triage the COVID-19 cases (Kwekha-Rashid et al., 2021).

To sum up and compare our work with different approaches taken by the researchers in the above studies over Covid-19 we can differentiate the following:

- The studies focused on compartmental epidemiology models, mainly the modifications of the SIR susceptible-infected-removed model. These methods aim

to deliver the long-term scenarios of the pandemic (i.e. 2 year horizon), based on strict assumptions i.e. that people who recovered from the disease are not going to get infected again. The goal of these models is quite different than the aim of the author's study, which is short-term prediction and explanation of the auxiliary variables actual impact on the new daily cases. The authors of SIR models aim to produce realistic predictions in the long horizon. Our goal is to accurately predict new daily cases in a horizon of 1-7 days and point out the variables that have the highest impact on the actual predictions.

- Autoregressive models taking into account solely lagged values of the new confirmed cases. In the following study, we include the component of lagged values of the new confirmed cases. Also, the additional auxiliary variables are considered to obtain more accurate results.
- Unsupervised machine learning models, i.e. to find the countries with similar dynamics of Covid-19 outbreak. A study with similar applications could be conducted with the NUTS-2 regions for the whole Poland. Nevertheless, it is out of the scope of the proposed research, which is focused on short-term predictions and the explainability of different factors considered in the modelling process. The unsupervised methods are designed to solve problems of a different nature than the considered supervised machine learning models (LASSO, SVR, LSTM, and GB trees).
- Supervised machine learning methods, focused on the short-term predictions of the new cases or similar statistics (i.e. fatalities). The research focuses on the whole spectrum of machine learning methods, either one or several different for comparison purpose. In our study we also focus on the choice of several machine learning methods:
 - the considered LASSO model is the linear regression with only one additional hyperparameter. It is the simplest among the considered methods, present to evaluate if the more complex methodologies significantly improve predictions,
 - the GB trees and SVR are the models that in different ways aim to take into account nonlinear relationships between variables,
 - the LSTM neural network is the most complex among the considered methods – an interesting aspect of the study is the comparison of the LSTM with GB trees/SVR and the LASSO (modified linear regression) in terms of stability and prediction power.

Although we consider 4 machine learning methods, the arbitrary choice of the considered methods is one of the limitations of our study. Hence, additionally we decided to choose models from 3 different levels of complexity to obtain a satisfying

range of results and evaluate if more complex approaches are better than the simpler ones. In this paper we propose a comparison of several different machine learning approaches with the setup, which based on our best knowledge is not presented in the literature:

- taking into consideration the complexity of the time series of new confirmed cases we propose different scenarios for the application of machine learning models: the models trained on the single train and test subsets as well as with 5-fold time series cross validation. The methods of different train and test data splits result in different hyperparameters chosen for the final model form,
- the study aims to compare the models trained on the times series data collected for the whole country, with the models trained on the data collected separately for each of the 16 voivodeships (NUTS 2 regions),
- studies presented in the current literature rarely compare the different machine learning approaches with each other when it comes to modelling Covid-19 data. Even if several machine learning models are proposed, the Long-Short Term Memory (LSTM) neural networks are not compared with other, less complex techniques. In this study, LSTM networks are in the scope with other methods,

Additionally, the study aims to deliver some insight into the impact of different factors on the actual new confirmed Covid-19 cases. The proposed models consider 38 variables collected from different sources. In different studies, it was analysed whether restrictions of movements can significantly influence the transmission of Covid-19 (Nouvellet et al., 2021). Therefore, the considered factors include:

- daily weather data,
- Covid-19 daily statistics (new confirmed cases, fatalities, tests, etc.)
- vaccinations data,
- Covid-19 Variants of Concern and Variants of interest data,
- place and time indicators,
- general policy and government restrictions,
- Covid-19 international indexes for Poland (i.e. containment health index),
- people mobility data.

The impact of the different considered factors on the actual number of new confirmed cases is an important aspect of the studies. We aim to use explanatory methods of complex machine learning models to detect the most important variables. The explanatory method is considered for the model with the best predictive power among the considered scenarios. We use the idea of the Shapley values (Shapley, 1953), successfully applied by Lundberg and Lee (2017), to explain the impact of factors on the model.

2. Models and methods

This study aims to predict the daily number of new infections in Poland based on the data from the two previous days. The compared machine learning models are Least Absolute Shrinkage and Selection Operator (LASSO) regression, Gradient Boosted (GB) regression trees, Support Vector Regression (SVR), and Long-Short Term Memory (LSTM) recurrent neural network. Each of these models is estimated for the data collected for the whole Poland as well as separately for 16 voivodeships (NUTS 2 regions). In the case of models estimated for the voivodeships, the sum of 16 predictions gives the prediction for new infections in Poland. In addition, each of the models is estimated in 2 variants of establishing hyperparameters, which gives 16 models in total (assuming that the set of hyperparameters for the dataset division variants is different for each model). All the models are estimated with code written in Python.

The first of the considered models is the LASSO Regression (Ranstam, Cook, 2018). The LASSO regression can be perceived as the modification of standard linear regression, which is made to reduce overfitting (the poor performance of the model on the dataset on which the model parameters are not trained). It also addresses the problem of the automated feature selection. The parameters of the LASSO regression are obtained by minimizing the modified error function (in this paper RMSE). The modification increases the value of the computed error function by as much as the sum of absolute values of model parameters multiplied by the hyperparameter α (Ranstam, Cook, 2018).

The exact value of α is established during the model training process. In this paper the LASSO regression is performed with Scikit-learn Python library (Pedregosa et al., 2011) with a function `sklearn.linear_model.Lasso`, where the LASSO regression parameters $\hat{\beta}$ are estimated by minimization of the equation:

$$\frac{1}{2n} ||\mathbf{X}\hat{\beta} - \mathbf{Y}||_2^2 + \alpha ||\hat{\beta}||_1, \quad (1)$$

where n is the number of samples based on which the LASSO parameters are estimated, $\hat{\beta}$ is the vector of estimated parameters, \mathbf{X} is the matrix of auxiliary variables from the sample, \mathbf{Y} is the vector of the modelled variable from the sample, α is a regularization hyperparameter with a value chosen by the researcher.

Because of the model error modification, all the considered auxiliary variables need to be normalized or standardized. The standardization and normalization min-max of the auxiliary feature is made with the following equations, respectively:

$$x_{is} = \frac{x_i - \bar{x}}{\sigma}, \quad (2)$$

$$x_{in} = \frac{x_i - \min(x)}{\max(x) - \min(x)}, \quad (3)$$

where x_i is i th instance of the considered variable, x_{is} is the standardized i th instance of the considered variable, \bar{x} is the mean of all instances of the considered dataset, σ is

the standard deviation of the considered dataset, $\min(x)$ is the minimum value of the considered dataset, $\max(x)$ is the maximum value of the considered dataset. The normalization min-max rescales the feature range to be $[0, 1]$. The mean, standard deviation as well as minimum and maximum values are always computed for the training subset. Then the values computed for the training subset are applied to the testing subset, which is done in order to avoid information leak between subsets. In the case of the LASSO regression, models with standardization, normalization min-max and no preprocessing of the data are tested.

The second considered model is Support Vector Regression (SVR) (Vapnik et al., 1994). It introduces the nonlinear relationships between the auxiliary features with the usage of Kernel functions (Sato et al., 2008). In this paper, the radial basis function kernel is considered, which can generalize the infinite-degree polynomial with the single hyperparameter $\gamma > 0$ which controls the influence of a single learning sample (Peng, Nagata, 2020), given by:

$$\kappa(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \quad (4)$$

where $\kappa(x_i, x_j)$ is the radial basis kernel of samples x_i and x_j .

Another feature of the SVR is the specific error function minimized during the algorithm learning. The differences between the fitted and real values of the model are accounted for in the computed error only if they are higher than a certain value of hyperparameter ε . Therefore, the minimized function is given by (Peng, Nagata, 2020):

$$L_\varepsilon(y_i, \hat{y}_i) = \begin{cases} |y_i - \hat{y}_i| - \varepsilon, & \text{if } |y_i - \hat{y}_i| > \varepsilon \\ 0, & \text{if } |y_i - \hat{y}_i| \leq \varepsilon \end{cases}. \quad (5)$$

The next SVR hyperparameter is a penalty, which works similar to the α in the LASSO regression, but instead of summing the absolute values of model parameters, it takes the squares of model parameters (Hastie et al., 2008). The variables for each of the considered SVR scenarios are standardized. In this paper, the SVR is performed with the Scikit-learn Python library (Pedregosa et al., 2011) with the function *sklearn.svm.SVR*.

The third considered model is Gradient Boosted (GB) regression trees (Friedman, 2001). Decision tree is a very popular machine learning algorithm, which in its basic structure divides data many times into segments (leaves). After dividing the data into segments, the arithmetic mean of the response variable is determined for each leaf (Hastie et al., 2008). The GB algorithm is an enhanced version of the decision tree model. GB trees with standardization or with no preprocessing of the data are considered.

In this paper, the GB tree algorithm is performed with the XGBoost Python library (Chen, Guestrin, 2016) with the function `xgboost.XGBRegressor`. The applied GB trees algorithm is as follows.

1. The subsample of the observations is randomly drawn from the train dataset, the size of the subsample is a hyperparameter defined by the researcher (i.e. 70%).
2. For the subsample drawn in the previous step the decision tree is fitted with the CART algorithm (Breiman et al., 1984). During each following split of the single segment into two separate leaves, the different, randomly drawn subsample of the auxiliary variables is considered (i.e. 80% of variables). The subsample size is a hyperparameter, defined by the researcher.
3. After the creation of a tree the fitted values $\hat{\mathbf{Y}}$ for each observation in the train dataset are calculated.
4. The fitted values are multiplied by learning rate hyperparameter η from range $[0,1]$, i.e. by 0.01. The residuals of the model are calculated with the equation:

$$\mathbf{r}_b = \mathbf{Y} - \eta \hat{\mathbf{Y}}_b. \quad (6)$$

5. Vector \mathbf{Y} is replaced by the residuals obtained in the previous step: $\mathbf{Y} = \mathbf{r}_b$.
6. The algorithm goes back to the first step. Steps 1-5 are repeated B times, where B is a defined hyperparameter, i.e. 500.
7. The final form of GB trees is given by:

$$\hat{\mathbf{Y}}_{boost} = \sum_{b=1}^B \eta \hat{\mathbf{Y}}_b. \quad (7)$$

The fourth considered model is Long-Short Term Memory (LSTM) recurrent neural network (Hochreiter, Schmidhuber, 1997). Recurrent neural network is a method widely used in sequential data modelling (Toharudin et al., 2021). The recurrent neural network is an iterative method that in each iteration estimates the fitted values and additionally takes into consideration the values obtained with the previous iterations of the model. The LSTM is a modified recurrent neural network addressing some issues regarding the learning of the network with the backpropagation algorithm: the vanishing or exploding gradient (Hochreiter, Schmidhuber, 1997). The neural networks introduce complex, nonlinear relationships between variables by the usage of multiple neurons (the number of neurons is a hyperparameter) with nonlinear activation functions (the type of activation function is a hyperparameter). In this paper one-layered LSTMs are considered. In the LSTM network, two types of activation functions are used. The first one is typically a sigmoid function (Chimmula, Zhang, 2020), which gives an output in the range $[0, 1]$ and is used for example to properly scale the output from the previous LSTM iteration. The sigmoid function is given by an equation:

$$Sigmoid(x) = \frac{1}{1+e^{-x}}. \quad (8)$$

The second one is similar as in the case of ordinary neural network and introduces nonlinearity to the structure. In this paper the Rectified Linear Unit (ReLU) function is chosen during the learning process, given by (He et al., 2015):

$$\text{ReLU}(x) = \max(0, x). \quad (9)$$

Other considered hyperparameters are: the distribution from which the weights are initialized (weights are the parameters of the network), the number of epochs (number of iterations based on which the backpropagation algorithm adjusts the weights), and the regularization hyperparameter. The variables for each of the considered LSTM scenarios are normalized with min-max normalization. The LSTM networks are built with the Keras Python library (Gulli, Pal, 2017) with *tf.keras.wrappers.scikit_learn.KerasRegressor* function, which is the implementation of the Scikit-learn (Pedregosa et al., 2011) regressor API for Keras Python library. The optimal weights of the LSTM networks are obtained with the usage of the adam (adaptive moment estimation) algorithm (Kingma, Ba, 2015).

For all 4 models the hyperparameters are established based on the 692 daily observations. In the case of models built for voivodeships the data are extended with 15 zero one variables indicating the voivodeship (NUTS 2 region) affiliation. In the case of the LSTM network, the auxiliary data from all the voivodeships for each day are accumulated in one data row and the output of the model for each row is a vector of 16 fitted/forecasted values of new infections (one for each voivodeship). The modified structure of the data for neural network correctly reflects the time dependencies between the variables, which is important in the LSTM concept.

3. Data and the split into subsets

The modelling of the new daily official confirmed cases of Covid-19 is a challenging issue. During the 24 considered months of the data (since the beginning of the pandemic in Poland) several trends connected to the different conditions have been observed, which can be observed in Figure 1.

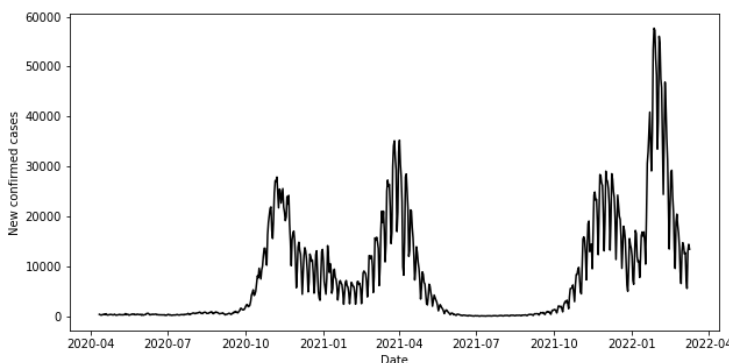


Figure 1: The dynamic of new official Covid-19 infections in Poland, data from 11th of April 2020 to 10th of March 2022

Table A1 in the appendix describes all the relevant data sources and Table A2 in the appendix explains each of the auxiliary variables. The incidental lack of data for some variables for single days are linearly interpolated. In the case of a lack of data on the borders of the considered time series, the nearest observation is assigned to the record with the missing data. The data used for modelling are lagged by 1 or 7 days. The Covid-19 data (new infections, new confirmed cases, Covid-19 variants, vaccinations) and weather data are lagged by one day. The general Covid-19 policy and mobility data are the variables lagged by 7 days because these variables are considered as additional conditions that affect the spread of the virus. The new infections are commonly noticed by affected people after a few days. For example, more strict gatherings restrictions are not supposed to influence the detected new infections the next day after the shift, but rather after at least a few days. The categorical variables are replaced by new 0-1 dummy variables. The official sources of Covid-19 data are connected to some limitations. Not all of the actual new infections of Covid-19 disease are recorded in the official statistics (Vaid et al., 2020). In this paper, the forecast of new confirmed cases is based solely on the official data from the government records. The concept of machine learning models is based on the division of all available data into train and test datasets. The train dataset is used for estimation of the model parameters and establishing the values of hyperparameters during the learning process. The test dataset contains samples which were not used by the researcher in any form during the learning process. Therefore, the predictions made on the test dataset allow assessing the model accuracy (Xu, Goodacre, 2018). In order to correctly compare the different machine learning methods with each other, the same division into train and test datasets should be applied for each method. The split that is often applied for common machine learning tasks in the literature is the train dataset equal to 80% of the available data and the test set equal to 20% (Hastie et al., 2008).

The whole dataset contains 699 records: the daily data of new confirmed cases from 11th April 2020 to 10th March 2022 and the data for the auxiliary variables from 4th April 2020 to 9th March 2022. The last 7 days of data are excluded from the dataset based on which the hyperparameters of the models are established (4th to 10th March 2022). The last 7 days of data are used for the evaluation of the predictive power of models. Therefore, the new confirmed cases based on which the model hyperparameters and parameters are established, are based on the period 11th April 2020 – 3rd March 2022 (692 observations). In machine learning, the hyperparameter is a value that affects the learning process of the given method (Hutter et al., 2014). The range of tested hyperparameters is defined by the researcher before the start of the whole process. The values of the parameters of each model are obtained with the training process.

The hyperparameters and parameters of the models are evaluated in two stages (Hastie et al., 2008). Firstly, the model is learned on the given training subset: for each combination of the hyperparameters, the parameters of the model are established, and then the prediction accuracy RMSE (Root Mean Squared Error) of the forecasts is calculated. The applied RMSEs are given by the following equations:

$$RMSE_{country\ level}^{goodness-of-fit} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}, \quad (10)$$

$$RMSE_{country\ level}^{ex\ post\ accuracy} = \sqrt{\frac{\sum_{i=n}^{n+m} (y_i - \hat{y}_i)^2}{m}}, \quad (11)$$

$$RMSE_{NUTS\ 2\ level}^{goodness-of-fit} = \sqrt{\frac{\sum_{i=1}^n (\sum_{j=1}^k y_{ij} - \sum_{j=1}^k \hat{y}_{ij})^2}{n}}, \quad (12)$$

$$RMSE_{NUTS\ 2\ level}^{ex\ post\ accuracy} = \sqrt{\frac{\sum_{i=n}^{n+m} (\sum_{j=1}^k y_{ij} - \sum_{j=1}^k \hat{y}_{ij})^2}{m}}, \quad (13)$$

where n is the number of observations based on which models are estimated, m is the number of observations for which predictions are made, k is the number of NUTS 2 regions, \hat{y}_i is modelled fitted value, \tilde{y}_i is a forecast, y_i is the real value. The RMSE indicates how the modelled values deviate from the real values on average.

In the next step, the hyperparameters of the model with the lowest value of prediction accuracy RMSE calculated for the testing subset are remembered and the parameters of the model with the remembered hyperparameters are estimated based on the whole considered dataset for the model creation purpose. Then, the performance of the model is established based on the 7 last observations of the dataset – the observations that are not involved in the model creation procedure. If the hyperparameters of the model are chosen based on the procedure of the k -fold cross-validation, then the dataset based on which the set of hyperparameters is established is divided into training and testing subsets k times. For each set of hyperparameters, the prediction accuracy RMSE is calculated for each of the k testing subsets (Hastie et al. 2008). Then for each set of hyperparameters, the average prediction accuracy RMSE calculated on testing subsets is computed and the hyperparameters with the lowest average prediction accuracy RMSE are chosen for parameters estimation based on the whole considered dataset for the model creation purpose. There are 2 scenarios of the division of the dataset into training and testing subsets. The first one is the single division of the dataset into training and testing subsets (where 90% of the observations is in the training subset) and the second one is the 5-fold cross validation adapted for the time series problem (Bergmeir, Benítez, 2012). Given the nature of the dependence of the following observations in the time series, the division of the dataset into training

and testing subsets should not be random, but rather established in a way that the order of the observations is not disrupted. The two ways of the division of the dataset are presented in Figure 2 and Figure 3.

training: 1-623	testing: 624-692
-----------------	------------------

Figure 2: The division of the dataset into single training and testing subsets

training: 1-117	testing: 118-232
training: 1-232	testing: 233-347
training: 1-347	testing: 348-462
training: 1-462	testing: 463-577
training: 1-577	testing: 578-692

Figure 3: The division of the dataset with 5-fold time series cross validation Method

4. Results

Table 1 presents the RMSEs computed for the whole learning subset of 692 observations and the validation subset of 7 observations.

The model characterized by the best prediction power in terms of RMSE among the considered models is SVR trained with a single split of data into training and testing subsets and trained on the data for the whole Poland. The comparison of the real and predicted values for the validation subset is presented in Figure 4.

Table 1: Results of the models learned in different scenarios

Model	Training method: cross validation (cv) or single training/testing split (s)	Data: Poland (pl) or voivodeships (voi)	RMSE for learning subset	RMSE for validation
GB	cv	pl	9.1	1276.8
	s	pl	8.2	1401.3
	cv	voi	465.1	1928.2
	s	voi	455.5	1926.5
LASSO	cv	pl	1688.1	1793.0
	s	pl	1662.7	1502.7
	cv	voi	2816.8	3199.0
	s	voi	2444.6	2366.7
LSTM	cv	pl	215.3	2894.6
	s	pl	1519.8	3710.8
	cv	voi	847.8	4767.6
	s	voi	902.1	4957.0
SVR	cv	pl	147.0	1301.9
	s	pl	780.6	1003.7
	cv	voi	617.3	1270.8
	s	voi	679.0	2522.4

To maintain the consistency between the approaches considering the predictions for the whole Poland and for the voivodeships, the final predictions made for NUTS-2 regions are summed-up and compared to the results obtained on the country level (as indicated in equations 10-13). Because the data for NUTS-2 regions are collected into one dataset with the voivodeship indicator, it means that models automatically tend to focus on the days and NUTS-2 regions with the higher number of new observed cases. For example, the correct prediction for an instance with 10 000 actual new confirmed cases for NUTS-2 region A is more important than a very close prediction for an instance with 50 new cases for NUTS-2 region B. Therefore, the consistency of the approach for analysis of the NUTS-2 regions with data on the country level is maintained. We can observe that the predictions made for the whole Poland and on the level of NUTS-2 regions do not deviate significantly from each other for the RMSE for the validation dataset. The hyperparameter ranges can seriously influence model performance.

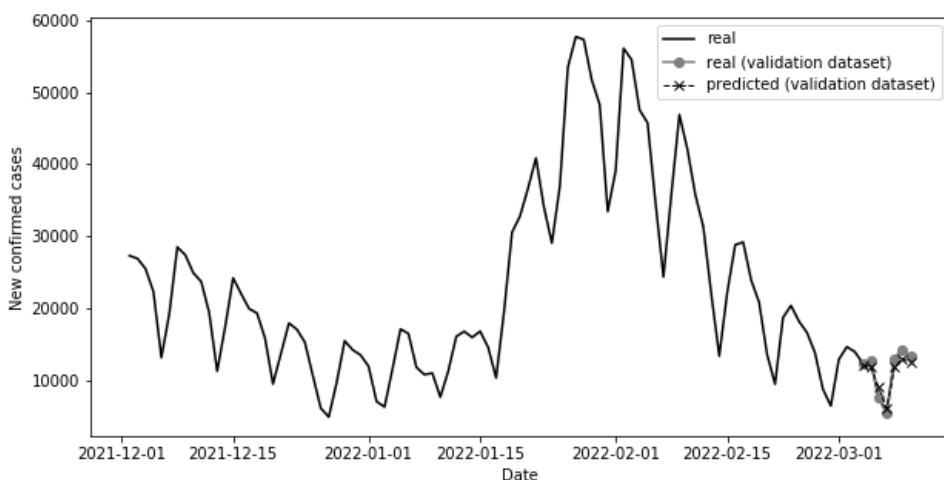


Figure 4: The prediction for 7 observations from the validation dataset by the Support Vector Regression with the lowest ex-post prediction accuracy RMSE

The mean absolute percentage error (MAPE) is given by the following equation:

$$MAPE = \frac{1}{m} \sum_{i=n+1}^{n+m} \frac{|y_i - \tilde{y}_i|}{|y_i|}, \quad (14)$$

where n is the number of observations based on which models are estimated, m is the number of observations for which predictions are made, \tilde{y}_i is a forecast, y_i is the real value. The MAPEs for all considered scenarios are presented in Figure 5.

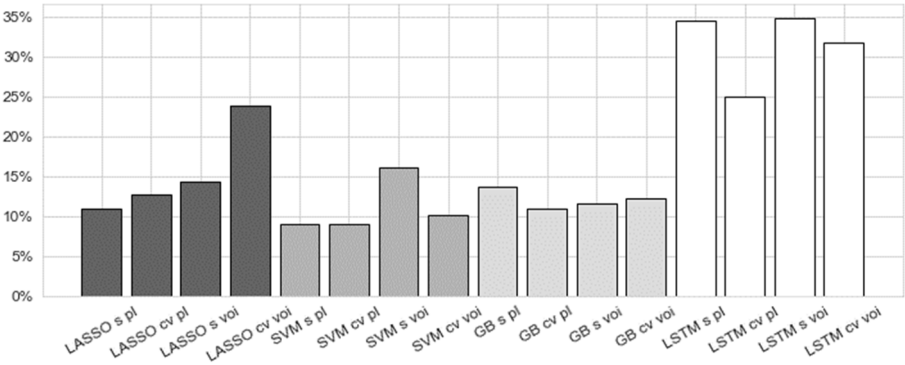


Figure 5: The MAPEs computed for validation dataset for all the considered scenarios

The boxplots of the ex-post predictions errors for 7 values from the validation dataset are presented in Figure 6. They indicate if the given method is characterized by the overfitting or underfitting of the predicted values and also indicate if the value of mean error measures is caused by consistent error values or rather single outlier observations.

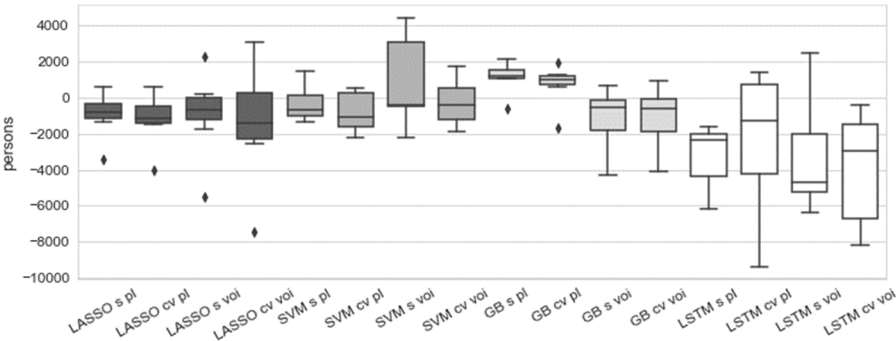


Figure 6: The boxplots of ex-post prediction errors for the validation dataset for all the considered scenarios

The errors produced by SVR and GB models are the most consistent, while in the case of LASSO and LSTM there are several outlier observations.

In addition to the choice of the model characterized by the lowest prediction RMSE we want to establish the impact of the considered factors on the actual predictions made by the model. We use the Shapley values to detect the variables that have the highest impact on the final numbers of new confirmed cases made by the SVR model established on the whole Poland data (with the hyperparameters choice based on the single train-test split of data). The idea of the Shapley values was originally proposed as

the concept of players' contribution in cooperative game theory (Shapley, 1953). In the context of machine learning models, we assume that each auxiliary variable is a "player" in a cooperative game, which contribute in a certain way to the final model prediction (Molnar, 2022). With the Shapley values, we want to estimate how much one of the concrete features impacts the deviation from the average prediction. The Shapley value is the average marginal contribution of a feature value across all possible feature subsets. To compute the exact Shapley value of the j – th feature for a given instance of data, all possible sets of feature values have to be considered (Molnar, 2022). If the overall number of features is relatively high, the computation of the exact Shapley value is very time-consuming. Therefore, we use the following method to estimate the Shapley value for a single feature (Štrumbelj, Kononenko, 2014):

1. We choose the number of iterations M , model f , the dataset X , single instance x , and the j – th feature for which the Shapley value is estimated.
2. For all $m = 1, \dots, M$:
 - a. we draw a random instance z from dataset X , other than x ,
 - b. choose a random permutation p of all the considered features, which includes the j – th feature,
 - c. generate random order of the features in a permutation p ,
 - d. the vectors of auxiliary features for instances x and z are as follow: $x_p = (x_{(1)}, \dots, x_{(j)}, \dots, x_{(k)})$, $z_p = (z_{(1)}, \dots, z_{(j)}, \dots, z_{(k)})$, where k is the total number of features in permutation p ,
 - e. re-train model f on all the instances from the original dataset on the permutation of features p ,
 - f. we construct the two new instances of data by combining the instances x and z : we replace the features placed to the right of j – th feature in instance x , including or excluding the j – th feature from the replacement,
 - g. the created instances are:

$$x_{p\ new} = (x_{(1)}, \dots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \dots, z_{(k)})$$
 and

$$z_{p\ new} = (x_{(1)}, \dots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \dots, z_{(k)}),$$
 - h. we compute the marginal contribution of the j – th feature on the prediction: $\phi_j^m = f(x_{p\ new}) - f(z_{p\ new})$.
3. We compute the Shapley value for instance x as the average marginal contribution:

$$\phi_j(x) = \frac{1}{M} \sum_{m=1}^M \phi_j^m.$$

The above algorithm is repeated for all the features. In this study estimation of the Shapley values is performed with the Shap Python library (Lundberg, Lee, 2017).

The importance of the given feature is computed as the average of the absolute Shapley values for all the considered instances. The distributions of the Shapley values in a form of violin charts for the ten most important features for all the instances from

the training dataset are presented in Figure 7. The respective information for the testing dataset is presented in Figure 8.

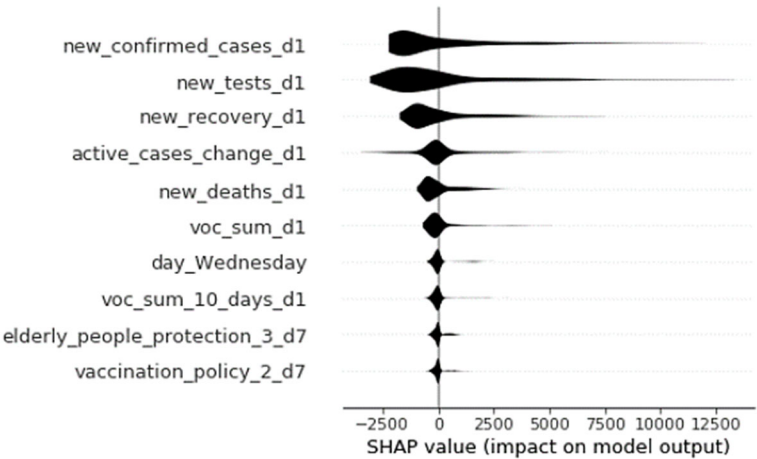


Figure 7: The distributions of the training dataset (692 instances) for the 10 features with the highest average of the absolute Shapley values



Figure 8. The distributions of the testing dataset (7 instances) for the 10 features with the highest average of the absolute Shapley values

The vertical line in Figure 7 and Figure 8 is the baseline (mean of the predictions of new daily confirmed cases). In the case of the training dataset the variables that have the highest impact on the final prediction for a given day are: new confirmed cases from

the previous day, new conducted tests in the previous day, new recoveries from the previous day, the change of active cases from the previous day, the new deaths from the previous day, new VOC and VOI cases reported in the previous day, time indicator for Wednesday, the sum of VOC and VOI occurrences from the last 10 following days (sum from the previous day), the indicator of elderly people protection from 7 days ago with level 3 (which means extensive restrictions for isolation and hygiene) , the indicator of vaccination policy from 7 days ago with level 2 (which means availability of vaccination for medical key workers and clinically vulnerable groups).

In the case of the testing dataset the list of the 10 most important features is quite similar, however, there are some new variables: time indicator for Tuesday, maximum wind speed from the previous day, the change from baseline for traffic congestion in groceries and pharmacies from 7 days ago.

There are some interesting aspects of the study:

- The time indicators for Tuesday and Wednesday are quite important for the final model results – the indicators for other days are less important.
- The high importance of the number of newly conducted tests may indicate that unfortunately, the true number of infections is much higher than officially reported. With the increasing number of tests, the new infections also increase.
- The overall number of vaccinated people is less important for the model than the overall vaccination policy, which may mean that the availability of vaccinations may change the people's behavior, which influenced the mobility.
- The wind speed may be important due to the indirect relationship of less frequent going out from home in the case of high wind speed. Also, the bad weather may indirectly affect the willingness to go out for a Covid-19 test.

The above interpretations of the results are only several of the possible ones.

5. Limitations and future work

One of the limitations of the studies is that it is based on the data from official government reports and there are a certain number of unobserved new infections. Another limitation is that the model can produce the predictions only for the next day, due to the consideration of variables lagged by one day. In future work, the application of variables lagged by more than one can be considered. Another limitation is the arbitrary choice of the concrete period based on which the model results are evaluated (the last 7 days of data). The important limitation of the study is the arbitrary choice of the number of days by which the auxiliary variables are lagged (1 or 7 days). Another limitation is the inclusion of auxiliary variables: the overall number of 38 variables is considered, but other indicators might be also included. The next limitation is the arbitrary choice of the searched ranges of hyperparameters, due to time-consuming

learning process. In future studies, the usage of data from other countries can be considered. Another limitation is the validity of the data. For example, the number of variants of concerns is dependent on the forwarding of the results to the public database GISAID from the different labs, which are not required to send the data. An additional limitation is connected to the methods proposed for the comparative studies (4 different machine learning models), which was an arbitrary choice.

6. Conclusions

We conclude that we propose the setup of 16 scenarios of model selection to detect the model with the best predictive power. The scenarios differ from each other by: machine learning model, the way of hyperparameters selection and the data setup (data for the whole Poland or for each of 16 Polish NUTS-2 regions). The model that produces the lowest error predictions for Covid-19 new daily infections in Poland is the Support Vector Regression model, with ex-post RMSE equal to 1003.7 cases. Ex-post RMSE is an average difference between the actual number of the new cases and the predictions for 7 days. The training process of the model is based on a single split of data into training and testing datasets. For the scenario characterized by the best predictive power, the impact of the auxiliary variables on the final results has been estimated with the Shapley values. Among the factors that have the highest impact on the final results are: Covid-19 statistics (confirmed cases, deaths, recoveries, active cases) from the previous day, Variants of Concern, time indicator for Wednesday, elderly people protection and the general vaccination policy. The machine learning models can help not only successfully predict the different Covid-19 characteristics in the short term periods, but also explain the factors that have the highest impact on the predictions for considered datasets.

Acknowledgment

The author is thankful to Michał Rogalski, who collected the daily Covid-19 data from the government reports, based on which most of the data were prepared: bit.ly/covid19-poland.

References

- Ahmad, A., Garhwal, S., Ray, S., K., Kumar, G., Malebary, S., J., Barukab, O., M., (2020). The number of confirmed cases of Covid-19 by using machine learning: methods and challenges, *Archives of Computational Methods in Engineering*, <https://doi.org/10.1007/s11831-020-09472-8>.

- Arino, J., Portet, S., (2020). A simple model for Covid-19, *Infectious Disease Modelling*, 5, pp. 309–315, <https://doi.org/10.1016/j.idm.2020.04.002>.
- Aydin, N., Yurdakul, G., (2020). Assessing countries' performances against Covid-19 via WSIDEA and machine learning algorithms, *Applied Soft Computing Journal*, 97, p. 106792, <https://doi.org/10.1016/j.asoc.2020.106792>.
- Barnett-Itzhaki, Z., Elbaz, M., Buttermann, R., Amar, D., Amitay, M., Racowskyc, C., Orvieto, R., Hauser, R., Baccarelli, A., Machtinger, R., (2020). Machine learning vs. classic statistics for the prediction of IVF outcomes, *Journal of Assisted Reproduction and Genetics*, 37, pp. 2405–2412, <https://doi.org/10.1007/s10815-020-01908-1>.
- Benvenuto, D., Giovanetti, M., Vasallo, L., Angeletti, S., Ciccozzi, M., (2020). Application of the ARIMA model on the Covid-2019 epidemic dataset, *Data in brief*, 29, p. 105340, <https://doi.org/10.1016/j.dib.2020.105340>.
- Bergmeir, C., Benítez, J. M., (2012). On the use of cross-validation for time series predictor evaluation, *Information Sciences*, 191, pp. 192–213, <https://doi.org/10.1016/j.ins.2011.12.028>.
- Blavatnik School of Government, University of Oxford, (2022). [online] Available at <<https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker>> [Accessed April 30, (2022)].
- Breiman, L., Friedman, J., Olshen, R., Stone, C., (1984). *Classification and regression trees*, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
- Chen, J., Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketznel, M., Bauwelick, M., Donkelaar, A., Hivdtfeldt, U., Katsouyanni, K., Et Al., (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide, *Environment International*, 130, p. 104934, DOI: 10.1016/j.envint.2019.104934.
- Chen, T., Guestrin, T., (2016). XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, DOI: 10.1145/2939672.2939785.
- Chen, Y., Lu, P., (2020). A time-dependent SIR model for Covid-19 with undetectable infected persons, *IEEE Transactions on Network Science and Engineering*, 7(4), pp. 3279–3294, DOI: 10.1109/TNSE.2020.3024723.
- Chimmula V., K., R., Zhang, L., (2020). Time series forecasting of Covid-19 transmission in Canada using LSTM networks, *Chaos, Solitons & Fractals*, 135, p. 109864, <https://doi.org/10.1016/j.chaos.2020.109864>.

- Cooper, I., Mondal A., Antonopoulos C. G., (2020). A SIR model assumption for the spread of Covid-19 in different communities, *Chaos, Solitons & Fractals*, 139, p. 110057, <https://doi.org/10.1016/j.chaos.2020.110057>.
- Daily Temperature In Capital Cities Of Voivodeships In Poland, (2022). [online] Available at: <<https://freemeteo.pl>> [Accessed March 20, (2022)].
- Demertzis, K., Tsiotas, D., Magafas, L., (2020). Modeling and forecasting the covid-19 temporal spread in Greece: an exploratory approach based on complex network defined splines, *International Journal of Environmental Research and Public Health*, 17, p. 4693, doi:10.3390/ijerph17134693.
- Fanelli, D., Piazza, F., (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France, *Chaos, Solitons & Fractals*, 134, p. 109761, <https://doi.org/10.1016/j.chaos.2020.109761>.
- Fong, S., J., Li, N., D., G., Crespo R., G., Herrera-Viedma, E., (2020). Finding an accurate early forecasting model from small dataset: a case of 2019-ncov novel coronavirus outbreak, *International Journal of Interactive Multimedia and Artificial Intelligence*, 6, pp. 132–139, DOI: 10.9781/ijimai.2020.02.002.
- Friedman, J., H., (2001). Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, 29 (5), pp. 1189–1232, DOI: 10.1214/aos/1013203451.
- Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., Colaneri, M., (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy, *Nature Medicine*, 26(6), pp. 855–860, <https://doi.org/10.1038/s41591-020-0883-7>.
- GISAID database, (2022). [online] Available at <<https://www.gisaid.org/>> [Accessed April 30, (2022)].
- Google Covid-19 Community Mobility Reports, (2022) [online] Available at: <<https://www.google.com/covid19/mobility/>> [Accessed March 20, (2022)].
- Gu, C., Zhu, J., Sun, Y., Zhou, K., Gu, J., (2020). The inflection point about covid-19 may have passed, *Science Bulletin*, 65(11), pp. 865–867, DOI: 10.1016/j.scib.2020.02.025.
- Gulli, A., Pal, S., (2017). *Deep learning with Keras*, Packt Publishing Ltd.
- Hastie, T., Tibshirani, R., Friedman, J., (2008). *The Elements of Statistical Learning*, Springer Science + Business Media LLC, New York.
- He, K., Zhang, X., Ren, S., Sun, J., (2015). Delving deep into rectifiers: surpassing human-level performance on ImageNet classification, *Proceedings of the 2015 IEEE*

- International Conference on Computer Vision (ICCV)*, Santiago, Chile, pp. 1026–1034, <https://doi.ieeecomputersociety.org/10.1109/ICCV.2015.123>.
- Hochreiter, S., Schmidhuber, J., (1997). Long Short-Term Memory, *Neural Computation*, 9(8), p. 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>.
- Hutter, F., Hoos, H., Leyton-Brown K., (2014). An efficient approach for assessing hyperparameter importance, *Proceedings of the 31st International Conference on Machine Learning*, 32(1), pp. 754–762.
- Jumin, E., Zaini, N., Ahmed A., Abdullah, S., Ismail, M., Sherif, M., (2020). Machine learning versus linear regression modelling approach for accurate ozone concentrations prediction, *Engineering Applications of Computational Fluid Mechanics*, 14(1), pp. 713–725, <https://doi.org/10.1080/19942060.2020.1758792>.
- Kermack, W. O., Mckendrick, A., G., (1927). A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society A*, 115(772), pp. 700–721, <https://doi.org/10.1098/rspa.1927.0118>.
- Kingma, D., Ba, J., (2015). ADAM: a method for stochastic optimization, *International Conference on Learning Representations 2015*, San Diego, USA, <https://arxiv.org/abs/1412.6980>.
- Kwekha-Rashid, A.S., Abduljabbar, H.N., Alhayani, B., (2021). Coronavirus disease (COVID-19) cases analysis using machine-learning applications, *Applied Nanoscience*, <https://doi.org/10.1007/s13204-021-01868-7>.
- Lundberg, S., M., Lee, S. I., (2017). A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4768–4777.
- Malki, Z., Atlam, E., Hassanien, A., E., Dagnew, G., Elhosseini M., A., Gad, I., (2020). Association between weather data and Covid-19 pandemic predicting mortality rate: machine learning approaches, *Chaos, Solitons & Fractals*, 138, p. 110137, <https://doi.org/10.1016/j.chaos.2020.110137>.
- Molnar C. (2022). *Interpretable machine learning. A guide for making black box models explainable*, Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, [online] Available at: <<https://christophm.github.io/interpretable-ml-book/>> [Accessed April 30, (2022)].
- Ministry of Health Republic of Poland, (2022). [online] Available at: <<https://www.gov.pl/web/coronavirus>> [Accessed March 20, (2022)].

- Nouvellet, P. Et Al., (2021). Reduction in mobility and COVID-19 transmission, *Nature Communications*, 12(1090), <https://doi.org/10.1038/s41467-021-21358-2>.
- Okuonghae, D., Oname, A., (2020). Analysis of a mathematical model for Covid-19 population dynamics in Lagos, Nigeria, *Chaos, Solitons & Fractals*, 139, p. 110032, <https://doi.org/10.1016/j.chaos.2020.110032>.
- Pedregosa, F. Et Al., (2011). Scikit-learn: machine learning in Python, *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Peng, Y., Nagata, M. H., (2020). An empirical overview of nonlinearity and overfitting in machine learning using Covid-19 data, *Chaos, Solitons & Fractals*, 139, p. 110055, <https://doi.org/10.1016/j.chaos.2020.110055>.
- Ranstam, J., Cook, J., A., (2018). LASSO regression, *British Journal of Surgery*, 105, p. 1348, <https://doi.org/10.1002/bjs.10895>.
- R Interface To Covid-19 Data Hub, (2022). [online] Available at <<https://cran.r-project.org/web/packages/COVID19/index.html>> [Accessed March 20, (2022)].
- Shapley, L. S., (1953). A value for n-person games, *Contributions to the Theory of Games*, 2 (28), pp. 307–317.
- Štrumbelj, E., Kononenko I., (2014). Explaining prediction models and individual predictions with feature contributions, *Knowledge and information systems*, 41(3), pp. 647–665.
- Toharudin, T., Pontoh, R. S., Caraka, R. E., Zahroh, S., Lee, Y., Chen, R., C., (2021). Employing long short-term memory and Facebook prophet model in air temperature forecasting, *Communications in Statistics – Simulation and Computation*, pp. 1–24, DOI: 10.1080/03610918.2020.1854302.
- Tomar, A., Gupta, N., (2020). Prediction for the spread of covid-19 in India and effectiveness of preventive measures, *Science of The Total Environment*, 728, p. 138762, <https://doi.org/10.1016/j.scitotenv.2020.138762>.
- Sato, J., R., Costafreda, S., Morettin, P., A., Brammer, M., J., (2008). Measuring time series predictability using Support Vector Regression, *Communications in Statistics – Simulation and Computation*, 37(6), pp. 1183–1197, <https://doi.org/10.1080/03610910801942422>.
- Vaid, S., Cakan, C., Bhandari, M., (2020). Using machine learning to estimate unobserved Covid-19 infections in North America, *The Journal of Bone and Joint Surgery Incorporated*, 102 (70), pp. 1–5, <http://dx.doi.org/10.2106/JBJS.20.00715>.

- Vapnik, V., Levin E., Cun Y. L., (1994). Measuring the vc-dimension of a learning machine, *Neural Computation*, 6(5), pp. 851–76, DOI: 10.1162/neco.1994.6.5.851.
- Wang, P., Zheng, X., Li, J., Zhu, B., (2020). Prediction of epidemic trends in Covid-19 with logistic model and machine learning technics, *Chaos, Solitons & Fractals*, 139, p. 110058, DOI: 10.1016/j.chaos.2020.110058.
- Xu, Y., Goodacre R., (2018). On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning, *Journal of Analysis and Testing*, 2, pp. 249–262.

Appendix

Table A1: Description of data source

Data	Description	Source
Weather	Daily temperature and wind speed data for the capital of voivodeship (in the case of data for whole Poland temperature for Warsaw).	https://freemeteo.pl (Daily temperature in capital cities of voivodeships in Poland, 2022).
Covid-19 data	Data for new Covid-19 confirmed cases, new fatalities, new recoveries, change of active cases, new tests, and new cases/tests ratio. The sum of people fully vaccinated (with two doses or with one of the Johnson & Johnson vaccination) and the sum of people vaccinated with the third dosage. Data are available for voivodeships and the whole Poland.	Data collected based on reports provided by the Ministry of Health, data from the WSSE, PSEZ, Voivodeship Offices, and those obtained in requests for access to public information: https://www.gov.pl/web/coronavirus (Ministry of Health Republic of Poland, 2022).
Covid-19 variants of concern and variants of interest data	The number of Covid-19 variants of concern (VOC) and variants of interest (VOI) reported by different labs analyzing Covid-19 tests in Poland.	The GISAID database: https://www.gisaid.org (Gisaid database, 2022).
Mobility data	The reports for movement trends of people in different places (change from baseline days). The baseline day is always the same day of the week. The value of movement trend in baseline day is the median value from the 5 weeks Jan 3 – Feb 6, 2020. Data are available for voivodeships and the whole Poland.	Google Covid-19 community mobility reports: https://www.google.com/covid19/mobility (Google Covid-19 community mobility reports, 2022).
General Covid-19 policy and government restrictions data	Available variables: closing of schools, closing of workplaces, cancelation of events, gatherings restrictions, closing of transport, stay-at-home restrictions, internal movement restrictions, international movement restrictions, information campaigns, testing policy, contact tracking, facial coverings, vaccination policy, elderly people protection, government response index, stringency index, containment health index, economic support index.	The policy measures from the R package are provided by Oxford Covid-19 Government Response Tracker (Blavatnik School of Government, University of Oxford, 2022): R Interface to COVID-19 Data Hub, 'Covid19' R package: https://cran.r-project.org/web/packages/COVID19/index.html (R interface to Covid-19 data hub, 2022).

Table A2: Description of auxiliary variables

Predictor
Covid-19 data:
- new confirmed cases from the previous day
- new fatalities from the previous day
- new recoveries from the previous day
- change of active cases - state for the previous day
- the number of new conducted tests - data from the previous day
- ratio of new confirmed cases to the number of conducted tests (from the previous day)
Vaccination data:
- the sum of people fully vaccinated (with two doses or with one of Johnson & Johnson vaccination) - state from the previous day
- the sum of people vaccinated with the third dosage - state from the previous day
Place and time indicators:
- weekday indicator: separate zero-one variable for weekday from the previous day
- voivodeship indicator: separate zero-one variable
Covid-19 variants of concern and variants of interest data:
- the number of VOC and VOI: VOC Omicron, VOC Alpha, VOC Delta, VOC Beta, VOC Gamma, VOI Eta, and VOI Lambda. Considered variables: the number of new VOC and VOI cases reported in the previous day and the sum of VOC and VOI occurrences from the last 10 following days
General Covid-19 policy and government restrictions data:
- school closing indicator from 7 days ago (4 levels)
- workplace closing from 7 days ago (4 levels)
- cancelation of events from 7 days ago (3 levels)
- gatherings restrictions from 7 days ago (5 levels)
- transport closing from 7 days ago (3 levels)
- stay home restrictions from 7 days ago (4 levels)
- internal movement restrictions from 7 days ago (3 levels)
- international movement restrictions from 7 days ago (5 levels)
- information campaigns from 7 days ago (3 levels)
- testing policy from 7 days ago (4 levels)
- contact tracking from 7 days ago (3 levels)
- facial coverings from 7 days ago (5 levels)
- vaccination policy from 7 days ago (6 levels)
- elderly people protection from 7 days ago (from no measure to extensive restrictions)
- government response index from 7 days ago
- stringency index from 7 days ago
- containment health index from 7 days ago
- economic support index from 7 days ago
Mobility:
The reports for movement trends of people in different places (change from baseline days). The baseline day is always the same day of the week. The value of movement trend in the baseline day is the median value from the 5 weeks Jan 3 – Feb 6, 2020. The indicator from 7 days ago is considered. The variables are: retail and recreations places, grocery and pharmacy, parks, transit stations, workplaces, residential.
Weather:
- 3 variables: maximum daily temperature, minimum daily temperature and a maximum speed of wind for the capital of voivodeship (in the case of data for whole Poland temperature for Warsaw) from the previous day

New generators for minimal circular generalised neighbour designs in blocks of two different sizes

Muhammad Nadeem¹, Khadija Noreen², H. M. Kashif Rasheed³,
Rashid Ahmed⁴, Mahmood Ul Hassan⁵

Abstract

Minimal neighbour designs (NDs) are used when a response of a treatment (direct effect) is affected by the treatment(s) applied in the neighbouring units. Minimal generalised NDs are preferred when minimal NDs cannot be constructed. Through the method of cyclic shifts (Rule I), the conditions for the existence of minimal circular generalised NDs are discussed, in which $v/2$ unordered pairs do not appear as neighbours. Certain generators are also developed to obtain minimal circular generalised NDs in blocks of two different sizes, where $k_2 = 3, 4$ and 5 . All these designs are constructed using i sets of shifts for k_1 and two for k_2 .

Key words: direct effects, neighbour effects, method of cyclic shifts, generalised NDs, GN2-designs.

Mathematics Subject Classification (2010): 05B05; 62K10; 62K05.

1. Introduction

There are several situations where response of a treatment (direct effect) is affected by the treatment(s) applied in neighbouring units. Such effects are known as neighbour effects which become the major source of bias. Such bias is minimized with the use of neighbour designs (NDs):

- If each pair of treatments appears once as neighbour then it is minimal ND.

¹ Department of Statistics, The Islamia University of Bahawalpur, Pakistan.
ORCID: <https://orcid.org/0000-0002-4034-3564>.

² Department of Statistics, The Islamia University of Bahawalpur, Pakistan.
ORCID: <https://orcid.org/0000-0002-0377-0030>.

³ Department of Statistics, The Islamia University of Bahawalpur, Pakistan.
ORCID: <https://orcid.org/0000-0002-5965-7803>.

⁴ Department of Statistics, The Islamia University of Bahawalpur, Pakistan.
E-mail: <mailto:rashid701@hotmail.com>. ORCID: <https://orcid.org/0000-0001-9703-7296>.

⁵ Department of Statistics, Stockholm University, Stockholm, Sweden.
ORCID: <https://orcid.org/0000-0003-2889-0263>.

- A block formed in a cycle in such a way that its first and last units are considered as adjacent neighbours is called a circular block. In circular blocks, each unit has one left-neighbour and one right-neighbour.
- A block in which each treatment appears either once or not at all is called a binary block.
- The circular generalized neighbour designs (CGNDs) in which $\lambda'_1 = 1$ and $\lambda'_2 = 0$ are called minimal CGNDs (MCGNDs) and are a better alternative to the minimal NDs.

Rees (1967) used neighbour designs in virus research. Partially NDs and generalized NDs (GNDs) should be used in the situations where NDs require a large number of experimental units. Misra *et al.* (1991) relaxed the condition of the constancy of λ' and constructed GNDs. Chaure and Misra (1996), Nutan (2007), Kedia and Misra (2008) constructed GN_2 - designs and GN_3 -designs. Ahmed *et al.* (2009), Zafaryab *et al.* (2010) and Shehzad *et al.* (2011) presented procedures to generate MCGNDs for limited cases. Iqbal *et al.* (2012) presented CGNDs for $k = 3$. Ahmed and Akhtar (2012) presented partially balanced NDs in circular blocks for some specific cases. In the literature, some minimal CGNDs might be constructed through i sets of shifts for k_1 and one set for k_2 , where $k_2 = 3, 4$ and 5 . These designs can also be constructed for other combinations of v , k_1 and k_2 using i sets of shifts for k_1 and two for k_2 which have not been constructed. In this article, generators are developed to obtain MCGNDs in two different block sizes for (i) $k_2 = 3$, (ii) $k_2 = 4$, (iii) $k_2 = 5$. All these designs are constructed through the method of cyclic shifts (Rule I) using i sets of shifts for k_1 and two sets for k_2 . In the proposed designs $v/2$ unordered pairs do not appear as neighbours while all others appear once.

In Section 2, the method of cyclic shifts (Rule I) is described for MCGNDs. In Section 3, conditions are discussed for the existence of MCGNDs in which $v/2$ unordered pairs do not appear as neighbours. In Section 4, generators are developed for MCGNDs in two different block sizes when $k_2 = 4, 5$ and 6 .

2. Method of Construction

The method of cyclic shifts was introduced by Iqbal (1991) for the construction of BIBDs, NDs, RMDs, Polygonal designs, etc. Its Rule I is explained here for MCGNDs.

Let $S_j = [q_{j1}, q_{j2}, \dots, q_{j(k-1)}]$ be l sets of shifts, with $j = 1, 2, \dots, l$ and $1 \leq q_{ji} \leq v-1$. If each of $1, 2, \dots, v-1$ appears once in S^* , where $S^* = [q_{j1}, q_{j2}, \dots, q_{j(k-1)}, (q_{j1} + q_{j2} + \dots + q_{j(k-1)}) \bmod v, v - q_{j1}, v - q_{j2}, \dots, v - q_{j(k-1)}, v - (q_{j1} + q_{j2} + \dots + q_{j(k-1)}) \bmod v]$ then it will provide minimal CND. If (i) $\lambda'_1 = 1$ and $\lambda'_2 = 0$, or (ii) $\lambda'_1 = 1$ and $\lambda'_2 = 2$ then the design is called MCGND.

Example 2.1. The following sets produce MCGND for $\nu = 24$, $k_1 = 5$ and $k_2 = 3$.
 $S_1 = [3,4,5,10], S_2 = [7,11], S_3 = [8,15]$

To generate the design from these sets of shifts, take ν blocks for $S_1 = [3,4,5,10]$. Assign 0, 1, ..., $\nu-1$ as the first unit element for each block respectively. Add 3 (mod ν) to the each first unit element to obtain the second unit elements. Similarly, add 4 (mod ν) to the each second unit element to obtain the third unit elements. Then add 5 and 10 in the similar way and get Table 1.

Table 1: Blocks generated from $S_1 = [3,4,5,10]$

B ₁	B ₂	B ₃	B ₄	B ₅	B ₆	B ₇	B ₈	B ₉	B ₁₀	B ₁₁	B ₁₂
0	1	2	3	4	5	6	7	8	9	10	11
3	4	5	6	7	8	9	10	11	12	13	14
7	8	9	10	11	12	13	14	15	16	17	18
12	13	14	15	16	17	18	19	20	21	22	23
22	23	0	1	2	3	4	5	6	7	8	9
B ₁₃	B ₁₄	B ₁₅	B ₁₆	B ₁₇	B ₁₈	B ₁₉	B ₂₀	B ₂₁	B ₂₂	B ₂₃	B ₂₄
12	13	14	15	16	17	18	19	20	21	22	23
15	16	17	18	19	20	21	22	23	0	1	2
19	20	21	22	23	0	1	2	3	4	5	6
0	1	2	3	4	5	6	7	8	9	10	11
10	11	12	13	14	15	16	17	18	19	20	21

Take ν more subjects for $S_2 = [7,11]$ and generate Table 2 in the similar way as obtained through S_1 .

Table 2: Blocks generated from $S_2 = [7,11]$

B ₂₅	B ₂₆	B ₂₇	B ₂₈	B ₂₉	B ₃₀	B ₃₁	B ₃₂	B ₃₃	B ₃₄	B ₃₅	B ₃₆
0	1	2	3	4	5	6	7	8	9	10	11
7	8	9	10	11	12	13	14	15	16	17	18
18	19	20	21	22	23	0	1	2	3	4	5
B ₃₇	B ₃₈	B ₃₉	B ₄₀	B ₄₁	B ₄₂	B ₄₃	B ₄₄	B ₄₅	B ₄₆	B ₄₇	B ₄₈
12	13	14	15	16	17	18	19	20	21	22	23
19	20	21	22	23	0	1	2	3	4	5	6
6	7	8	9	10	11	12	13	14	15	16	17

Take ν more subjects for $S_3 = [8,5]$ and generate Table 3 in the similar way as obtained through S_1 .

Table 3: Blocks generated from $S_3 = [8,5]$

B ₄₉	B ₅₀	B ₅₁	B ₅₂	B ₅₃	B ₅₄	B ₅₅	B ₅₆	B ₅₇	B ₅₈	B ₅₉	B ₆₀
0	1	2	3	4	5	6	7	8	9	10	11
8	9	10	11	12	13	14	15	16	17	18	19
23	0	1	2	3	4	5	6	7	8	9	10
B ₆₁	B ₆₂	B ₆₃	B ₆₄	B ₆₅	B ₆₆	B ₆₇	B ₆₈	B ₆₉	B ₇₀	B ₇₁	B ₇₂
12	13	14	15	16	17	18	19	20	21	22	23
20	21	22	23	0	1	2	3	4	5	6	7
11	12	13	14	15	16	17	18	19	20	21	22

Table 1, 2 and 3 jointly present the MCGND for $v = 24$, $k_1 = 5$ and $k_2 = 3$, using 72 blocks. In this design 12 unordered pairs (0,12), (1,13), ..., (11,23) among the total 276 do not appear as neighbours while all remaining 264 unordered pairs appear once.

3. Conditions for the Existence of MCGNDs in Which $v/2$ Pairs Do Not Appear as Neighbours

Condition 3.1: Let $m = (v-2)/2$ and $A = [1, 2, \dots, m]$. If $m \pmod{4} \equiv 0$ then A will provide MCGNDs for $v = 2ik_1 + 4k_2 + 2$ in which $(m+1)$ unordered pairs do not appear as neighbours while all other appear exactly once.

Condition 3.2: Let $m = (v-2)/2$ and $m \pmod{4} \equiv 3$. Then, $A = [1, 2, \dots, (3m-1)/4, (3m+7)/4, (3m+11)/4, \dots, m, 5(m+1)/4]$ will provide MCGNDs for $v = 2ik_1 + 4k_2 + 2$ in which $(m+1)$ unordered pairs do not appear while all others appear once.

4. MCGNDs in Blocks of Two Different Sizes

Here, some generators are developed through Rule I to obtain MCGNDs in blocks of two different sizes using i sets for k_1 and two for k_2 . In these designs $v/2$ pairs do not appear as neighbours while all other appear once. $(i+2)$ sets are obtained as follows:

1. Divide values of 'A' selected from Section 3, in i classes of size k_1 and two of size k_2 such that the sum of values in each class is divisible by v .
2. Deleting any one value from each class will result in $(i+2)$ sets of shifts to produce MCGND.

4.1. MCGNDs when $k_2 = 3$

Generator 4.1.1. Construct MCGNDs for k_2 for $v = 2ik_1 + 14$, $k_1 = 4l + 2$, $k_2 = 3$, i odd, $m \pmod{4} \equiv 0$ and l integer. Here:

- Consider $A = [1, 2, \dots, m]$.

Example 4.1.1. The following sets produce MCGND for $v = 26$, $k_1 = 6$ and $k_2 = 3$.

$$S_1 = [2, 3, 4, 5, 11], S_2 = [9, 10], S_3 = [8, 12]$$

Designs constructed through this method for $v \leq 100$, $k_1 = 6, 10, 14$ and 18 are presented in Table 4 in Appendix.

Generator 4.1.2. Construct MCGNDs for k_2 for $v = 2ik_1 + 14$, $k_1 \pmod{4} \equiv 1$, $k_2 = 3$, $I \pmod{4} \equiv 1$, $m \pmod{4} \equiv 3$. Here:

- Consider $A = [1, 2, \dots, (3m-1)/4, (3m+7)/4, (3m+11)/4, \dots, m, 5(m+1)/4]$.

Example 4.1.2. The following sets produce MCGND for $v = 24$, $k_1 = 5$ and $k_2 = 3$.

$$S_1 = [3, 4, 5, 11], S_2 = [8, 10], S_3 = [7, 15]$$

Designs constructed through this method for $v \leq 100$, $k_1 = 5, 9, 13$ and 17 are presented in Table 5 in Appendix.

4.2. MCGNDs when $k_2 = 4$

Generator 4.2.1. Construct MCGNDs for k_2 for $\nu = 2ik_1 + 18$, $k_1 = 4l + 2$, $k_2 = 4$, i even, $m \pmod{4} \equiv 0$. Here:

- Consider $A = [1, 2, \dots, m]$.

Example 4.2.1. The following sets produce MCGND for $\nu = 42$, $k_1 = 6$ and $k_2 = 4$.

$$S_1 = [3, 4, 5, 8, 20], S_2 = [11, 13, 16, 17, 18], S_3 = [7, 14, 15], S_4 = [10, 12, 19]$$

Designs constructed through this method for $\nu \leq 100$, $k_1 = 6, 10, 14$ and 18 are presented in Table 6 in Appendix.

Generator 4.2.2. Construct MCGNDs for k_2 for $\nu = 2ik_1 + 18$, $k_1 \pmod{4} \equiv 1$, $k_2 = 4$, $i \pmod{4} \equiv 3$, $m \pmod{4} \equiv 3$. Here:

- Consider $A = [1, 2, \dots, (3m-1)/4, (3m+7)/4, (3m+11)/4, \dots, m, 5(m+1)/4]$.

Example 4.2.2. The following sets produce MCGND for $\nu = 48$, $k_1 = 5$ and $k_2 = 4$.

$$S_1 = [4, 5, 17, 19], S_2 = [8, 9, 14, 16], S_3 = [10, 11, 12, 13], S_4 = [22, 23, 30], S_5 = [7, 15, 20]$$

Designs constructed through this method for $\nu \leq 100$, $k_1 = 5, 9$ and 13 are presented in Table 7 in Appendix.

4.3. MCGNDs when $k_2 = 5$

Generator 4.3.1. Construct MCGNDs for $\nu = 2ik_1 + 22$, $k_1 = 4l + 2$, $k_2 = 5$, i odd, $m \pmod{4} \equiv 0$ and l integer. Here:

- $A = [1, 2, \dots, m]$.

Example 4.3.1. The following sets produce MCGND for $\nu = 34$, $k_1 = 6$ and $k_2 = 5$.

$$S_1 = [3, 4, 5, 7, 13], S_2 = [6, 8, 9, 10], S_3 = [12, 14, 15, 16]$$

Designs constructed through this method for $\nu \leq 100$, $k_1 = 6, 10, 14$ and 18 are presented in Table 8 in Appendix.

Generator 4.3.2. Construct MCGNDs for $\nu = 2ik_1 + 22$, $k_1 \pmod{4} \equiv 1$, $k_2 = 5$, $i \pmod{4} \equiv 1$, $m \pmod{4} \equiv 3$. Here:

- $A = [1, 2, \dots, (3m-1)/4, (3m+7)/4, (3m+11)/4, \dots, m, 5(m+1)/4]$.

Example 4.3.2. The following sets produce MCGND for $\nu = 32$, $k_1 = 5$ and $k_2 = 5$.

$$S_1 = [2, 5, 9, 15], S_2 = [10, 13, 14, 20], S_3 = [4, 6, 8, 11]$$

Designs constructed through this method for $\nu \leq 100$, $k_1 = 5, 9, 13$ and 17 are presented in Table 9 in Appendix.

Acknowledgement

Authors are thankful to the Reviewer for valuable suggestions.

References

- Ahmed, R., Akhtar, M. and Tahir, M. H., (2009). Economical generalized neighbor designs of use in Serology. *Computational Statistics and Data Analysis*, 53, pp. 4584–4589.
- Ahmed, R., Akhtar, M., (2012). Designs partially balanced for neighbor effects. *Aligarh Journal of Statistics*, 32, pp. 41–53.
- Chaure, N. K., Misra, B. L., (1996). On construction of generalized neighbor design. *Sankhya Series B*, 58, pp. 45–253.
- Iqbal, I., (1991). Construction of experimental design using cyclic shifts, Ph.D. Thesis, University of Kent at Canterbury, U.K.
- Iqbal, I., Tahir, M. H., Aggarwal, M. L., Ali, A. and Ahmed, I. (2012). Generalized neighbor designs with block size 3. *Journal of Statistical Planning and Inference*, 142, pp. 626–632.
- Kedia, R. G., Misra, B. L., (2008). On construction of generalized neighbor design of use in serology. *Statistics and Probability Letters*, 18, pp. 254–256.
- Misra, B. L., Bhagwandas and Nutan S. M., (1991). Families of neighbor designs and their analysis, *Communications in Statistics - Simulation and Computation*, 20, pp. 427–436.
- Nutan, S. M., (2007). Families of Proper Generalized Neighbor Designs. *Journal of Statistical Planning and Inference*, 137, pp. 1681–1686.
- Rees, D. H., (1967). Some designs of use in serology. *Biometrics*, 23, pp. 779–791.
- Shehzad, F., Zafaryab, M. and Ahmed, R., (2011). Some series of proper generalized neighbor designs. *Journal of Statistical Planning and Inference*, 141, pp. 3808–3818.
- Zafaryab, M., Shehzad, F. and Ahmed, R., (2010). Proper generalized neighbor designs in circular blocks. *Journal of Statistical Planning and Inference*, 140, pp. 3498–3504.

Appendix

Table 4: MCGNDs for $\nu = 2ik_1 + 14$, $k_1 = 4l + 2$, $k_2 = 3$, i odd, l integer and $\nu \leq 100$

ν	k_1	k_2	Sets of Shifts
26	6	3	[1,2,3,4,5,11]+[7,9,10]+[6,8,12]
50	6	3	[1,3,4,6,12,24]+[2,7,8,9,11,13]+ [14,15,16,17,18,20]+[10,19,21]+[5,22,23]
74	6	3	[2,4,5,6,23,34]+[7,8,9,10,11,29]+ [1,12,13,15,16,17]+ [18,19,22,26,31,32]+ [14,24,25,27,28,30]+ [20,21,33]+[3,35,36]
98	6	3	[1,2,3,6,38,48]+[9,10,11,12,26,30]+ [13,14,15,16,18,22]+ [7,8,19,20,21,23]+ [25,27,28,29,43,44]+ [24,31,34,35,33,39]+ [4,32,37,40,41,42]+ [17,36,45]+[5,46,47]
34	10	3	[1,2,3,4,6,7,8,12,9,16]+[10,11,13]+[5,14,15]
74	10	3	[2,4,5,6,7,8,21,29,32,34]+[9,17,12,13,14,15,18,19,20,11]+ [1,16,22,23,24,25,26,27,28,30]+ [10,31,33]+[3,35,36]
42	14	3	[1,2,3,4,6,7,8,9,10,12,13,14,17,20]+[11,15,16]+[5,18,19]
98	14	3	[1,2,3,4,7,8,9,10,12,13,14,20,45,48]+ [6,15,16,17,18,19,21,22,23,24,27,35,26,25]+ [28,29,30,31,32,33,36,37,34,38,39,40,41,42]+[11,43,44]+[5,46,47]
50	18	3	[1,2,3,4,6,7,8,9,24,12,13,20,15,14,17,16,11,18]+[10,19,21]+[5,22,23]

Table 5: MCGNDs for $\nu = 2ik_1 + 14$, $k_1 \pmod{4} \equiv 1$, $k_2 = 3$, $i \pmod{4} \equiv 1$ and $\nu \leq 100$

ν	k_1	k_2	Sets of Shifts
24	5	3	[1,3,4,5,11]+[6,8,10]+[2,7,15]
64	5	3	[1,2,6,27,28]+[4,5,17,18,20]+ [10,12,13,14,15]+ [7,8,11,16,22]+ [19,21,23,25,40]+ [9,26,29]+[3,30,31]
32	9	3	[1,2,4,5,6,8,11,14,13]+[7,10,15]+[3,9,20]
40	13	3	[1,2,3,6,7,8,9,12,10,13,14,16,19]+[4,11,25]+[5,17,18]
48	17	3	[1,2,3,4,6,7,8,10,11,15,13,14,16,12,17,23,30]+[9,19,20]+[5,21,22]

Table 6: MCGNDs for $\nu = 2ik_1 + 18$, $k_1 = 4l + 2$, $k_2 = 4$, i even and l integer and $\nu \leq 100$

ν	k_1	k_2	Sets of Shifts
42	6	4	[2,3,4,5,8,20]+[9,11,13,16,17,18]+[6,7,14,15]+[1,10,12,19]
66	6	4	[1,3,4,6,20,32]+[7,8,9,11,12,19]+[16,17,18,26,27,28]+ [10,21,22,25,23,31]+ [13,14,15,24]+[2,5,29,30]
90	6	4	[5,18,34,40,41,42]+[7,8,9,14,15,37]+[6,11,12,16,17,28]+ [1,4,19,21,22,23]+ [25,26,27,29,30,43]+[13,31,32,33,35,36]+ [3,10,38,39]+[2,20,24,44]
58	10	4	[3,4,5,6,8,9,13,23,17,28]+[10,11,12,15,16,18,19,20,26,27]+ [1,14,21,22]+[2,7,24,25]
98	10	4	[1,2,3,4,6,7,8,9,10,48]+[13,14,15,16,17,18,19,20,23,41]+ [12,21,22,24,28,30,33,39,40,45]+ [32,34,35,36,37,38,43,44,46,47]+ [11,27,29,31]+[5,25,26,42]
74	14	4	[3,4,5,6,7,9,10,12,19,20,29,33,31,34]+ [2,14,15,16,17,21,24,27,22,23,25,28,26,36]+[1,11,30,32]+[8,13,18,35]
90	18	4	[1,2,3,4,5,7,8,11,12,13,14,15,16,18,19,39,40,43]+ [20,22,23,24,25,26,27,28,29,30,34,31,33,35,32,36,41,44]+[6,9,37,38]+ [10,17,21,42]

Table 7: MCGNDs for $\nu = 2ik_1 + 18$, $k_1 \pmod{4} \equiv 1$, $k_2 = 4$, $i \pmod{4} \equiv 3$ and $\nu \leq 100$

ν	k_1	k_2	Sets of Shifts
48	5	4	[3,4,5,17,19]+[1,8,9,14,16]+[2,10,11,12,13]+[21,22,23,30]+[6,7,15,20]
88	5	4	[2,4,5,34,43]+[7,8,9,22,42]+[3,12,13,28,32]+[14,17,18,19,20]+ [6,10,23,24,25]+[29,31,37,38,41]+[16,30,36,39,55]+[11,15,27,35]+ [1,21,26,40]
72	9	4	[4,6,7,8,10,20,26,28,35]+[9,12,13,14,15,16,17,18,30]+ [2,3,5,19,23,24,22,25,21]+[1,11,29,31]+[32,33,34,45]
96	13	4	[1,2,3,4,5,6,7,8,9,10,11,12,18]+[15,16,17,19,20,21,22,23,24,25,26,27,33]+[29,30,3 1,32,34,35,37,38,39,40,42,46,47]+[13,14,28,41]+[43,44,45,60]

Table 8: MCGNDs for $\nu = 2ik_1 + 22$, $k_1 = 4l + 2$, $k_2 = 5$, i odd, l integer, and $\nu \leq 100$

ν	k_1	k_2	Sets of Shifts
34	6	5	[2,3,4,5,7,13]+[1,6,8,9,10]+[11,12,14,15,16]
58	6	5	[1,2,3,4,22,26]+[7,8,9,10,11,13]+[16,17,19,20,21,23]+ [5,6,14,15,18]+[12,24,25,27,28]
82	6	5	[2,4,5,6,29,36]+[9,10,11,12,18,22]+[7,13,14,15,16,17]+[1,3,8,21,23,26]+ [19,24,25,28,31,37]+[30,32,33,34,35]+[20,27,38,39,40]
42	10	5	[2,3,4,6,7,8,9,10,15,20]+[1,5,11,12,13]+[14,16,17,18,19]
82	10	5	[1,2,3,4,5,6,7,8,9,37]+[11,13,14,15,16,17,18,19,20,21]+ [10,22,23,24,25,26,27,28,29,32]+[30,31,33,34,36]+[12,35,38,39,40]
50	14	5	[2,3,4,6,7,8,10,11,12,13,15,16,24,19]+[1,5,9,17,18]+[14,20,21,22,23]
58	18	5	[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,21,17,16]+[19,20,23,26,28]+ [18,22,24,25,27]

Table 9: MCGNDs for $\nu = 2ik_1 + 22$, $k_1 \pmod{4} \equiv 1$, $k_2 = 5$, $i \pmod{4} \equiv 1$ and $\nu \leq 100$

ν	k_1	k_2	Sets of Shifts
32	5	5	[1,2,5,9,15]+[7,10,13,14,20]+[3,4,6,8,11]
72	5	5	[3,5,9,10,45]+[4,8,17,21,22]+[2,13,15,19,23]+[6,12,16,18,20]+ [1,7,11,24,29]+[25,26,28,30,35]+[14,31,32,33,34]
40	9	5	[1,5,6,7,8,9,12,13,19]+[2,3,10,11,14]+[4,16,17,18,25]
48	13	5	[1,2,3,4,5,6,7,8,9,10,11,13,17]+[12,15,16,23,30]+[14,19,20,21,22]
56	17	5	[1,2,3,4,5,6,7,8,9,16,12,13,23,19,15,11,14]+[17,18,20,22,35]+ [10,24,25,26,27]

On representativeness, informative sampling, nonignorable nonresponse, semiparametric prediction and calibration

Abdulahakeem Eideh¹

Abstract

Informative sampling refers to a sampling design for which the sample selection probabilities depend on the values of the model outcome variable. In such cases the model holding for the sample data is different from the model holding for the population data. Similarly, nonignorable nonresponse refers to a nonresponse mechanism in which the response probability depends on the value of a missing outcome variable. For such a nonresponse mechanism the model holding for the response data is different from the model holding for the population data. In this paper, we study, within a modelling framework, the semi-parametric prediction of a finite population total by specifying the probability distribution of the response units under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. Furthermore, based on the relationship between response distribution and population distribution, we introduce a new measure of the representativeness of a response set and a new test of nonignorable nonresponse and informative sampling, jointly. Finally, a calibration estimator is obtained when the sampling design is informative and the nonresponse mechanism is nonignorable.


Key words: calibration, representative measure, response distribution, nonignorable nonresponse, informative sampling design.

1. Introduction

Informative sampling refers to sampling design for which the sample selection probabilities depend on the values of the model outcome variable (or the model outcome variable is correlated with design variables not included in working model). In such cases the model holding for the sample data (after sampling) is different from the model holding for the population data (before sampling); see Pfeiffermann et al. (1998). In the same way, nonignorable nonresponse refers to nonresponse mechanism in which the response probability depends on the value of a missing outcome variable;

¹ Department of Mathematics, Al-Quds University, Abu-Dees Campus, Al-Quds, Palestine.

E-mail: msabdul@staff.alquds.edu. ORCID: <https://orcid.org/0000-0002-9077-5795>.

© Abdulhakeem Eideh. Article available under the CC BY-SA 4.0 licence 

see Little (1982). For such nonresponse mechanism the model holding for the response data (after responding) is different from the model holding for the population data; Eideh (2007, 2012). From the literature review on survey sampling, it is clear that ignoring informative sampling or nonignorable nonresponse yield biased descriptive and analytics inferences about finite population parameters; see, for example, Chambers and Skinner (2003) and Eideh (2009). In recent articles, Eideh (2016, 2020) considers parametric prediction of finite population total under informative sampling design and nonignorable nonresponse. The author proved that, the failure to account informative sampling and nonignorable nonresponse in the analysis of survey data leads to biased inferences about the population of interest. In this paper, we study, within a modeling framework, the semi-parametric prediction of finite population total, by specifying the probability distribution of the observed measurements under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. Furthermore, based on the relationship between response distribution and population distribution, we introduced a new measure of representativeness of a response set, called generalized R-indicator, and a new test of nonignorable nonresponse and informative sampling, jointly.

The paper is structured as follows. Section 2 is devoted to notations. In Section 3 we review the definition of sample, sample-complement, response, and nonresponse distributions, and relationships between their mathematical expectations. Section 4 describes estimation of response probabilities under nonignorable nonresponse. In Section 5 a new test of nonignorable nonresponse and informative sampling was developed. In Section 6 we discuss different ways to generalize a measure of representativeness. Section 7 is devoted to the basic idea of prediction. In Section 8 we present the methodology of the semiparametric prediction of finite population total under informative sampling and nonignorable nonresponse. Finally, Section 9 provides the conclusions.

2. Notations

Let $U = \{1, \dots, N\}$ denote a finite population consisting of N units. Let y be the study or outcome variable of interest and let y_i be the value of y for the i -th population unit. A probability sample s is drawn from U according to a specified sampling design. The sample size is denoted by n . Let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, $i \in U$ be the values of a vector of auxiliary variables, x_1, \dots, x_p , and $\mathbf{z} = \{z_1, \dots, z_N\}$ be the values of known design variables, used for the sample selection process not included in the model under consideration. In what follows, we consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s) > 0$, and sampling weight $w_i = 1/\pi_i$; $i = 1, \dots, N$.

In practice, the π_i 's may depend on the population values $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. We express this dependence by writing: $\pi_i = \Pr(i \in s | \mathbf{x}, \mathbf{y}, \mathbf{z})$ for all units $i \in U$. Denote by $\mathbf{I} = (I_1, \dots, I_N)'$ the N by 1 sample indicator (vector) variable, such that $I_i = 1$ if unit $i \in U$ is selected to the sample and $I_i = 0$ if otherwise. Therefore, $s = \{i | i \in U, I_i = 1\}$ and its complement is $\bar{s} = c = \{i | i \in U, I_i = 0\}$. We consider the population values y_1, \dots, y_N as random variables, which are independent realizations from a distribution with probability density functions (pdf) $f_p(y_i | \mathbf{x}_i; \theta)$, indexed by a vector of parameters θ .

In addition to the effect of complex sample design, one of the major problems in the analysis of survey data is that of missing values. Denote by $\mathbf{R} = (R_1, \dots, R_N)'$ the N by 1 response indicator (vector) variable such that $R_i = 1$ if unit $i \in s$ is observed and $R_i = 0$ if otherwise. We assume that these random variables are independent of one another and of the sample selection mechanism. The response set is defined accordingly as $r = \{i \in s | R_i = 1\}$ and the nonresponse set by $\bar{r} = \{i \in s | R_i = 0\}$. We assume probability sampling so that $\pi_i = \Pr(i \in s) > 0$ for all units $i \in U$. Let $\psi_i = \Pr(i \in r | \mathbf{x}, \mathbf{y}, \mathbf{z}) > 0$ and $\varphi_i = 1/\psi_i$ be the response probability and response weights for all units $i \in s$. Let $O = \{(\mathbf{x}_i, I_i), i \in U\}, \{\pi_i, R_i, i \in s\} \cup \{(y_i, \mathbf{x}_i), i \in r\}$ and N, n , and m , be the available information from the sample and response sets. Furthermore, the following notations are frequently used in the paper: let $f_p, E_p(\cdot); f_s, E_s(\cdot); f_{\bar{s}}, E_{\bar{s}}(\cdot); f_r, E_r(\cdot)$; and $f_{\bar{r}}, E_{\bar{r}}(\cdot)$ denote the probability density functions and mathematical expectations of the population, sample, and sample-complement, response and nonresponse distributions, respectively.

3. Key equations

This section is based on Eideh (2020) and the references therein. The methodology in this paper is based on the following equations:

$$E_p(y_i | x_i) = E_r(\varphi_i w_i y_i | x_i) / E_r(\varphi_i w_i | x_i), \quad (1)$$

$$E_s(y_i | x_i) = \frac{E_r(\varphi_i y_i | x_i)}{E_r(\varphi_i | x_i)}, \quad (2)$$

$$E_{\bar{s}}(y_i | x_i) = E_r\{\varphi_i (w_i - 1) y_i | x_i\} / E_r\{\varphi_i (w_i - 1) | x_i\}, \quad (3)$$

$$E_{\bar{r}}(y_i | x_i) = E_r\{(\varphi_i - 1) y_i | x_i\} / E_r\{(\varphi_i - 1) | x_i\}, \quad (4)$$

$$f_r(y_i | x_i) = \frac{E_r(\varphi_i w_i | x_i)}{E_r(\varphi_i w_i | x_i, y_i)} f_p(y_i | x_i). \quad (5)$$

Consequently, $E_p(y_i|x_i)$, $E_s(y_i|x_i)$, $E_{\bar{s}}(y_i|x_i)$, $E_r(y_i|x_i)$ and $E_{\bar{r}}(y_i|x_i)$ can be estimated based on $\{x_i, y_i, \hat{\phi}_i, w_i; i \in r\}$. For estimation of φ_i , see Section 4.

4. Estimation of response probabilities under nonignorable nonresponse

Under nonignorable nonresponse, the values of y_i for $i \in r$ are available, but for $i \notin r$ are not available, so we cannot fit the following nonresponse model:

$$\psi_i = Pr(R_i = 1|i \in s, x_i, y_i) = \frac{\exp(\gamma_0 + \gamma_1 x_i + \gamma_2 y_i)}{1 + \exp(\gamma_0 + \gamma_1 x_i + \gamma_2 y_i)} \quad (6)$$

directly using the maximum likelihood method. A recent approach of estimation ψ_i under nonignorable nonresponse is discussed by Sverchkov (2008) using missing information principle. Assume $R_i: Bernoulli(\psi_i(x_i, y_i, \gamma))$, then

$$f(r_i|x_i, y_i) = (\psi_i(x_i, y_i, \gamma))^{r_i} (1 - \psi_i(x_i, y_i, \gamma))^{1-r_i}. \quad (7)$$

The maximum likelihood estimator of γ satisfies:

$$\frac{\partial l(\gamma)}{\partial \gamma} = \sum_{i \in r} \frac{\partial \log(\psi_i(x_i, y_i, \gamma))}{\partial \gamma} + \sum_{i \in \bar{r}} \frac{\partial \log(1 - \psi_i(x_i, y_i, \gamma))}{\partial \gamma} = 0. \quad (8)$$

Using (4), the observed log-likelihood equation is:

$$\begin{aligned} & \sum_{i \in r} \frac{\partial \log(\psi_i(x_i, y_i, \gamma))}{\partial \gamma} + \sum_{i \in \bar{r}} \frac{\left\{ E_r \left(\frac{[\varphi_i(x_i, y_i, \gamma) - 1] \partial \log(1 - \psi_i(x_i, y_i, \gamma))}{\partial \gamma} \right) \right\} | x_i}{E_r[(\varphi_i(x_i, y_i, \gamma) - 1) | x_i]} \\ &= \sum_{i \in r} \frac{\partial \log(\psi_i(x_i, y_i, \gamma))}{\partial \gamma} + \sum_{i \in \bar{r}} \frac{\int \frac{[\varphi_i(x_i, y_i, \gamma) - 1] \partial \log(1 - \psi_i(x_i, y_i, \gamma))}{\partial \gamma} f_r(y_i | x_i) dy_i}{\int (\varphi_i(x_i, y_i, \gamma) - 1) f_r(y_i | x_i) dy_i} = 0. \end{aligned} \quad (9)$$

Hence, $\hat{\psi}_i = \psi_i(\hat{\gamma}) = \psi_i(x_i, y_i, \hat{\gamma}) = Pr(i \in r | x_i, y_i, \hat{\gamma})$ and then $\hat{\varphi}_i = 1/\hat{\psi}_i$. From now on, to simplify notation, we will use ψ_i to denote ψ_i or $\hat{\psi}_i$.

5. New test of nonignorable nonresponse and informative sampling, jointly

According to (5), the response distribution $f_r(y_i|x_i)$ of $y_i, i \in r$, is different from the population distribution, $f_p(y_i|x_i)$, unless $E_r(\varphi_i w_i | x_i, y_i) = E_r(\varphi_i w_i | x_i)$ for all units $i \in U$, that is when the sampling design is noninformative and nonresponse

mechanism is ignorable. In such cases, the model holding for the response data (after sampling) is the same as the model holding for the population data (before sampling). The main target of inference is estimation $E_p(y_i|x_i)$. According to Eideh (2020), we have:

$$E_r(y_i|x_i) = E_p \left\{ \frac{E_s(\psi_i|x_i, y_i, \gamma)}{E_s(\psi_i|x_i, \theta, \eta, \gamma)} \frac{E_p(\pi_i|x_i, y_i, \gamma)}{E_p(\pi_i|x_i, \theta, \gamma)} y_i \middle| x_i \right\} \neq E_p(y_i|x_i). \quad (10)$$

This relationship illustrates that the failure to account nonignorable nonresponse and informative sampling design can bias the inference. So, testing the ignorability of nonresponse and the informativeness of sampling design is necessary, which is the aim of this section.

Recall that if $E_r(\varphi_i w_i|x_i, y_i) = E_r(\varphi_i w_i|x_i)$, for all units $i \in U$, then $f_r(y_i|x_i) = f_p(y_i|x_i)$. Consequently, in the spirit of equation (5), we introduce the following new test of ignorable nonresponse and informative sampling, jointly, by testing

$$H_0: E_r(\varphi_i w_i|x_i, y_i) = E_r(\varphi_i w_i|x_i) \text{ versus } H_1: E_r(\varphi_i w_i|x_i, y_i) \neq E_r(\varphi_i w_i|x_i) \quad (11)$$

at α level of significance.

In addition to that, testing of noninformative sampling design and nonresponse mechanism is missing completely at random, and can be approached by testing:

$$H_0: E_r(\varphi_i w_i|x_i, y_i) = \text{constant} \text{ versus } H_1: E_r(\varphi_i w_i|x_i, y_i) \neq \text{constant} \quad (12)$$

at α level of significance.

Particular cases:

(a) If the sampling design is noninformative, that is, the sample selection process can be ignored, then the test of nonignorable nonresponse is determined by testing:

$$H_0: E_r(\varphi_i|x_i, y_i) = E_r(\varphi_i|x_i) \text{ versus } H_1: E_r(\varphi_i|x_i, y_i) \neq E_r(\varphi_i|x_i) \quad (13)$$

at α level of significance.

(b) If nonresponse mechanism is ignorable, then the test of informativeness can be conducted by testing:

$$H_0: E_r(w_i|x_i, y_i) = E_r(w_i|x_i) \text{ versus } H_1: E_r(w_i|x_i, y_i) \neq E_r(w_i|x_i) \quad (14)$$

at α level of significance.

The above hypotheses can be tested by using the general regression test approach, by specifying the full model. For example, in (11) and (12), assume the full model is given by:

$$E_r(\varphi_i w_i | x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 y_i. \quad (15)$$

Then, (11) becomes $H_0: \beta_2 = 0$, and (12) says $H_0: \beta_1 = \beta_2 = 0$.

6. Generalized Measure of Representativeness

Schouten et al. (2009) proposed an indicator which we call an R-indicator ('R' for representativeness), for the similarity between the response to a survey and the sample or the population under investigation. This similarity can be referred to as "representative response". The R-indicator that they proposed employs estimated response probabilities.

Definition 1 (strong): A response subset is representative with respect to the sample if the response propensities ρ_i are the same for all units in the population. That is,

$$\rho_i = \Pr(R_i = 1 | I_i = 1) = \rho \text{ for all units } i \in U \quad (16)$$

and if the response of a unit is independent of the response of all other units.

Under the assumption that the individual response propensities ρ_i are known, Schouten et al. (2009) defined the R-indicator as:

$$R(\rho) = 1 - 2S(\rho), \quad (17)$$

where

$$S(\rho) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (\rho_i - \bar{\rho})^2}, \bar{\rho} = \frac{1}{N} \sum_{i=1}^N \rho_i. \quad (18)$$

One may view R as a lack of the association measure. When $R(\rho) = 1$ there is no relation between any survey item and the missing-data mechanism. The R-indicator takes values on the interval $[0, 1]$ with the value 1 being strong representativeness and the value 0 being the maximum deviation from strong representativeness.

Schouten et al. (2009) pointed that "in practice these propensities are unknown. Furthermore, in a survey, we only have information about the response behaviour of sample units. We, therefore, have to find alternatives to the indicators R. Let $\hat{\rho}_i$ denote an estimator for ρ_i which uses all or a subset of the available auxiliary variables.

Methods that support such estimation are, for instance, logistic or probit regression models". The authors replace R by the estimators \hat{R} :

$$\hat{R}(\rho) = 1 - 2\hat{S}(\rho), \hat{S}(\rho) = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \frac{I_i}{\pi_i} (\hat{\rho}_i - \hat{\rho})^2}, \hat{\rho} = \frac{1}{N} \sum_{i=1}^N \frac{I_i}{\pi_i} \rho_i. \quad (19)$$

The R -indicator introduced by Schouten et al. (2009) assumed that the sampling design is noninformative and nonresponse mechanism is ignorable. In this section we develop a new indicator of representative response of the survey and the population when the sampling design is informative and the nonresponse mechanism nonignorable. It should be pointed here that, under nonignorable nonresponse, we cannot compute the propensity scores for all units in the sample, see Section 4, consequently, the formulas defined in equation (19) are not applicable in such cases.

For simplicity, assume no auxiliary variables are available. In the essence of equation (5), let us consider the following four cases.

Case 1: Sampling design is informative and nonresponse mechanism is nonignorable (in), then

$$f_r(y_i) = \frac{E_r(\varphi_i w_i)}{E_r(\varphi_i w_i | y_i)} f_p(y_i). \quad (20)$$

In this case, the response distribution represents the population distribution if $E_r(\phi_i w_i) = E_r(\varphi_i w_i | y_i)$. That is, $E_r(\varphi_i w_i | y_i) = \text{constant}$. In this case, we introduce the following definition.

Definition 2: A response set is representative with respect to the population if the product of the response weights and sampling weights $\varphi_i w_i = d_i$ are the same for all units in the population and if the response of a unit is independent of the response of all other units. That is, $\varphi_i w_i = d_i = d$ for all units $i \in U$.

Thus, we define a generalized R -indicator as follows:

$$\hat{R}(d) = 1 - 2\hat{S}(d), \quad (21)$$

where

$$\hat{S}(d) = \sqrt{\frac{1}{\sum_{i=1}^m \hat{d}_i} \sum_{i=1}^m \hat{d}_i (\hat{d}_i - \hat{\hat{d}})^2} \quad (22)$$

and

$$\hat{\hat{d}} = \frac{1}{\sum_{i=1}^m \hat{d}_i} \sum_{i=1}^m \hat{d}_i (\hat{d}_i), \quad \hat{\varphi}_i w_i = \hat{d}_i. \quad (23)$$

Case 2: Sampling design is noninformative and nonresponse mechanism is nonignorable (nn), then

$$f_r(y_i) = \frac{E_r(\varphi_i)}{E_r(\varphi_i|y_i)} f_p(y_i). \quad (24)$$

In this case, the response distribution represents the population distribution if $E_r(\phi_i) = E_r(\varphi_i|y_i)$. That is, $E_r(\varphi_i|y_i) = \text{constant}$. In this constant, we introduce the following definition.

Definition 3: A response set is representative with respect to the population if the response weights φ_i are the same for all units in the population and if the response of a unit is independent of the response of all other units. That is, $\varphi_i = \varphi$ for all units $i \in U$.

Thus, we define a generalized R-indicator as follows:

$$\hat{R}_{nn}(\varphi) = 1 - 2\hat{S}_{nn}(\varphi), \quad (25)$$

where

$$\hat{S}_{nn}(\varphi) = \sqrt{\frac{1}{\sum_{i=1}^m \hat{\varphi}_i} \sum_{i=1}^m \hat{\varphi}_i (\hat{\varphi}_i - \hat{\varphi}_{nn})^2} \quad (26)$$

and

$$\hat{\varphi}_{nn} = \frac{1}{\sum_{i=1}^m \hat{\varphi}_i} \sum_{i=1}^m \hat{\varphi}_i^2. \quad (27)$$

Case 3: Sampling design is informative and nonresponse mechanism is ignorable (ii), then

$$f_r(y_i) = \frac{E_r(w_i)}{E_r(w_i|y_i)} f_p(y_i). \quad (28)$$

In this case, the response distribution represents the population distribution if $E_r(w_i) = E_r(w_i|y_i)$. That is, $E_r(w_i|y_i) = \text{constant}$. In this case, we introduce the following definition.

Definition 4: A response set is representative with respect to the population if the sampling weighs w_i are the same for all units in the population and if the response of a unit is independent of the response of all other units. That is, $w_i = w$ for all units $i \in U$.

Thus, we define a generalized R-indicator as follows:

$$\hat{R}_{ii}(w) = 1 - 2\hat{S}_{ii}(w), \quad (29)$$

where

$$\hat{S}_{ii}(w) = \sqrt{\frac{1}{\sum_{i=1}^m w_i} \sum_{i=1}^m w_i (w_i - \hat{w})^2} \quad (30)$$

and

$$\hat{w} = \frac{1}{\sum_{i=1}^m w_i} \sum_{i=1}^m w_i^2. \quad (31)$$

Case 4: Sampling design is noninformative and nonresponse mechanism is ignorable (ii), then

$$f_r(y_i) = f_p(y_i). \quad (32)$$

In this case, the response distribution represents the population distribution, and no need a measure of a representative subset.

Research in this highly interested generalized R-indicator is in progress.

7. Prediction of Finite Population Total

This section is devoted to the basics of the prediction of finite population total, taking into account informative sampling design and nonignorable nonresponse mechanism. Assume single-stage population model. Let

$$T = \sum_{i=1}^N y_i = \sum_{i \in S} y_i + \sum_{i \in \bar{S}} y_i = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} y_i + \sum_{i \in \bar{S}} y_i \quad (33)$$

be the finite population total that we want to predict using the data from the response set and possibly values of auxiliary variables. Let $\hat{T} = \hat{T}(O)$ define the predictor of T based on the available information, from the sample and response set $O = \{(I_i), i \in U\}, \{\pi_i, R_i, i \in S\} \cup \{(y_i), i \in r\}$ and N, n , and m . The mean square error (MSE) of \hat{T} given O with respect to the population pdf is defined by:

$$MSE_p(\hat{T}) = E_p \left\{ (\hat{T} - T)^2 | O \right\} = \{ \hat{T} - E_p(T|O) \}^2 + Var_p(T|O). \quad (34)$$

It obvious that (16) is minimized when $\hat{T} = E(T|O)$. Hence, the minimum mean squared error best linear unbiased predictor (BLUP) of $T = \sum_{i=1}^N y_i$ is given by:

$$T^* = E_p(T|O) = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} E_{\bar{r}}(y_i|O) + \sum_{i \in \bar{S}} E_{\bar{S}}(y_i|O). \quad (35)$$

For more information about the parametric prediction of finite population total under informative sampling and nonignorable nonresponse; see Eideh (2020).

In the next section we consider the semiparametric prediction of finite population total under informative sampling.

8. Semiparametric Prediction of Finite Population Total under Informative Sampling and Nonignorable Nonresponse

Sverchkov and Pfeiffermann (2004) studied the semiparametric prediction of finite population totals under informative sampling. In this section we develop the semiparametric prediction of finite population total under informative sampling and nonignorable nonresponse. According to (35), the prediction of finite population total requires predication of $\sum_{i \in \bar{r}} y_i$, and $\sum_{i \in \bar{s}} y_i$. That is, to predict T we need to predict values for $\{y_i, i \in \bar{r}\}$ and $\{y_i, i \in \bar{s}\}$, based on the prediction of nonsampled and nonresponse models. Fuller (2009, p. 282) pointed that “The analysis of data with unplanned nonresponse requires the specification of a model for the nonresponse. Models for nonresponse address two characteristics: the probability of obtaining a response and the distribution of the characteristic. In one model it is assumed that the probability of response can be expressed as a function of auxiliary data. The assumption of a second important model is that the expected value of the unobserved variable is related to observable auxiliary data. In some situations models constructed under the two models lead to the same estimator. Similarly, specifications containing models for both components can be developed.”

8.1. General Theory

Assume that

(a) the sample-complement model (or nonsampled model or imputation model for non-sampled units) takes the form:

$$y_i = S_{\beta}(\mathbf{x}_i) + \varepsilon_i, \text{ for all } i \in \bar{s}, \quad (36)$$

$$E_{\bar{s}}(\varepsilon_i | \mathbf{x}_i) = 0, \quad E_{\bar{s}}(\varepsilon_i^2 | \mathbf{x}_i) = \text{Var}_{\bar{s}}(\varepsilon_i | \mathbf{x}_i) = \sigma_{\varepsilon}^2 v(\mathbf{x}_i), \quad \text{and} \quad E_{\bar{s}}(\varepsilon_j \varepsilon_k | \mathbf{x}_i) = \text{Cov}(\varepsilon_j, \varepsilon_k | \mathbf{x}_i) = 0, j \neq k.$$

(b) and the response-complement model (or nonresponse model or missing data model or imputation model for nonrespondent units) is:

$$y_i = Z_{\alpha}(\mathbf{x}_i) + \tau_i, \text{ for all } i \in \bar{r}, \quad (37)$$

$E_{\bar{r}}(\tau_i | \mathbf{x}_i) = 0$, $E_{\bar{r}}(\tau_i^2 | \mathbf{x}_i) = \text{Var}_{\bar{r}}(\tau_i | \mathbf{x}_i) = \sigma_{\tau}^2 u(\mathbf{x}_i)$, and $E_{\bar{r}}(\tau_j \tau_k | \mathbf{x}_i) = \text{Cov}(\tau_j, \tau_k | \mathbf{x}_i) = 0$, $j \neq k$, where $S_{\beta}(\mathbf{x}_i)$ and $Z_{\alpha}(\mathbf{x}_i)$ are known functions of \mathbf{x}_i that depend on unknown vector parameters β and α , respectively. The variances $\sigma_{\varepsilon}^2 v(\mathbf{x}_i)$ and $\sigma_{\tau}^2 u(\mathbf{x}_i)$ are assumed known except for σ_{ε}^2 and σ_{τ}^2 .

For the prediction process, we need the estimation of $S_{\beta}(\mathbf{x}_i)$ and $Z_{\alpha}(\mathbf{x}_i)$.

(i) Estimation of $S_{\beta}(x_i)$:

Method 1: using (3), we have

$$S_{\beta}(x_i) = \arg \min_{S_{\beta}(x_i)} E_s \left\{ \frac{(y_i - S_{\beta}(x_i))^2}{v(x_i)} \middle| x_i \right\} = \arg \min_{S_{\beta}(x_i)} E_r \left\{ c_i \frac{(y_i - S_{\beta}(x_i))^2}{v(x_i)} \middle| x_i \right\}, \quad (38)$$

where $c_i = \{\varphi_i(w_i - 1)/E_r(\varphi_i(w_i - 1)|x_i)\}$.

Hence, the vector β can be estimated by:

$$\hat{\beta}_1 = \arg \min_{\beta} \sum_{i \in r} \left(\hat{c}_i \frac{(y_i - S_{\beta}(x_i))^2}{v(x_i)} \right), \quad (39)$$

where $\hat{c}_i = \{\hat{\varphi}_i(w_i - 1)/\hat{E}_r(\hat{\varphi}_i(w_i - 1)|x_i)\}$.

Method 2: using (3) and (36), and assume that $E_r(\varphi_i(w_i - 1)|x_i) \approx E_r(\varphi_i(w_i - 1))$, we have:

$$E_s \left\{ \frac{(y_i - S_{\beta}(x_i))^2}{v(x_i)} \middle| x_i \right\} = E_r \left\{ \frac{\varphi_i(w_i - 1)}{E_r(\varphi_i(w_i - 1))} \frac{(y_i - S_{\beta}(x_i))^2}{v(x_i)} \right\}. \quad (40)$$

Hence,

$$\hat{\beta}_2 = \arg \min_{\beta} \sum_{i \in r} \left(\hat{\varphi}_i(w_i - 1) \frac{(y_i - S_{\beta}(x_i))^2}{v(x_i)} \right), \quad (41)$$

since $E_r(\varphi_i(w_i - 1))$ is constant.

(ii) Estimation of $Z_{\alpha}(x_i)$

Method 1: using (4), we have:

$$Z_{\alpha}(x_i) = \arg \min_{Z_{\alpha}(x_i)} E_r \left\{ k_i \frac{(y_i - Z_{\alpha}(x_i))^2}{u(x_i)} \middle| x_i \right\}, \quad (42)$$

where $k_i = \{(\varphi_i - 1)/E_r((\varphi_i - 1)|x_i)\}$.

Hence, the vector $\hat{\alpha}$ can be estimated by

$$\hat{\alpha}_1 = \arg \min_{\beta} \sum_{i \in r} \left(\hat{k}_i \frac{(y_i - Z_{\alpha}(x_i))^2}{u(x_i)} \right), \quad (43)$$

where $\hat{k}_i = \{(\hat{\varphi}_i - 1)/\hat{E}_r((\hat{\varphi}_i - 1)|x_i)\}$.

Method 2: using (3) and (37), and assume that $E_r((\varphi_i - 1)|\mathbf{x}_i) \approx E_r(\varphi_i - 1)$, we have:

$$E_{\bar{r}} \left\{ \frac{(y_i - Z_{\alpha}(\mathbf{x}_i))^2}{u(\mathbf{x}_i)} \middle| \mathbf{x}_i \right\} = E_r \left\{ \frac{(\varphi_i - 1)}{E_r(\varphi_i - 1)} \frac{(y_i - Z_{\alpha}(\mathbf{x}_i))^2}{u(\mathbf{x}_i)} \right\}. \quad (44)$$

Thus,

$$\hat{\alpha}_2 = \arg \min_{\hat{\alpha}} \sum_{i \in r} \left((\hat{\varphi}_i - 1) \frac{(y_i - Z_{\hat{\alpha}}(\mathbf{x}_i))^2}{u(\mathbf{x}_i)} \right). \quad (45)$$

Hence,

$$\hat{T}_{in,1} = \hat{E}_p(T|O) = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} Z_{\hat{\alpha}_1}(\mathbf{x}_i) + \sum_{i \in \bar{s}} S_{\hat{\beta}_1}(\mathbf{x}_i) \quad (46)$$

and

$$\hat{T}_{in,2} = \hat{E}_p(T|O) = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} Z_{\hat{\alpha}_2}(\mathbf{x}_i) + \sum_{i \in \bar{s}} S_{\hat{\beta}_2}(\mathbf{x}_i). \quad (47)$$

The benefits of using the predictor $\hat{T}_{in,2}$ over using the predictor $\hat{T}_{in,1}$ is that $\hat{T}_{in,2}$ does not require the identification and estimation of $\varphi(\mathbf{x}_i) = E_r((\varphi_i - 1)|\mathbf{x}_i)$. On the other hand, in situations where this expectation can be estimated properly, the predictor $\hat{T}_{in,1}$ is likely to be more accurate since the weights $k_i = \{(\varphi_i - 1)/E_r((\varphi_i - 1)|\mathbf{x}_i)\}$ will often be less variable than the weights $(\varphi_i - 1)$. This is because the weights $k_i = \{(\varphi_i - 1)/E_r((\varphi_i - 1)|\mathbf{x}_i)\}$ only account for the net effect of the response process on the target conditional distribution $f_{\bar{r}}(y_i|\mathbf{x}_i, \theta, \eta, \gamma)$ whereas the weights $(\varphi_i - 1)$ account for the effect of the response process on the joint distribution $f_{\bar{r}}(y_i, \mathbf{x}_i; \theta, \eta, \gamma)$.

8.1.1. Particular cases

For illustration we use method 2 only under different famous models in survey sampling.

Case 1: Common Mean Model:

Sample-complement model: $E_{\bar{s}}(y_i) = \mu_{\bar{s}}$ and $Var_{\bar{s}}(y_i) = \sigma_{\bar{s}}^2$.

Response-complement model: $E_{\bar{r}}(y_i|\mathbf{x}_i) = \mu_{\bar{r}}$ and $Var_{\bar{r}}(y_i) = \sigma_{\bar{r}}^2$.

After some algebra, the weights under the 4 different combinations of sampling design (informative (i), noninformative (n)) and nonresponse mechanism (ignorable (i), nonignorable (n)) are summarized in Table 1.

Table 1: w_{ir} - Homogeneous Model, $\hat{T}_{in,2} = \sum_{i \in r} w_{ir} y_i$

SD-NM	w_{ir}
ii	$1 + \frac{(n-m)}{m} + (N-n) \frac{w_i - 1}{\sum_{i \in r} (w_i - 1)}$
in	$1 + (n-m) \frac{(\hat{\phi}_i - 1)}{\sum_{i \in r} (\hat{\phi}_i - 1)} + (N-n) \frac{\hat{\phi}_i (w_i - 1)}{\sum_{i \in r} \hat{\phi}_i (w_i - 1)}$
ni	$\frac{N}{m}$
nn	$1 + (n-m) \frac{(\hat{\phi}_i - 1)}{\sum_{i \in r} (\hat{\phi}_i - 1)} + (N-n) \frac{\hat{\phi}_i}{\sum_{i \in r} \hat{\phi}_i}$

Note that $\sum_{i \in r} w_{ir} = \sum_{i \in U} 1 = N$.

Case 2: Simple linear regression model:

Sample-complement model: $E_{\bar{s}}(y_i | x_i) = \beta_0 + \beta_1 x_i$ and $Var_{\bar{s}}(y_i) = \sigma_{\bar{\epsilon}}^2$.

Response-complement model: $E_{\bar{r}}(y_i | x_i) = \alpha_0 + \alpha_1 x_i$ and $Var_{\bar{r}}(y_i) = \sigma_{\bar{\epsilon}}^2$.

After some algebra, the weights under the 4 different combinations of sampling designs (informative (i), noninformative (n)) and nonresponse mechanism (ignorable (i), nonignorable (n)) are summarized in Table 2.

Table 2: w_{ir} - Simple linear regression model, $\hat{T}_{in,2} = \sum_{i \in r} w_{ir} y_i$

SD-NM	w_{ir}	\bar{x}_{φ^*}	\bar{x}_{w^*}
ii	$1 + \frac{(n-m)}{m} + (n-m) (\bar{x}_{\bar{r}} - \bar{x}_{\varphi^*}) \frac{(x_i - \bar{x}_{\varphi^*})}{\sum_{i \in r} (x_i - \bar{x}_{\varphi^*})^2}$ $(N-n) \frac{(w_i - 1)}{\sum_{i \in r} i (w_i - 1)} +$ $(N-n) (\bar{x}_{\bar{s}} - \bar{x}_{w^*}) \frac{(w_i - 1) (x_i - \bar{x}_{w^*})}{\sum_{i \in r} (w_i - 1) (x_i - \bar{x}_{w^*})^2}$	$\frac{\sum_{i \in r} x_i}{m}$	$\frac{\sum_{i \in r} (w_i - 1) x_i}{\sum_{i \in r} (w_i - 1)}$
in	$1 + (n-m) \frac{(\hat{\phi}_i - 1)}{\sum_{i \in r} (\hat{\phi}_i - 1)} +$ $(n-m) (\bar{x}_{\bar{r}} - \bar{x}_{\varphi^*}) \frac{(\hat{\phi}_i - 1) (x_i - \bar{x}_{\varphi^*})}{\sum_{i \in r} (\hat{\phi}_i - 1) (x_i - \bar{x}_{\varphi^*})^2}$ $(N-n) \frac{\hat{\phi}_i (w_i - 1)}{\sum_{i \in r} \hat{\phi}_i (w_i - 1)} +$ $(N-n) (\bar{x}_{\bar{s}} - \bar{x}_{w^*}) \frac{\hat{\phi}_i (w_i - 1) (x_i - \bar{x}_{w^*})}{\sum_{i \in r} \hat{\phi}_i (w_i - 1) (x_i - \bar{x}_{w^*})^2}$	$\frac{\sum_{i \in r} (\hat{\phi}_i - 1) x_i}{\sum_{i \in r} (\hat{\phi}_i - 1)}$	$\frac{\sum_{i \in r} \hat{\phi}_i (w_i - 1) x_i}{\sum_{i \in r} \hat{\phi}_i (w_i - 1)}$
ni	$1 + \frac{(n-m)}{m} + (n-m) (\bar{x}_{\bar{r}} - \bar{x}_{\varphi^*}) \frac{(x_i - \bar{x}_{\varphi^*})}{\sum_{i \in r} (x_i - \bar{x}_{\varphi^*})^2}$ $\frac{(N-n)}{m} + (N-n) (\bar{x}_{\bar{s}} - \bar{x}_{w^*}) \frac{(x_i - \bar{x}_{w^*})}{\sum_{i \in r} (x_i - \bar{x}_{w^*})^2}$	$\frac{\sum_{i \in r} x_i}{m}$	$\frac{\sum_{i \in r} x_i}{m}$

Table 2: w_{ir} - Simple linear regression model, $\hat{T}_{in,2} = \sum_{i \in r} w_{ir} y_i$ (cont.)

SD-NM	w_{ir}	\bar{x}_{φ^*}	\bar{x}_{w^*}
nn	$1 + (n-m) \frac{(\hat{\varphi}_i - 1)}{\sum_{i \in r} (\hat{\varphi}_i - 1)} +$ $(n-m)(\bar{x}_{\bar{r}} - \bar{x}_{\varphi^*}) \frac{(\hat{\varphi}_i - 1)(x_i - \bar{x}_{\varphi^*})}{\sum_{i \in r} (\hat{\varphi}_i - 1)(x_i - \bar{x}_{\varphi^*})^2}$ $(N-n) \frac{\hat{\varphi}_i}{\sum_{i \in r} \hat{\varphi}_i} + (N-n)(\bar{x}_{\bar{s}} - \bar{x}_{w^*}) \frac{\hat{\varphi}_i(x_i - \bar{x}_{w^*})}{\sum_{i \in r} \hat{\varphi}_i(x_i - \bar{x}_{w^*})^2}$	$\frac{\sum_{i \in r} (\hat{\varphi}_i - 1)x_i}{\sum_{i \in r} (\hat{\varphi}_i - 1)}$	$\frac{\sum_{i \in r} \hat{\varphi}_i x_i}{\sum_{i \in r} \hat{\varphi}_i}$

Note that $\sum_{i \in r} w_{ir} x_i = \sum_{i \in U} x_i$.

Case 3: Simple ratio model:

Sample-complement model: $E_{\bar{s}}(y_i | x_i) = \beta x_i$ and $Var_{\bar{s}}(y_i) = \sigma_{\varepsilon}^2 x_i$.

Response-complement model: $E_{\bar{r}}(y_i | x_i) = \alpha x_i$; and $Var_{\bar{r}}(y_i) = \sigma_{\tau}^2 x_i$.

After some algebra, the weights under the 4 different combinations of sampling designs (informative (i), noninformative (n)) and nonresponse mechanism (ignorable (i), nonignorable (n)) are summarized in Table 3.

Table 3: w_{ir} - Simple ratio (or proportional) model, $\hat{T}_{in,2} = \sum_{i \in r} w_{ir} y_i$

SD-NM	w_{ir}	\bar{x}_{φ^*}	\bar{x}_{w^*}
ii	$1 + \frac{(n-m)}{m} \frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}} +$ $(N-n) \frac{(w_i - 1)}{\sum_{i \in r} (w_i - 1)} \frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$	$\frac{\sum_{i \in r} x_i}{m}$	$\frac{\sum_{i \in r} (w_i - 1)x_i}{\sum_{i \in r} (w_i - 1)}$
in	$1 + (n-m) \frac{(\hat{\varphi}_i - 1)}{\sum_{i \in r} (\hat{\varphi}_i - 1)} \frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}} +$ $(N-n) \frac{\hat{\varphi}_i (w_i - 1)}{\sum_{i \in r} \hat{\varphi}_i (w_i - 1)} \frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$	$\frac{\sum_{i \in r} (\hat{\varphi}_i - 1)x_i}{\sum_{i \in r} (\hat{\varphi}_i - 1)}$	$\bar{x}_{w^*} = \frac{\sum_{i \in r} \hat{\varphi}_i (w_i - 1)x_i}{\sum_{i \in r} \hat{\varphi}_i (w_i - 1)}$
ni	$1 + \frac{(n-m)}{m} \frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}} + \frac{(N-n)}{m} \frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$	$\frac{\sum_{i \in r} x_i}{m}$	$\frac{\sum_{i \in r} x_i}{m}$
nn	$1 + (n-m) \frac{(\hat{\varphi}_i - 1)}{\sum_{i \in r} (\hat{\varphi}_i - 1)} \frac{\bar{x}_{\bar{r}}}{\bar{x}_{\varphi^*}} +$ $(N-n) \frac{\hat{\varphi}_i}{\sum_{i \in r} \hat{\varphi}_i} \frac{\bar{x}_{\bar{s}}}{\bar{x}_{w^*}}$	$\frac{\sum_{i \in r} (\hat{\varphi}_i - 1)x_i}{\sum_{i \in r} (\hat{\varphi}_i - 1)}$	$\frac{\sum_{i \in r} \hat{\varphi}_i x_i}{\sum_{i \in r} \hat{\varphi}_i}$

Note that $\sum_{i \in r} w_{ir} x_i = \sum_{i \in U} x_i$.

8.2. Bias correction method- multiple regression

According to (35), with auxiliary variables, the prediction of finite population total requires computation of $\sum_{i \in \bar{s}} E_s(y_i | \mathbf{x}_i)$ and $\sum_{i \in \bar{r}} E_r(y_i | \mathbf{x}_i)$. Here, we use the “bias correction method” proposed by (Chambers 2003), see Chambers and Clark (2012, page 114).

Computation of $\sum_{i \in \bar{s}} E_s(y_i | \mathbf{x}_i)$:

$$\begin{aligned} \sum_{i \in \bar{s}} E_s(y_i | \mathbf{x}_i) &= \sum_{i \in \bar{s}} \{E_{\bar{s}}(y_i | \mathbf{x}_i) + E_s(y_i | \mathbf{x}_i) - E_s(y_i | \mathbf{x}_i)\} \\ &= \sum_{i \in \bar{s}} E_s(y_i | \mathbf{x}_i) + \sum_{i \in \bar{s}} \{E_{\bar{s}}(y_i | \mathbf{x}_i) - E_s(y_i | \mathbf{x}_i)\} \\ &\cong \sum_{i \in \bar{s}} E_s(y_i | \mathbf{x}_i) + N - n \frac{1}{N-n} \sum_{i \in \bar{s}} E_{\bar{s}}(y_i - E_s(y_i | \mathbf{x}_i)). \end{aligned} \quad (48)$$

Now, using (3), $E_{\bar{s}}(y_i - E_s(y_i | \mathbf{x}_i))$ can be estimated by

$$\hat{E}_{\bar{s}}(y_i - \hat{E}_s(y_i | \mathbf{x}_i)) = \frac{1}{\sum_{i \in \bar{r}} \hat{\varphi}_i(w_i - 1)} \sum_{i \in \bar{r}} \hat{\varphi}_i(w_i - 1) \left(y_i - \hat{E}_r \left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i | \mathbf{x}_i)} y_i \right) \right). \quad (49)$$

Also, using (2), $E_s(y_i | \mathbf{x}_i)$ can be estimated by

$$\hat{E}_s(y_i | \mathbf{x}_i) = \hat{E}_r \left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i | \mathbf{x}_i)} y_i \right). \quad (50)$$

Computation of $\sum_{i \in \bar{r}} E_r(y_i | \mathbf{x}_i)$:

Similarly,

$$\begin{aligned} \sum_{i \in \bar{r}} E_r(y_i | \mathbf{x}_i) &= \sum_{i \in \bar{r}} \{E_{\bar{r}}(y_i | \mathbf{x}_i) + E_r(y_i | \mathbf{x}_i) - E_r(y_i | \mathbf{x}_i)\} \\ &= \sum_{i \in \bar{r}} E_r(y_i | \mathbf{x}_i) + \sum_{i \in \bar{r}} \{E_{\bar{r}}(y_i | \mathbf{x}_i) - E_r(y_i | \mathbf{x}_i)\} \\ &\cong \sum_{i \in \bar{r}} E_r(y_i | \mathbf{x}_i) + n - m \frac{1}{n-m} \sum_{i \in \bar{r}} E_{\bar{r}}(y_i - E_r(y_i | \mathbf{x}_i)). \end{aligned} \quad (51)$$

But, using (4), $E_{\bar{r}}(y_i - E_r(y_i | \mathbf{x}_i))$ can be estimated by

$$\hat{E}_{\bar{r}}(y_i - \hat{E}_r(y_i | \mathbf{x}_i)) = \frac{1}{\sum_{i \in \bar{r}} (\hat{\varphi}_i - 1)} \sum_{i \in \bar{r}} (\hat{\varphi}_i - 1) (y_i - \hat{E}_r(y_i | \mathbf{x}_i)). \quad (52)$$

Hence,

$$\begin{aligned} \hat{T}_{3,in} &= \sum_{i \in \bar{r}} y_i + \sum_{i \in \bar{r}} \hat{E}_r(y_i | \mathbf{x}_i) + (n - m) \frac{1}{\sum_{i \in \bar{r}} (\hat{\varphi}_i - 1)} \sum_{i \in \bar{r}} (\hat{\varphi}_i - 1) (y_i - \hat{E}_r(y_i | \mathbf{x}_i)) \\ &\quad + \sum_{i \in \bar{s}} \hat{E}_r \left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i | \mathbf{x}_i)} y_i \right) + (N - n) \frac{1}{\sum_{i \in \bar{r}} \hat{\varphi}_i(w_i - 1)} \sum_{i \in \bar{r}} \hat{\varphi}_i(w_i - 1) \left(y_i - \hat{E}_r \left(\frac{\hat{\varphi}_i}{\hat{E}_r(\hat{\varphi}_i | \mathbf{x}_i)} y_i \right) \right). \end{aligned} \quad (53)$$

8.3. Generalized regression estimator (GREG) under informative sampling and nonignorable nonresponse

Assume that

$$E_p(y_i) = \mathbf{x}_i' \boldsymbol{\beta} = x_{i1}\beta_1 + \cdots + x_{ip}\beta_p, \quad (54)$$

then the GREG estimator is:

$$\hat{T} = \sum_{i=1}^N \hat{E}_p(y_i) = \sum_{i=1}^N \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\phi W}, \quad \hat{\boldsymbol{\beta}}_{\phi W} = (\sum_{i \in r} \phi_i w_i \mathbf{x}_i' \mathbf{x}_i)^{-1} (\sum_{i \in r} \phi_i w_i \mathbf{x}_i' y_i). \quad (55)$$

Justification: under (54) and using (1), we can show that

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} E_p(y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 = \arg \min_{\boldsymbol{\beta}} \sum_{i \in r} \phi_i w_i (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2. \quad (56)$$

Therefore,

$$\hat{\boldsymbol{\beta}}_{\phi W} = (\sum_{i \in r} \phi_i w_i \mathbf{x}_i' \mathbf{x}_i)^{-1} (\sum_{i \in r} \phi_i w_i \mathbf{x}_i' y_i) = (\mathbf{x}'(\Phi W) \mathbf{x})^{-1} \mathbf{x}'(\Phi W) \mathbf{y}, \quad (57)$$

where $= \text{diag}(\phi_1 w_1, \dots, \phi_r w_r) \mathbf{x} = (\mathbf{x}'_1, \dots, \mathbf{x}'_r)'$. Then

$$\hat{T} = \sum_{i=1}^N \hat{E}_p(y_i) = \sum_{i=1}^N \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\phi W}. \quad (58)$$

Now, if $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})' = (1, \tilde{\mathbf{x}}_i)'$, then

$$\hat{T} = \sum_{i=1}^N \hat{E}_p(y_i) = (\sum_{i=1}^N [1, \tilde{\mathbf{x}}_i']) \begin{bmatrix} \hat{\beta}_{1\phi W} \\ \hat{\boldsymbol{\beta}}'_{\phi W} \end{bmatrix} = \sum_{i \in r} \phi_i w_i g_i y_i. \quad (59)$$

where $\tilde{\mathbf{x}}_U = (\tilde{x}_{2U}, \dots, \tilde{x}_{pU})'$, $\tilde{x}_{jU} = \frac{1}{N} \sum_{i=1}^N x_{ij}$, $\hat{\beta}_{1\phi W} = \bar{y}_{\phi W} - \hat{\boldsymbol{\beta}}'_{\phi W} \tilde{\mathbf{x}}_{\phi W}$,

$$\bar{x}_{j\phi W} = \frac{\sum_{i \in r} \phi_i w_i x_{ij}}{\sum_{i \in r} \phi_i w_i} \bar{y}_{\phi W} = \frac{\sum_{i \in r} \phi_i w_i y_i}{\sum_{i \in r} \phi_i w_i}, \text{ and} \quad (60)$$

$$g_i = N \left\{ \frac{1}{\sum_{i \in r} \phi_i w_i} + (\tilde{\mathbf{x}}_U - \tilde{\mathbf{x}}_{\phi W})' (\sum_{i \in r} \phi_i w_i \mathbf{x}_i' \mathbf{x}_i)^{-1} \mathbf{x}_i \right\}.$$

Not that if $\tilde{\mathbf{x}}_U - \tilde{\mathbf{x}}_{\phi W} = 0$, that is, $\tilde{x}_{jU} = \bar{x}_{j\phi W}$, or

$$\frac{1}{N} \sum_{i=1}^N x_{ij} = \frac{\sum_{i \in r} \phi_i w_i x_{ij}}{\sum_{i \in r} \phi_i w_i}, \quad (61)$$

then

$$\hat{T} = \left(\frac{N}{\sum_{i \in r} \phi_i w_i} \right) \sum_{i \in r} \phi_i w_i y_i. \quad (62)$$

Furthermore, $\hat{T} = \sum_{i \in r} \varphi_i w_i g_i y_i$ (59) belongs to the class of calibration estimators since $\sum_{i \in r} \varphi_i w_i g_i \mathbf{x}_i = \mathbf{X}_U$, see Deville and Särndal (1992). According to this, we can derive a calibration estimator of the finite population total $T = \sum_{i=1}^N y_i$ when sampling design is informative and nonresponse mechanism is nonignorable as follows.

It should be noted here that equation (61) can be considered as calibration constraint when sampling design is informative and nonresponse mechanism is nonignorable. That is, the calibration estimator of $T = \sum_{i=1}^N y_i$ can be obtained by minimizing

$$\sum_{i \in r} \frac{(w_i^{cal} - \varphi_i w_i)^2}{\varphi_i w_i} \quad (63)$$

with respect to w_i^{cal} , subject to the constraint

$$\sum_{i \in r} w_i^{cal} \mathbf{x}_i = \mathbf{X}_U. \quad (64)$$

Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)'$ be the Lagrange multiplier, so the Lagrange function is:

$$\psi(w_1^{cal}, \dots, w_n^{cal}; \boldsymbol{\lambda}) = \sum_{i \in r} \frac{(w_i^{cal} - \varphi_i w_i)^2}{\varphi_i w_i} - 2(\sum_{i \in r} w_i^{cal} \mathbf{x}_i - \mathbf{X}_U)' \boldsymbol{\lambda}. \quad (65)$$

Differentiating (65) with respect to w_i^{cal} and $\boldsymbol{\lambda}$, and then equating the derivatives to zero, we get the calibration weights:

$$w_i^{cal} = \varphi_i w_i (1 + \mathbf{x}_i' \boldsymbol{\lambda}). \quad (66)$$

where $\boldsymbol{\lambda}$ is determined by the constraint $\sum_{i \in r} w_i^{cal} \mathbf{x}_i = \mathbf{X}_U$, which is equal to

$$\boldsymbol{\lambda} = (\sum_{i \in r} \varphi_i w_i \mathbf{x}_i' \mathbf{x}_i)^{-1} (\sum_{i \in r} w_i^{cal} \mathbf{x}_i - \mathbf{X}_U). \quad (67)$$

If $(\sum_{i \in r} \varphi_i w_i \mathbf{x}_i' \mathbf{x}_i)$ is invertible, then the calibration estimator of $T = \sum_{i=1}^N y_i$ is

$$\hat{T}_{cal} = \sum_{i \in r} \varphi_i w_i g_i^{cal} y_i = \sum_{i \in r} w_i^{cal} y_i \quad (68)$$

where $w_i^{cal} = \varphi_i w_i g_i^{cal}$ and g_i^{cal} is given by:

$$g_i^{cal} = 1 + (\tilde{\mathbf{x}}_U - \tilde{\mathbf{x}}_{\varphi w})' (\sum_{i \in r} \varphi_i w_i \mathbf{x}_i' \mathbf{x}_i)^{-1} \mathbf{x}_i. \quad (69)$$

Variance of $\hat{T}_{cal} = \sum_{i \in r} \varphi_i w_i g_i^{cal} y_i = \sum_{i \in r} w_i^{cal} y_i$.

Following Deville and Särndal (1992), the estimated variance of \hat{T}_{cal} (equation 68) is given by:

$$\hat{V}(\hat{T}_{cal}) = \sum_{i \in r} \sum_{j \in r} \left(1 - \frac{(\psi_i \pi_i)(\psi_j \pi_j)}{(\psi_{ij} \pi_{ij})} \right) (w_i^{cal} e_i)(w_j^{cal} e_j), \quad (70)$$

where $\psi_{ij} = Pr(i, j \in r)$, $\pi_{ij} = Pr(i, j \in s)$ and $e_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\varphi w}$.

9. Conclusions

In this paper, we study, within a modeling framework, the semi-parametric prediction of finite population total, by specifying the probability distribution of the observed measurements under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases. Furthermore, based on the relationship between response distribution and population distribution, we introduced a new measure of representativeness of a response set and a new test of nonignorable nonresponse and informative sampling, jointly. In addition to that, generalized regression (GREG) and calibration estimators under informative sampling and nonignorable nonresponse are derived.

The paper is purely mathematical and focuses on the role of informativeness of sampling design and informativeness of nonresponse in adjusting various predictors for bias reduction. Further experimentation (simulation and real data problem) with this kind of semiparametric predictors, generalized measures of representativeness, tests of nonignorable nonresponse, informativeness of sampling design, and calibration estimators are therefore highly recommended. The author hopes that the new mathematical results obtained will encourage further theoretical, empirical and practical research in these directions.

Acknowledgements

I would like to thank the editor-in-chief and the referees for their valuable comments. The research was partially supported by a grant from DAAD (Deutscher Akademischer Austauschdienst German Academic Exchange Service) – Research Stays for University Academics and Scientists, 2018. Also, the author would like to thank Professor Timo Schmid for hosting me during my visit to Free University Berlin.

References

- Chambers, R., Skinner, C., (2003). *Analysis of Survey Data*. New York: John Wiley.
- Deville, J. C., Särndal, C. E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, pp. 376–382.
- Eideh, A. H., (2020). Parametric Prediction of Finite Population Total under Informative Sampling and Nonignorable Nonresponse. *Statistics In Transition new series*, March 2020, vol. 21, no.1, pp. 13–35.

- Eideh, A. H., (2016). Estimation of Finite Population Mean and Superpopulation Parameters when the Sampling Design is Informative and Nonresponse Mechanism is Nonignorable. *Pakistan Journal of Statistics and Operation Research (PJSOR)*. *Pak.j.stat.oper.res*, vol. XII, no. 3, 2016, pp. 467–489.
- Eideh A. H., (2012). Estimation and Prediction under Nonignorable Nonresponse via Response and Nonresponse Distributions. *Journal of Indian Society of Agriculture Statistics*, 66(3), pp. 359–380.
- Eideh A. H., (2009). On the use of the Sample Distribution and Sample Likelihood for Inference under Informative Probability Sampling. *DIRASAT (Natural Science)*, vol. 36 (2009), no.1, pp. 18–29.
- Eideh A. H., (2007). Method of Moments Estimators of Finite Population Parameters in Case of Full Response and Nonresponse. *Contributed Paper for the 56th Biennial Session of the International Statistical Institute*, August 22–29, Lisboa, Portugal, ISI 2007 Book of Abstracts, p. 430.
- Fuller W. A., (2009). *Sampling Statistics*. Wiley.
- Little, R. J. A., (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, pp. 237–250.
- Pfeffermann, D., Krieger, A. M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, pp. 1087–1114.
- Schouten, B., Cobben, F. and Bethlehem, J., (2009). Indicators for the representativeness of survey Response. *Survey Methodology*, 35, pp. 101–113.
- Sverchkov, M., (2008). A New Approach to Estimation of Response Probabilities When Missing Data Are Not Missing at Random. *Proceedings of the Survey Research Methods Section*, pp. 867–874.
- Sverchkov, M., Pfeffermann, D., (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, pp. 79–92.

A new confidence interval for the odds ratio

Zofia Zielińska-Kolasińska,¹ Wojciech Zieliński²

ABSTRACT

We consider the problem of interval estimation of the odds ratio. An asymptotic confidence interval is widely applied in economics, medicine, sociology, etc. Unfortunately, this confidence interval has a poor coverage probability, significantly smaller than the nominal confidence level. In this paper, a new confidence interval is proposed. Its construction requires only information on the sizes of samples and the sample odds ratio. The coverage probability of the proposed confidence interval is at least the nominal confidence level.

Key words: confidence interval, odds ratio.

1. Introduction

In many practical sciences such as economy, medicine, sociology, etc. dichotomous variate is observed. Such variate is to be compared in two independent groups. Commonly used is the difference of two fractions (the risk difference), the ratio of two proportions (the relative risk) and the odds ratio. The relative risk and the odds ratio are relative measurements for comparison of two variates, while the risk difference is an absolute measurement.

The odds ratio is one of the parameters commonly used in such comparisons, especially in two-arm binomial experiments. This indicator was firstly applied by Cornfield (1951). The literature devoted to the analysis of odds ratio and its estimators is very rich, see, e.g. Encyclopedia of Statistical Sciences (2006) Volume 9, pp. 5722–5726 and the literature therein.

However, the problem is in the interval estimation. In general, there are two approaches to the problem. The first one consists in the analysis of 2×2 tables (Edwards (1963), Gart (1971), Thomas (1971)). The second approach is based on logistic model in which the odds ratio has a direct relationship with the regression coefficient (Gart (1971), McCullagh (1980), Morris (1988)). Wang, Shan (2015) constructed exact confidence interval for the odds ratio based on another approach. Namely, they applied the so-called rank function. A very exhaustive review of different confidence intervals for the odds ratio may be found in Andrés et.al (2020). Unfortunately, all those confidence intervals have one very important disadvantage: their real probability of coverage is significantly smaller than the nominal one. It is in contradiction to the Neyman (1934 p. 562) definition of a confidence interval. Hence, the risk of a wrong conclusion (i.e. overestimation or underestimation) is greater than the assumed one and unluckily remains unknown.

¹Poland. E-mail: zzk@egrp.pl, ORCID: <https://orcid.org/000-0001-8845-758X>.

²Department of Econometrics and Statistics, Warsaw University of Life Sciences, Warsaw, Poland.
E-mail: wojciech_zielinski@sggw.edu.pl, ORCID: <https://orcid.org/0000-0003-0749-8764>.

The most commonly used in applications is an asymptotic interval for odds ratio derived from logistic model (formula (4) in Section 3). This asymptotic interval is widely used in different statistical packages. There are also many internet scripts for calculating the asymptotic confidence interval (see, e.g. <http://www.hutchon.net/ConfidOR.htm>). Unfortunately, this confidence interval has some statistical disadvantages discussed in Section 3. To avoid those disadvantages a new confidence interval is proposed. The idea of construction is similar to the idea of construction of the confidence interval for the difference of two probabilities of success (the risk difference) proposed by Zieliński (2020a). It is based on the exact distribution of the sample odds ratio, hence it works for large as well as for small samples. The coverage probability of that confidence interval is at least the nominal confidence level.

In Section 2 a new confidence interval is constructed. In Section 3 some disadvantages of the asymptotic confidence interval are discussed. An example of application is given in Section 4. Final conclusions are given in Section 5.

2. A new confidence interval

Consider two independent r.v.'s ξ_A and ξ_B distributed as $Bin(n_A, p_A)$ and $Bin(n_B, p_B)$, respectively. The problem is in estimating the odds ratio:

$$OR = \frac{(p_A/(1-p_A))}{(p_B/(1-p_B))} = \frac{p_A}{(1-p_A)} \cdot \frac{(1-p_B)}{p_B}.$$

Let n_{A1} and n_{B1} be observed numbers of successes. The data are usually organized in a 2×2 table:

	success	failure	
Group A	n_{A1}	n_{A0}	n_A
Group B	n_{B1}	n_{B0}	n_B
	n_1	n_0	n

The standard estimator of OR is as follows:

$$\widetilde{OR} = \frac{n_{A1}}{n_A - n_{A1}} \cdot \frac{n_B - n_{B1}}{n_{B1}}. \quad (1)$$

The estimator \widetilde{OR} is undefined for $n_{A1} = n_A$ or $n_{B1} = 0$. The probability of the nonexistence of \widetilde{OR} equals

$$P_{p_A, p_B} \{ \xi_A = n_A \text{ or } \xi_B = 0 \} = p_A^{n_A} + (1-p_B)^{n_B} - p_A^{n_A} (1-p_B)^{n_B}.$$

For a given odds ratio equal to $r > 0$

$$p_B = \frac{p_A}{p_A + r(1-p_A)} \text{ and } 1-p_B = \frac{r(1-p_A)}{p_A + r(1-p_A)}$$

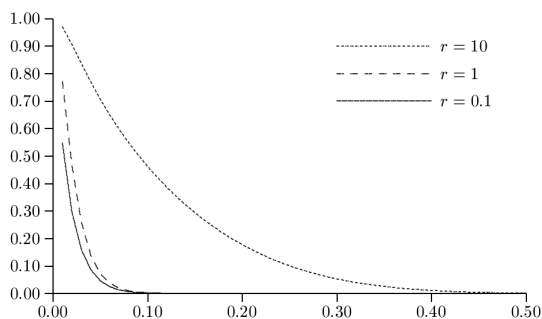


Figure 1: The probability $P_{r,p_A} \{\xi_A = n_A \text{ or } \xi_B = 0\}$

and

$$P_{r,p_A} \{\xi_A = n_A \text{ or } \xi_B = 0\} = p_A^{n_A} + (1 - p_A^{n_A}) \left(\frac{r(1 - p_A)}{p_A + r(1 - p_A)} \right)^{n_B}.$$

In Figure 1 the above probability (y axis) is shown for different values of true odds ratio r with respect to probability p_A (x axis). In Figure 1 $n_A = 60$ and $n_B = 70$ were taken.

It is seen that the probability of nonexistence of \widetilde{OR} is quite high, especially for small values of the probability p_A . To eliminate that phenomena another approach is needed.

Usually, the problem of estimating an odds ratio is considered in the following statistical model:

$$(\{0, 1, \dots, n_A\} \times \{0, 1, \dots, n_B\}, \{Bin(n_A, p_A) \cdot Bin(n_B, p_B), (p_A, p_B) \in (0, 1) \times (0, 1)\}).$$

Since we are interested in estimating the odds ratio OR , consider now a new statistical model. This model is the one-parameter model: the odds ratio is an unknown parameter

$$(\mathcal{X}, \{F_r, 0 \leq r \leq +\infty\}),$$

where

$$\mathcal{X} = \left\{ \frac{n_{A1}}{n_A - n_{A1}} \cdot \frac{n_B - n_{B1}}{n_{B1}} : n_{A1} \in \{0, 1, \dots, n_A\}, n_{B1} \in \{0, 1, \dots, n_B\} \right\}.$$

The cumulative distribution functions (CDF) $F_r(\cdot)$ are defined as follows.

Since the estimator \widetilde{OR} given by formula (1) is undefined for $n_{A1} = n_A$ or $n_{B1} = 0$ we

extend the definition of the estimator of the odds ratio in the following way:

$$\widehat{OR} = \begin{cases} 0, & \text{for } (n_{A1} = 0, n_{B1} \geq 1) \text{ or } (n_{A1} \leq n_A - 1, n_{B1} = n_B), \\ +\infty, & \text{for } (n_{A1} = n_A, 1 \leq n_{B1} \leq n_B - 1) \text{ or } (n_{A1} \geq 1, n_{B1} = 0), \\ 1, & \text{for } (n_{A1} = 0, n_{B1} = 0) \text{ or } (n_{A1} = n_A, n_{B1} = n_B), \\ \text{formula (1),} & \text{elsewhere.} \end{cases} \quad (2)$$

The probability of observing $\xi_A = n_{A1}$ and $\xi_B = n_{B1}$ equals

$$P_{p_A, p_B} \{n_{A1}, n_{B1}\} = \binom{n_A}{n_{A1}} p_A^{n_{A1}} (1 - p_A)^{n_A - n_{A1}} \binom{n_B}{n_{B1}} p_B^{n_{B1}} (1 - p_B)^{n_B - n_{B1}}.$$

Equivalently

$$P_{r, p_A} \{n_{A1}, n_{B1}\} = r^{n_B - n_{B1}} \binom{n_A}{n_{A1}} \binom{n_B}{n_{B1}} \frac{p_A^{n_{A1} + n_{B1}} (1 - p_A)^{n_A + n_B - n_{A1} - n_{B1}}}{(p_A + r(1 - p_A))^{n_B}}.$$

Let

$$F_{r, p_A}(t) = P_{r, p_A} \{ \widehat{OR} \leq t \} = \sum_{n_{A1}=0}^{n_A} \sum_{n_{B1}=0}^{n_B} P_{r, p_A} \{n_{A1}, n_{B1}\} \mathbf{1} \left(\widehat{OR}(n_{A1}, n_{B1}) \leq t \right),$$

where $\mathbf{1}(q) = 1$ when q is true and $= 0$ elsewhere. For any given $p_A \in (0, 1)$ the family $\{F_{r, p_A}, r > 0\}$ is stochastically ordered, i.e.

$$F_{r_1, p_A}(\cdot) \geq F_{r_2, p_A}(\cdot) \text{ for } r_1 \leq r_2.$$

It follows from the fact that for a given n_{A1}, n_{B1} and p_A the probability $P_{r, p_A} \{n_{A1}, n_{B1}\}$ is the decreasing function of odds ratio r .

Let γ be the given confidence level and let \hat{r} be the observed odds ratio. For a given p_A the confidence interval for r takes on the form

$$(Left(\hat{r}, p_A), Right(\hat{r}, p_A)), \quad (3)$$

where

$$\begin{cases} Left(\hat{r}, p_A) = \max \{r : G_{r, p_A}(\hat{r}) \geq (1 + \gamma)/2\}, \\ Right(\hat{r}, p_A) = \min \{r : F_{r, p_A}(\hat{r}) \leq (1 - \gamma)/2\}. \end{cases}$$

Here, $G_{r, p_A}(t)$ denotes the probability $P_{r, p_A} \{ \widehat{OR} < t \}$.

The coverage probability by the construction is at least a given confidence level γ . In Figure 2 the coverage probability for $p_A = 0.5$ and $n_A = 50, n_B = 10$ is presented ($\gamma = 0.95$). For other values of $p_A \in (0, 1)$ the graphs are similar. On the x -axis the value r of the odds ratio is given and on the y -axis the probability of coverage is shown.

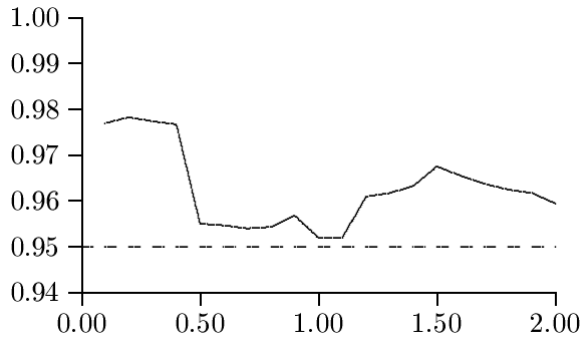


Figure 2: Coverage probability of (3)

Since the probability p_A is unknown it should be treated as a nuisance parameter. In statistics two methods of eliminating a nuisance parameter are common: estimation of such parameter or integration over nuisance parameter. In what follows we choose the second method, i.e. appropriate integration:

$$P_r\{n_{A1}, n_{B1}\} = \int_0^1 P_{r,p_A}\{n_{A1}, n_{B1}\} w(p_A) dp_A,$$

where $w: (0, 1) \rightarrow R^+$ is a weighting function such that $\int_0^1 w(u) du = 1$. The function w may be chosen quite arbitrary. The choice is interpreted as an *a priori* knowledge of probability p_A . The most common is the choice of function w proportional to $(u-a)^{\alpha-1}(b-u)^{\beta-1}$ for positive α and β and $0 \leq a < b \leq 1$. In what follows $\alpha = \beta = 1$, $a = 0$ and $b = 1$ is taken. So, it is assumed that the probability p_A may be any number from the interval $(0, 1)$.

We obtain

$$\begin{aligned} P_r\{n_{A1}, n_{B1}\} &= \int_0^1 P_{r,p_A}\{n_{A1}, n_{B1}\} dp_A \\ &= (n_A + n_B)! \frac{\binom{n_A}{n_{A1}} \binom{n_B}{n_{B1}}}{\binom{n_A + n_B}{n_{A1} + n_{B1}}} \left(\frac{1}{r}\right)^{n_{B1}} {}_2\tilde{F}_1\left[n_B, n_{A1} + n_{B1} + 1; n_A + n_B + 2; 1 - \frac{1}{r}\right], \end{aligned}$$

where

$${}_2\tilde{F}_1[x, y; z; t] = \frac{1}{\Gamma(z-y)\Gamma(y)} \int_0^1 u^{y-1} (1-u)^{z-y-1} (1-ut)^{-x} du \quad (\text{for } z > y > 0)$$

is the regularized confluent hypergeometric function. The CDF of \widehat{OR} equals (for $t \geq 0$)

$$F_r(t) = P_r\{\widehat{OR} \leq t\} = \sum_{n_{A1}=0}^{n_A} \sum_{n_{B1}=0}^{n_B} P_r\{n_{A1}, n_{B1}\} \mathbf{1}(\widehat{OR}(n_{A1}, n_{B1}) \leq t),$$

where $\mathbf{1}(q) = 1$ when q is true and $= 0$ elsewhere.

Since \widehat{OR} is given by formula (2) the CDF may be written as

$$\begin{aligned} F_r(t) &= \sum_{n_{A1}=0}^{n_A-1} \sum_{n_{B1}=h(n_{A1})}^{n_B} P_r\{n_{A1}, n_{B1}\} \\ &= (n_A + n_B)! \sum_{n_{A1}=0}^{n_A-1} \sum_{n_{B1}=h(n_{A1})}^{n_B} \frac{\binom{n_A}{n_{A1}} \binom{n_B}{n_{B1}}}{\binom{n_A+n_B}{n_{A1}+n_{B1}}} \left(\frac{1}{r}\right)^{n_{B1}} {}_2\widetilde{F}_1\left[n_B, n_{A1} + n_{B1} + 1; n_A + n_B + 2; 1 - \frac{1}{r}\right], \end{aligned}$$

where

$$h(n_{A1}) = \begin{cases} \left\lceil \frac{n_B}{t\left(\frac{n_A}{n_{A1}} - 1\right) + 1} \right\rceil, & \text{for } n_{A1} \geq 1, \\ 0, & \text{for } n_{A1} = 0, \end{cases}$$

(here $\lceil x \rceil$ denotes the smallest integer not less than x).

The family $\{F_r, r \geq 0\}$ is stochastically ordered, i.e. for a given $t > 0$

$$F_{r_1}(t) \geq F_{r_2}(t) \text{ for } r_1 \leq r_2.$$

It follows from the fact that for a given n_{A1}, n_{B1} and p_A the probability $P_{r,p_A}\{n_{A1}, n_{B1}\}$ is the decreasing function of odds ratio r and hence $P_r\{n_{A1}, n_{B1}\}$ is also decreasing in r .

Let $G_r(t)$ denote the probability $P_r\{\widehat{OR} < t\}$. Let γ be the given confidence level and let \hat{r} be the observed odds ratio. The confidence interval for r takes on the form

$$(Left(\hat{r}), Right(\hat{r})), \quad (4)$$

where

$$Left(\hat{r}) = \begin{cases} 0, & \hat{r} = 0, \\ 0, & \text{if } \lim_{r \rightarrow 0} G_r(\hat{r}) < (1 + \gamma)/2, \\ r_*, & r_* = \max\{r : G_r(\hat{r}) \geq (1 + \gamma)/2\}, \end{cases}$$

and

$$Right(\hat{r}) = \begin{cases} \infty, & \hat{r} = \infty, \\ \infty, & \text{if } \lim_{r \rightarrow \infty} F_r(\hat{r}) > (1 - \gamma)/2, \\ r^*, & r^* = \min\{r : F_r(\hat{r}) \leq (1 - \gamma)/2\}. \end{cases}$$

Theorem. For $n_A > \frac{2}{1-\gamma} - 1$ the confidence interval for the odds ratio is two-sided and is one-sided otherwise.

For the proof see Appendix 1.

If \hat{r} is the observed odds ratio then the confidence interval for r takes on the following

form:

$$\begin{aligned} \text{for } \hat{r} \in [0, 1) : & \begin{cases} \langle 0, r^* \rangle, & \text{for } n_A \leq \frac{2}{1-\gamma} - 1, \\ (r_*, r^*), & \text{for } n_A > \frac{2}{1-\gamma} - 1, \end{cases} \\ \text{for } \hat{r} \in [1, +\infty) : & \begin{cases} (r_*, +\infty), & \text{for } n_A \leq \frac{2}{1-\gamma} - 1, \\ (r_*, r^*), & \text{for } n_A > \frac{2}{1-\gamma} - 1, \end{cases} \end{aligned}$$

where r_* and r^* are given by formula (4). Minimal sample sizes n_A for which two-sided confidence interval exists are given in Table 1.

Table 1: Minimal sample size

γ	0.9	0.95	0.99	0.999
n_A	20	40	200	2000

For a given $r > 0$ the coverage probability, by construction, equals at least γ . Figure 3 shows the coverage probability for $n_A = 60$, $n_B = 70$ and $\gamma = 0.95$. On the x -axis the value r of the odds ratio is given and on the y -axis the probability of coverage is shown. The coverage probabilities are calculated, not simulated.

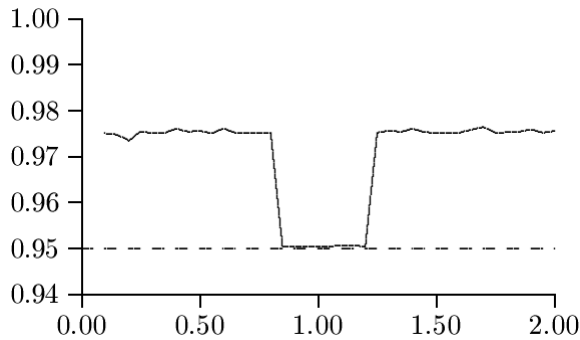


Figure 3: Coverage probability of (4)

Remark. The above considerations are made for A versus B . It is obvious that

$$OR(A \text{ vs } B) = \frac{1}{OR(B \text{ vs } A)}.$$

It is easily seen that the new confidence interval has the following natural property:

$$Left(A \text{ vs } B) = \frac{1}{Right(B \text{ vs } A)} \quad \text{and} \quad Right(A \text{ vs } B) = \frac{1}{Left(B \text{ vs } A)}.$$

In the case of considering B versus A in the Theorem, the sample size n_A should be changed to n_B .

3. Standard confidence interval

Estimating the odds ratio is one of the crucial problems in medicine, biometrics, etc. The most widely used confidence interval at the confidence level γ is of the form

$$\left(\widetilde{OR} \cdot \exp \left(u_{\frac{1-\gamma}{2}} \sqrt{\frac{1}{n_{A1}} + \frac{1}{n_{A0}} + \frac{1}{n_{B1}} + \frac{1}{n_{B0}}} \right), \widetilde{OR} \cdot \exp \left(u_{\frac{1+\gamma}{2}} \sqrt{\frac{1}{n_{A1}} + \frac{1}{n_{A0}} + \frac{1}{n_{B1}} + \frac{1}{n_{B0}}} \right) \right), \quad (5)$$

where u_δ denotes the δ quantile of $N(0, 1)$ distribution. In the above formula the estimator \widetilde{OR} is given by (1). Unfortunately, this confidence interval has at least three disadvantages. They are as follows.

1. Confidence interval (5) does not exist if at least one of n_{A0} , n_{A1} , n_{B0} or n_{B1} equals zero or \widetilde{OR} does not exist. The probability of such an event may be quite large, so in many real experiments it may happen (cf. Figure 1) that the confidence interval is undefined.

2. The coverage probability of c.i. (5) is less than the nominal one. In Figure 4 the coverage probability is shown for $n_A = 60$, $n_B = 70$ and $\gamma = 0.95$ (the value r of odds ratio is given on the x -axis and the coverage probability is given on the y -axis). The probability of wrong conclusion, i.e. of overestimation or underestimation is greater than the assumed 0.05. It means that the true value of odds ratio may be smaller than the left end of the confidence interval (4) or greater than its right end. The risk of such an event is greater than the nominal 0.05 and unfortunately remains unknown. Note that this is in contradiction to Neyman (1934, p. 562) definition of a confidence interval.

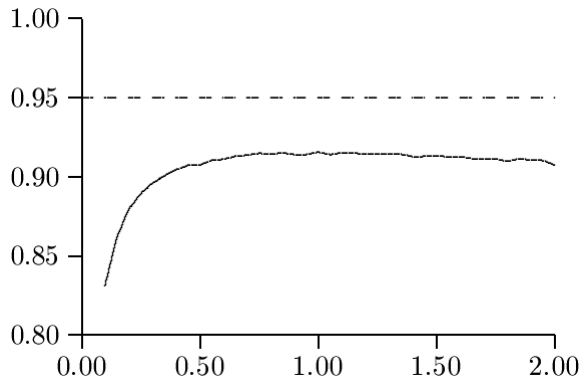


Figure 4: Coverage probability of (5)

3. The standard asymptotic confidence interval requires the knowledge of sample sizes as well as sample proportions in each sample. Unfortunately, it may lead to misunderstandings. Namely, suppose that six experiments were conducted. In each experiment two samples of sizes sixty and seventy, respectively, were drawn ($n_1 = 60$, $n_2 = 70$). The resulting

numbers of successes are shown in Table 2 (the first two columns).

Table 2: Confidence intervals in six experiments

n_{A1}	n_{B1}	\widetilde{OR}	<i>left</i>	<i>right</i>
6	14	0.4444	0.1592	1.2410
8	18	0.4444	0.1776	1.1122
15	30	0.4444	0.2095	0.9428
24	42	0.4444	0.2199	0.8985
36	54	0.4444	0.2078	0.9506
48	63	0.4444	0.1627	1.2141

It is seen that the sample odds ratio (the third column) is the same in all experiments, but the confidence intervals are quite different. Moreover, for example, in the first experiment it may be claimed that the population odds in groups *A* and *B* may be treated as equal, while in the fourth one such a conclusion should not be drawn.

4. An example of application

The aim of the study was to compare the chances of survival of trading companies in Mazowieckie voivodship versus Warsaw (Poland). The question was about the chances of surviving during the first ten years of activity (Zieliński 2020b).

Let p_A denote the probability of surviving the first ten years of activity of a firm established in Mazowieckie voivodship, and let p_B denote the appropriate probability for a firm established in Warsaw. We are interested in the estimation of the odds ratio, i.e. $(p_A/(1 - p_A))/(p_B/(1 - p_B))$.

From the REGON (*National Business Registry Number*) registry it is known that 32760 firms started their activity in 2007. Among them 17130 were established in Mazowieckie voivodship, while 15630 were established in Warsaw. Among firms established in 2007 the random sample of size 320 was taken and it was observed how many of those firms were still active in 2017. The data are given in Table 3.

Table 3: Random sample of firms

	Active	Nonactive	
Mazowieckie	96	74	170
Warsaw	85	65	150

On the basis of those data the odds ratio would be estimated.

Note that the estimator of the odds ratio is defined for random variables distributed as binomial. In our investigation we deal with random variables distributed as hypergeometric. It is well known that hypergeometric distribution may be approximated by an appropriate binomial distribution. Some remarks on consequences of such approximation may be found in Zieliński (2011). In what follows, it is assumed that binomial approximation to the hypergeometric one is fairly enough.

The estimate of odds for Mazowieckie voivodship equals $(96/170)/(74/170) = 1.297$. It means that almost 30% more of the firms established in 2007 were still working than were nonactive. A similar indicator for Warsaw equals 1.308.

The estimate of odds ratio for Mazowieckie voivodship versus Warsaw equals $1.292/1.308 = 0.992$. The confidence interval (4) at 95% confidence level is $(0.437, 2.049)$. Since this confidence interval covers 1, it may be expected that for the firms established in 2007 the chances of surviving the first ten years of activity for Mazowieckie voivodship and for Warsaw are similar.

The above conclusion may of course be wrong. It must be stressed that the risk of over- or under-estimation is at most 5%, in contradiction to the standard confidence interval.

Simple calculations show that the standard confidence interval (5) at 95% confidence level for odds ratio is $(0.989, 1.544)$. This confidence interval is narrower than (4), but unfortunately the risk of not covering the true value of the odds ratio is greater than assumed 5% and remains unknown.

Table 4: Number of firms in REGON registry in 2007

	Active	Nonactive	
Mazowieckie	9448	7682	17130
Warsaw	9607	6023	15630

In the presented example we are very lucky since we have full information about the number of firms established in 2007 which survived until 2017. Hence, we may calculate the exact value of odds ratio for that population. Those data are presented in Table 4 (data comes from the REGON registry).

The exact value of odds ratio in that population equals $(9448/7682)/(9607/6023) = 0.771$. Note that the new confidence interval (4) covers this value, while the standard asymptotic confidence interval does not.

5. Conclusions

In this paper a new confidence interval for the odds ratio is proposed. The confidence interval is based on the exact distribution of the sample odds ratio, hence it works for large as well as for small samples. The coverage probability of that confidence interval is at least the nominal confidence level, in contrast to the asymptotic confidence intervals known in the literature. It must be noted that the information on the sample sizes and the sample odds ratio is sufficient for constructing the new confidence interval. Unfortunately, no closed formulae for the ends of the confidence interval are available. However, for given n_A , n_B and observed \widehat{OR} the ends may be easily numerically computed with the aid of the standard software such as R, Mathematica, etc. (see Appendix 2).

Since the proposed confidence interval may be applied for small as well as for large sample sizes, it may be recommended for practical use.

References

- Baumol, W. J., (2015). *Macroeconomics: Principles and policy*, Cengage Learning, Inc.
- Andrés, M.A., Mato, S.A., Tejedor, H.I., (2020). Pseudo-Bayesian test for the comparison of two proportions, *Metron*, 49 (1-4), pp. 151–162.
- García-Pérez M.A., Núñez-Antón V. (2020) Asymptotic versus exact methods in the analysis of contingency tables: Evidence-based practical recommendations, *Stat Methods Med Res.*, 29(9), pp. 2569–2582.
- Cornfield, J., (1951). A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix, *JNCI: Journal of the National Cancer Institute*, 11, pp. 1269–1275, DOI: 10.1093/jnci/11.6.1269.
- Edwards, A.W.F., (1963). The Measure of Association in a 2×2 Table. *Journal of the Royal Statistical Society, Ser. A.* 126, pp. 109–114, DOI: 10.2307/2982448.
- Encyclopedia of Statistical Sciences, (2006). Wiley & Sons.
- García-Pérez M.A., Núñez-Antón V., (2020) Asymptotic versus exact methods in the analysis of contingency tables: Evidence-based practical recommendations. *Stat Methods Med Res.*, 29(9), pp. 2569–2582.
- Gart, J.J., (1971). The comparison of proportions: a review of significance tests, confidence intervals, and adjustments for stratification. *Review of the International Statistical Institute*, 39, pp. 148–169.
- Lawson, R., (2004). Small Sample Confidence Intervals for the Odds Ratio. *Communications in Statistics - Simulation and Computation*, 33, pp. 1095–1113, DOI: 10.1081/SAC-200040691.
- Morris, J.A., Gardner M.J., (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal*, 296, pp. 1313–6, DOI: 10.1136/bmj.296.6632.1313.
- McCullagh, P., (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society, Ser. B.* 42, pp. 109–142.
- Neyman, J. (1934). On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97, pp. 558–625.
- Thomas, D.G., (1971). Algorithm AS-36: exact confidence limits for the odds ratio in a 2×2 table. *Applied Statistics*, 20, pp. 105–110.
- Wang, W., Shan G., (2015) Exact Confidence Intervals for the Relative Risk and the Odds Ratio. *Biometrics*, 71, pp. 985-995, DOI: 10.1111/biom.12360.

- Zieliński, W., (2011) Comparison of confidence intervals for fraction in finite populations. *Quantitative Methods in Economics*, XII, pp. 177–182.
- Zieliński, W., (2020a). A new exact confidence interval for the difference of two binomial proportions. *REVSTAT-Statistical Journal*, 18, pp. 521–530.
- Zieliński, W., (2020b). A New Confidence Interval for the Odds Ratio: an Application to the Analysis of the Risk of Survival of an Enterprise. *The 14th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, pp. 185–191.

Appendix 1

A few remarks before the proof.

Remark 1. For $1 \leq n_{A1} \leq n_A - 1$ and $1 \leq n_{B1} \leq n_B - 1$

$$P_r\{n_{A1}, n_{B1}\} \rightarrow \begin{cases} 0, & \text{as } r \rightarrow 0 \\ 0, & \text{as } r \rightarrow +\infty \end{cases}$$

Proof of Remark 1. For $1 \leq n_{A1} \leq n_A - 1$ and $1 \leq n_{B1} \leq n_B - 1$

$$\begin{aligned} P_{r,p_A}\{n_{A1}, n_{B1}\} &\propto p_A^{n_{A1}}(1-p_A)^{n_A-n_{A1}} \cdot \left(\frac{p_A}{p_A+r(1-p_A)}\right)^{n_{B1}} \left(\frac{r(1-p_A)}{p_A+r(1-p_A)}\right)^{n_B-n_{B1}} \\ &\rightarrow \begin{cases} 0, & \text{as } r \rightarrow 0 \\ 0, & \text{as } r \rightarrow +\infty \end{cases} \end{aligned}$$

Hence, $P_r\{n_{A1}, n_{B1}\} \rightarrow 0$ as $r \rightarrow 0$ or $r \rightarrow \infty$.

Remark 2. $P_r\{\widehat{OR} = 0\} \rightarrow \begin{cases} \frac{n_A}{n_A+1}, & \text{as } r \rightarrow 0 \\ 0, & \text{as } r \rightarrow +\infty \end{cases}$

Proof of Remark 2. Note that $\widehat{OR} = 0$ if and only if $(n_{A1} = 0 \text{ and } n_{B1} \geq 1)$ or $(1 \leq n_{A1} \leq n_A - 1 \text{ and } n_{B1} = n_B)$. Hence,

$$\begin{aligned} P_{r,p_A}\{\widehat{OR} = 0\} &= (1-p_A)^{n_A} \sum_{n_{B1} \geq 1} \binom{n_B}{n_{B1}} p_B^{n_{B1}} (1-p_B)^{n_B-n_{B1}} + p_B^{n_B} \sum_{n_{A1}=1}^{n_A-1} \binom{n_A}{n_{A1}} p_A^{n_{A1}} (1-p_A)^{n_A-n_{A1}} \\ &= (1-p_A)^{n_A} \left(1 - \left(\frac{r(1-p_A)}{p_A+r(1-p_A)}\right)^{n_B}\right) + \left(\frac{p_A}{p_A+r(1-p_A)}\right)^{n_B} (1-p_A^{n_A} - (1-p_A)^{n_A}) \\ &\rightarrow \begin{cases} (1-p_A)^{n_A} + (1-p_A^{n_A} - (1-p_A)^{n_A}) = 1-p_A^{n_A}, & \text{as } r \rightarrow 0 \\ 0, & \text{as } r \rightarrow +\infty \end{cases} \end{aligned}$$

We obtain

$$P_r\{\widehat{OR} = 0\} = \int_0^1 P_{r,p_A}\{\widehat{OR} = 0\} dp_A \rightarrow \begin{cases} \frac{n_A}{n_A+1}, & \text{as } r \rightarrow 0 \\ 0, & \text{as } r \rightarrow +\infty \end{cases}$$

Remark 3. $P_r\{\widehat{OR} = 1\} \rightarrow \begin{cases} \frac{1}{n_A+1}, & \text{as } r \rightarrow 0 \\ \frac{1}{n_A+1}, & \text{as } r \rightarrow +\infty \end{cases}$

Proof of Remark 3. Note that $\widehat{OR} = 1$ iff $n_{A1}n_B = n_{B1}n_A$. Hence,

$$\begin{aligned} & P_{r,p_A}\{\widehat{OR} = 1\} \\ &= (1-p_A)^{n_A}(1-p_B)^{n_B} + p_A^{n_A}p_B^{n_B} + \sum_{n_{A1}=1}^{n_A-1} P_{r,p_A}\{n_{A1}, n_{B1}\} \\ &= (1-p_A)^{n_A} \left(\frac{r(1-p_A)}{p_A+r(1-p_A)} \right)^{n_B} + p_A^{n_A} \left(\frac{p_A}{p_A+r(1-p_A)} \right)^{n_B} + \sum_{n_{A1}=1}^{n_A-1} P_{r,p_A}\{n_{A1}, n_{B1}\} \\ &\rightarrow \begin{cases} p_A^{n_A}, & \text{as } r \rightarrow 0 \\ (1-p_A)^{n_A}, & \text{as } r \rightarrow +\infty \end{cases} \end{aligned}$$

We obtain

$$P_r\{\widehat{OR} = 1\} = \int_0^1 P_{r,p_A}\{\widehat{OR} = 1\} dp_A \rightarrow \begin{cases} \frac{1}{n_A+1}, & \text{as } r \rightarrow 0 \\ \frac{1}{n_A+1}, & \text{as } r \rightarrow +\infty \end{cases}$$

Theorem. For $n_A > \frac{2}{1-\gamma} - 1$ the confidence interval for r is two-sided and is one-sided otherwise.

Proof. For $0 < t < 1$ we have

$$P_r\{\widehat{OR} \leq t\} = P_r\{\widehat{OR} = 0\} + P_r\{0 < \widehat{OR} \leq t\} \rightarrow \begin{cases} \frac{n_A}{n_A+1}, & \text{as } r \rightarrow 0 \\ 0, & \text{as } r \rightarrow +\infty \end{cases}$$

If $\frac{n_A}{n_A+1} > \frac{1+\gamma}{2}$, i.e. $n_A > \frac{2}{1-\gamma} - 1$, the confidence interval is two-sided. Otherwise, the c.i. is one-sided with the left end equal to 0.

For $1 \leq t < +\infty$ we have

$$P_r\{\widehat{OR} \leq t\} = P_r\{\widehat{OR} < 1\} + P_r\{\widehat{OR} = 1\} + P_r\{1 < \widehat{OR} < +\infty\} \rightarrow \begin{cases} 1, & \text{as } r \rightarrow 0 \\ \frac{1}{n_A+1}, & \text{as } r \rightarrow +\infty \end{cases}$$

If $\frac{1}{n_A+1} < \frac{1-\gamma}{2}$, i.e. $n_A > \frac{2}{1-\gamma} - 1$, the confidence interval is two-sided. Otherwise, the c.i. is one sided with the right end equal to $+\infty$.

Appendix 2

An exemplary R code for calculating the confidence interval for the odds ratio is enclosed.

```
OR=function(n,m){
  ifelse(m[1]==0 & m[2]==0,0,
  ifelse(m[1]==n[1] & m[2]==n[2],2*(n[1]-1)*(n[2]-1),
  ifelse(m[2]==0,2*(n[1]-1)*(n[2]-1),
  ifelse(m[1]==n[1],2*(n[1]-1)*(n[2]-1),m[1]*(n[2]-m[2])/(n[1]-m[1])/m[2])
  )))}

f=function(rr,k1,k2,pA){dbinom(k1,n[1],pA)*dbinom(k2,n[2],pA/(pA+rr*(1-pA)))}

nieostr=function(rr,tt){
  line<-0
  prawd=c()
  for (k1 in 0:(n[1]-1)){
    RS=round(n[2]/(tt*(n[1]/k1-1)+1),2)
    Niod=ifelse(k1==0,ifelse(tt<1,1,0),ceiling(RS))
    for (k2 in Niod:n[2])
      {mrob=c(k1,k2)
      line=line+1;
      prawd[line]=integrate(f,0,1,rr=rr,k1=k1,k2=k2,subdivisions = 1000L,
        stop.on.error = FALSE)$value;}}
  td=sum(prawd)}

ostr=function(rr,tt){
  line<-0
  prawd=c()
  for (k1 in 0:(n[1]-1)){
    RS=round(n[2]/(tt*(n[1]/k1-1)+1),2)
    Osod=ifelse(k1==0,ifelse(tt<=1,1,0),ifelse(RS==trunc(RS),RS+1,ceiling(RS)))
    for (k2 in Osod:n[2])
      {mrob=c(k1,k2)
      line=line+1;
      prawd[line]=integrate(f,0,1,rr=rr,k1=k1,k2=k2,subdivisions = 1000L,
        stop.on.error = FALSE)$value;}}
  tg=sum(prawd)}

CI=function(n,m,level){
  orobs<-OR(n,m)
  eps=1e-4
  ifelse(orobs<1,
  {ifelse(n[1]<=2/(1-level)-1,
  {L=0;
  P=uniroot(function(t){ostr(t,orobs)-(1-level)/2}, lower = orobs,
    upper = 2*(n[1]-1)*(n[2]-1),tol = eps)$root},
  {L=uniroot(function(t){nieostr(t,orobs)-(1+level)/2}, lower = 0.00000001,
    upper = orobs, tol = eps)$root;
  P=uniroot(function(t){ostr(t,orobs)-(1-level)/2}, lower = orobs,
    upper = 2*(n[1]-1)*(n[2]-1), tol = eps)$root}}),
  {ifelse(n[1]<=2/(1-level)-1,
  {L=uniroot(function(t){nieostr(t,orobs)-(1+level)/2}, lower = 0.00000001,
    upper = orobs, tol = eps)$root;
  P=Inf},
  {L=uniroot(function(t){nieostr(t,orobs)-(1+level)/2}, lower = 0.00000001,
    upper = orobs, tol = eps)$root;
  P=uniroot(function(t){ostr(t,orobs)-(1-level)/2}, lower = orobs,
```

```

    upper = 2*(n[1]-1)*(n[2]-1), tol = eps)$root}})
)
print(paste("Confidence interval for odds ratio (",round(L,5),"",round(P,5),"
  at the confidence level ", level,sep=""),quote=FALSE)
print(paste("Sample odds ratio equals ",round(oro,4), "; n1=",n[1],"
  n2=",n[2],sep=""),quote=FALSE)}

#Example of usage
n=c(60,70) # input  $n_A$  and  $n_B$ 
m=c(7,63) # input  $n_{A1}$  and  $n_{B1}$ 
CI(n,m,level=0.95)

```

Determinants of livestock products export in Ethiopia

Ermyas Kefelegn¹

Abstract

Ethiopia has one of the largest livestock populations in Africa. In 2016–2017, the share of live animals, leather, and meat in the total export of the country reached 9.6%.

This paper aims to identify the determinants of the export of Ethiopian livestock products by means of vector autoregressive and vector error correction models.

Multivariate time series is used to model the association between the products of the Ethiopian livestock export included in the study. Vector autoregressive and vector error correction models are used for modelling and inference.

The results indicated the existence of a long term correlation between the volume of live animals, meat and leather exports. The volume of meat export is significantly affected by a lag occurring in the export of live animals in the short-run. Therefore, 3.7% of the short-run imbalance in the volume of leather export is adjusted each quarter.

It is suggested that the exporters of livestock products should properly utilise the Ethiopian livestock resources. On the other hand, the government should offer different forms of support to exporters, especially those focusing on exporting value-added products.

Key words: livestock export, VAR, VECM.

1. Introduction

Ethiopia has one of the largest livestock populations in Africa. According to recent estimates, the country has about 57.8 million heads of cattle, 28.9 million sheep, 29.7 million goats and 47.1 million poultry, plus an assortment of horses, donkeys and camels (CSA, 2015; MOA, 2015). The economic contribution of the livestock sub-sector in Ethiopia is about 11% of the total Gross Domestic Product (GDP) and 24% of the agricultural GDP (NBE, 2016).

The government of Ethiopia encourages investments in meat processing, especially focusing on exporting value-added products abroad. In 2016/17 export earnings from leather and leather products decreased by 1.1 percent due to a 1.6 percent fall in export volume despite 0.5 percent rise in international price. Consequently, the share of leather

¹ Department of statistics, Woldia University, Ethiopia. E-mail: ermyas62@gmail.com.
ORCID: <https://orcid.org/0009-0005-4496-2089>.



& leather products in total export revenue stood at 3.9% (NBE, 2016). Formally, Ethiopia exports approximately 200,000 livestock annually. Djibouti, Egypt, Somalia, Sudan, Saudi Arabia, Yemen and United Arab Emirate are the major importers of Ethiopian live animals. The key question that this paper attempts to address is “what factors determine the volume of Ethiopian livestock products export?” so this study is designed to identify factors that determine the volume of Ethiopian livestock products export. It also identifies whether there exist an association between the volume of meat, leather and live animals export or not.

Several studies about livestock products export and related variables are done using univariate time series analysis. Univariate time series analysis is important but it is inadequate for the analysis of interaction and co-movements of several time series simultaneously. In contrast, multivariate time series (MVTs) analysis involves a vector of time series that will be modelled simultaneously. MVTs deals with the interaction, co-movements and bi-directional causality of several time series. To the best of the authors’ knowledge, little information (study) is available in multivariate time series analysis about livestock products export. So, this study is important in filling this gap.

2. Literature review

In their study, using monthly data disaggregated by markets of destination and sectors, Cho, Sheldon and McCorrison (2002) found that there is a strong negative impact of the exchange rate uncertainty on livestock trade compared to other sectors for a simple bilateral trade flows across countries.

The study by De Grawue and Bellefoid (1986), Steinherr (1989) demonstrates that the trade volume is more affected by the long run changes in the exchange rate than by the short run exchange rate fluctuations. This confirms the result obtained by Sheldon and McCorrison (2002).

A major problem with the leather sector is the by-product status of hides and skins. Cattle, goats and sheep are mainly used for meat (Aklilu, Yacob 2002). Thus, the product, i.e. hides and skins, becomes available when meat is needed, not when it is appropriate for leather processing. This shows the direct dependence of leather sector on meat.

Livestock are shipped across borders without letters of credit or pre-arranged sale contracts, with the trade being managed through cross-border clan relationships that face high transaction costs, including significant risks of confiscation, theft and disease as they transport and trade in animals (Wassie 2015). Somaliland exporters rely on their agents based in Yemen.

3. Data and methodology

3.1. Source and type of data

The study used secondary data obtained from the National Bank of Ethiopia over the period from 2002 first quarter to 2017 third quarter. The data have information about the quarterly volume of live animals, meat and leather export of Ethiopia; and the quarterly exchange rate.

The vector of endogenous (response) variables (Y_t) is the quarterly volume of Ethiopian livestock products export. Specifically, $Y_t = (y_{1t}, y_{2t}, y_{3t})'$, where y_{1t} , y_{2t} , and y_{3t} represents the export volume of live animals, meat, and leather at time (quarter) t , respectively. The lagged values of quarterly volume of Ethiopian livestock products are used as independent variables in our VAR specification together with the exogenous covariate quarterly exchange rate (birr against the US dollar).

3.2. Methodology

3.2.1. Vector Auto regressive (VAR) Models

The study used multivariate time series to model the association between Ethiopian livestock products export. VAR model is one of the most successful, flexible and easy to use models for the analysis of multivariate time series. It is a natural extension of the univariate autoregressive model to dynamic multivariate time series.

Stationarity

The first step for an appropriate analysis is to determine whether the time series under consideration are stationary or not. Many economic and financial time series exhibit

a trending behaviour or non-stationarity in the mean. Due to non-stationarity, regressions with time series data are very likely to result in spurious results. The test of stationarity that has become widely popular over the past several years, namely, the Augmented Dickey- Fuller (ADF) test due to Dickey and Fuller (1979), is used to test for the existence of unit roots.

Since there are three endogenous variables, in this study a trivariate VAR (p) model is applied for quarterly Ethiopian livestock products export. The basic p - lag vector autoregressive (VAR (p)) model with an exogenous variable (X_t) has the form:

$$Y_t = C + \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \dots + \pi_p Y_{t-p} + G X_t + \varepsilon_t, \quad t = 1, 2, \dots, T \quad (1)$$

where π_i are $(n \times n)$ coefficient matrices, $G \sim (n \times n)$ is a parameter matrix and ε_t is $(n \times 1)$ unobservable zero mean white noise vector process with time invariant covariance matrix Σ .

For example, a trivariate VAR (1) model with an exogenous variable (X_t) has the form:

$$\begin{bmatrix} Y_{1t} \\ Y_{2t} \\ Y_{3t} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} + \begin{bmatrix} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \\ \pi_{31} & \pi_{32} & \pi_{33} \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \\ Y_{3,t-1} \end{bmatrix} + \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} X_t + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix} \quad (2)$$

where $Y_t = (Y_{1t}, Y_{2t}, Y_{3t})'$ is a vector of Ethiopian livestock products export, and X_t is the exchange rate at time t . The g_i represents the effect of the current exchange rate on the contemporaneous volume of export. The diagonal elements π_{ij} of matrix π represent the effect own one-period-lagged livestock product export on the respective contemporaneous export, while the off diagonal elements π_{ij} ($i \neq j$) represent the mean effects across Ethiopian livestock products export.

Estimation of the order of the VAR model

The lag length for the VAR (p) model may be determined using model selection criteria. The general approach is to fit VAR (p) models with orders $p = 0, 1, \dots, P_{\max}$ and choose the value of p which minimizes the Akaike (AIC), Schwarz – Bayesian (BIC) and Hannan – Quinn (HQ). Model selection criteria for VAR (p) models have the form:

$$IC(p) = \ln|\Sigma_p| + C_T \cdot \varphi(n, p) \quad (3)$$

where, IC = Information Criteria, $\Sigma_p = T^{-1} \sum \varepsilon_t \varepsilon_t'$ is the residual covariance matrix from a VAR (p) model, C_T is a sequence indexed by the sample size T , and $\varphi(n, p)$ is a penalty function which penalizes large VAR (p) models.

3.2.2. Co-integration Analysis

Two sets of variables are said to be cointegrated if a linear combination of these variables has a lower order of integration. For example, cointegration exists if a set of variables, each of which is integrated of order one ($I(1)$), have linear combinations that are $I(0)$. The order of integration $I(1)$ tells us that first differences transform the non-stationary variables into stationarity series. The presence of co-integration is an evidence of a long-run equilibrium relationship between the series under consideration. In this study the Johansson (1991) procedure was applied to test for the presence of cointegration relationships.

The starting point in Johansen's procedure in determining the number of co integrating vectors is re-parameterizing the VAR representation of Y_t .

$$Y_t = \pi_1 Y_{t-1} + \pi_2 Y_{t-2} + \dots + \pi_p Y_{t-p} + DX_t + \varepsilon_t \quad (4)$$

as a vector error correction model (VECM):

$$\Delta Y_t = \pi Y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + DX_t + \varepsilon_t \quad (5)$$

where $\pi = \sum_{i=1}^p \pi_i - I_n$, $\Gamma_i = -\sum_{j=i+1}^p \pi_j$ and I_n is the identity matrix (Reinsel, 1993)

The rank of the matrix π represents the number of cointegrating vectors in the system which can be determined using the Johnson Maximum likelihood method. If the rank (π) = 0, then there are

no cointegrating vectors, and we analyse the system using VAR technique by differencing the non-stationary series. If π has full rank, i.e. rank (π) = n , then Y_t has no unit root (Y_t is stationary in level). In such cases VAR methodology is applied to the system in level. Finally, if rank (π) = r , where $0 < r < n$, then there exist r cointegrating vectors that are stationary, and the system is analysed as VECM.

3.2.3. Vector Error Correction Modelling (VECM)

The finding that many time series may contain a unit root has spurred the development of the theory of non-stationary time series analysis. A VEC model is a restricted VAR designed for use with no stationary series that are known to be co-integrated. It restricts the long-run behaviour of the endogenous variables to converge to their co-integrating relationships while allowing for short-run adjustment dynamics. The co-integration term is known as the error correction term since the deviation from long-run equilibrium is corrected gradually through a series of partial short-run adjustments.

Granger's representation theorem (Granger, 1969) asserts that if the coefficient matrix π in equation (5) has reduced rank $r < n$, then there exist $(n \times r)$ matrices α and β each with rank r such that $\pi = \alpha \beta'$, where α is matrix of speed of adjustments and β is matrix of parameters which determines the co-integrating relationships matrix of long-run coefficients.

4. Results and Discussion

4.1. Time plots

The data in this study consist of log of quarterly volume (net weight) of live animal export (LLA), leather export (LLR) and meat export (LME) in tons; and the quarterly exchange rate (birr against US dollar). The time period covered is from the first quarter of 2002 to third quarter of 2017.

The time plot of each of the series is shown in Figure 1. From the time plot we can observe that even though the volume of livestock products export highly declined at different period due to the bans imposed by importing countries as a result of

outbreaks of livestock diseases, all the series show an increasing trend over the study period.

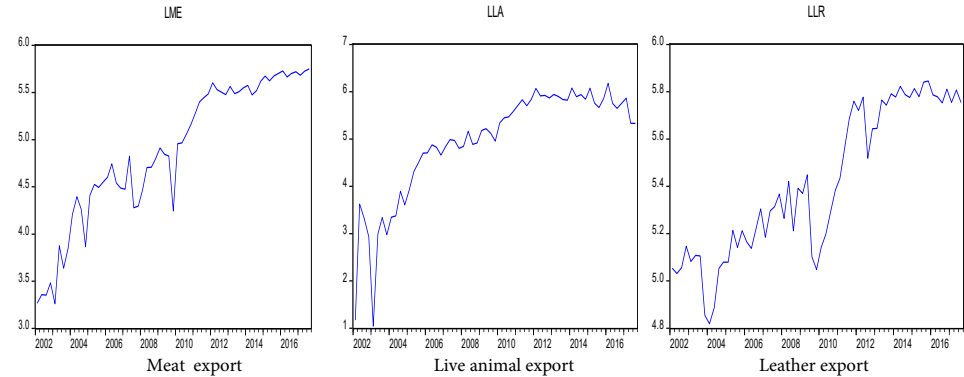


Figure 1: Time plot of the series

4.2. stationarity test

4.2.1. seasonality test

There are two reasons that might cause our data to be affected by seasonality. The first one is the quarterly nature of the data itself. Secondly, consumption of meat in Ethiopia is highly seasonal (with peaks around religious holidays), which in turn affects the export of livestock products. Therefore, before directly testing for the stationarity of the series, we have to check for the periodicity of the data. The result of seasonality test for the series is presented in Table 1.

Table 1: Seasonality test for the series

Series	F- statistic	M 7	Kruskal - Wallis statistic_(p – value)
LLA	1.942	1.955	5.776 (0.12304)
LLR	4.590	1.224	7.601 (0.05502)
LME	0.565	3.000	1.753 (0.62510)

The seasonality test results show that there is no evidence of stable seasonality for all of the series under consideration ($p\text{-value} > 0.05$). Moreover, the fact that the M7 statistics are all greater than one is further evidence that no seasonal adjustment is necessary (Lothian and Morry, 1978).

4.2.2. Unit root test

Before we attempt to fit a suitable model, we have to test for the presence of unit root(s) so that the order of integration of each series could be determined. The result of ADF test both at level and first difference for each series are presented in Tables 2.

Table 2: Unit root test results

Variables	At level		First difference	
	ADF statistics	P value	ADF statistics	P value
LLA	-2.519	0.317	-7.698	0.000
LLR	-1.635	0.766	-10.867	0.000
LME	-2.845	0.188	-8.704	0.000
LER	-0.040	0.950	-3.0257	0.038
Critical values (5%)		-2.9108	-2.9108	

* MacKinnon (1996) one-sided p-values.

The results in Table 2 indicate that the null hypothesis that the series in levels contains unit root could not be rejected for all of the four variables, while the null hypothesis is rejected for the first difference of the series. This implies that the time series under consideration are all integrated of order one (I (1)).

4.3. VAR model specification and estimation

For subsequent modelling choices, specifying the lag length has strong implications. The lag order selection results indicated that the appropriate lag length for the VAR model is one (1). Furthermore, the Wald lag exclusion test shows that the chosen lag is optimal and suitable for the data set. Accordingly, we adopt VAR (1) model for prediction and forecasting purposes.

Table 3: Vector auto regression estimates

Standard errors in (), t-statistics in [] & p-value in { }

Specification	D_LLA	D_LLRL	D_LME
D_LLA(-1)	-0.255833 *	-0.005862	-0.098513 *
	(0.11103)	(0.02736)	(0.05327)
	[-2.30422]	[-0.21425]	[-1.84922]
	{0.0173}	{0.8233}	{0.0453}
D_LLRL(-1)	-0.356147	-0.260173 *	-0.147339
	(0.52483)	(0.12934)	(0.25182)
	[-0.67860]	[-2.01154]	[-0.58510]
	{0.4798}	{0.0373}	{0.5423}
D_LME(-1)	0.238370	0.079182	-0.241890 *
	(0.27461)	(0.06768)	(0.13176)
	[0.86803]	[1.17001]	[-1.83580]
	{0.3663}	{0.2237}	{0.0471}
C	0.028549	0.013864	0.040123
	(0.06441)	(0.01587)	(0.03090)
	[0.44325]	[0.87343]	[1.29831]
	{0.6442}	{0.3633}	{0.1772}
D_LER	1.679694	-0.224150	2.512451
	(5.01018)	(1.23473)	(2.40396)
	[0.33526]	[-0.18154]	[1.04513]
	{0.7268}	{0.8500}	{0.2769}

* represents significant variables.

The volumes of live animal, leather and meat export are significantly explained by their own past volume of export. This implies that for a percent increase in one time lagged volume of live animal and leather export their volume of export is decreased by 0.255 and 0.26 percent respectively.

The volume of meat export is also explained by past volume of live animal export; a one percent increase in one-time lagged volume of live animal leads to a 0.24percent decrease in the volume of meat export. This result is in line with the findings of Gebergziaher (2015).

4.4. Co-integration analysis

Since the variables are integrated of the same order, we proceed to co integration test. The main purpose of co- integration analysis is to model the long-run relationship between the underlying variables. The results of c-integrating tests for LLA, LME and LLR are reported in Table 4.

Table 4: Johansen co-integration test results

Number of co integrating vector	Eigen value	Trace test			Maximum eigenvalue test		
		Statistic	critical value	Prob.**	Statistic	critical value	Prob **
None *(**)	0.255284	33.13406	29.79707	0.0199	28.87368	21.13162	0.0033
At most 1	0.172862	15.15419	15.49471	0.0562	11.15595	14.26460	0.1466
At most 2(**)	0.056959	3.577366	3.841466	0.0586	5.335259	3.841466	0.0209
Normalized co integrating coefficients (standard error in () and t-statistic in [])							
LLA	LLR	LME					
1.00000	5.267334	-3.909227					
	(1.43794)	(0.62645)					
	[3.66312]	[-6.24032]					

* for Trace test and (**) for maximum eigen value test.

From the Johansen co-integration test, the rank of the co-integration matrix was found to be one. In other words, there is one linear combination of the three I(1) series that is stationary, that is, there exists a long-run causal relationship among LLA, LLR, LME. A study by Gebergziaher (2015) also reports that there is a long-run association between livestock export.

The long-run model is given by:

$$LLA = 3.91 LME - 5.27 LLR - 14.39$$

The long-run equation shows that a one percent increases in the volume of meat export induces, on average, an increase of about 3.91 percent in the volume of live animals in the long-run.

4.4.1. Model estimation

Having concluded that the series under consideration are cointegrated, we proceed to estimate the short-run behaviour and the adjustment to the long-run equilibrium, which is represented by VECM. The results of the fitted VEC model are presented in Table 5 below.

Table 5: Vector error correction estimates

Standard errors in () & t-statistics in []

Error Correction:	D(LLA)	D(LLR)	D(LME)
CointEq1	-0.104701 (0.06410) [-1.63344]	-0.037182 * (0.01538) [-2.41777]	0.047054 (0.03085) [1.52542]
D(LLA(-1))	-0.201626 * (0.108564) [-1.857206]	0.013388 (0.02743) [0.48806]	-0.122875 * (0.05502) [-2.23321]
D(LLR(-1))	-0.162598 (0.53058) [-0.30645]	-0.191439 (0.12730) [-1.50386]	-0.234323 (0.25534) [-0.91770]
D(LME(-1))	-0.022891 (0.31434) [-0.07282]	-0.013599 (0.07542) [-0.18031]	-0.124475 (0.15128) [-0.82284]
C	0.044204 (0.06419) [0.68865]	0.019423 (0.01540) [1.26122]	0.033087 (0.03089) [1.07111]
D_LER	0.056511 (5.03619) [0.01122]	-0.800585 (1.20830) [-0.66257]	3.241934 (2.42362) [1.33764]

* represents significant variables

The results of the fitted VEC model show that the volume of meat export is significantly affected by lagged value of the volume of live animals export in the short-run. Furthermore, the vector error correction models show that 3.7% of the short-run disequilibrium in the volume of leather export is adjusted within one quarter, while the remaining shocks are adjusted in the subsequent quarters. On the other hand, the volume of live animals export is significantly affected by its own lagged values in the short-run.

4.4.2. Structural analysis

4.4.2.1. Forecast error variance decomposition

The decomposition is used to understand the proportion of the fluctuation in a series explained by its own shocks as well as shocks from other variables. The results of the decomposition analysis are presented in Table 6 below.

Table 6: Variance decomposition of LLA

Variance Decomposition of LLA:				
Period	S.E.	LLA	LLR	LME
1	0.835363	100.0000	0.000000	0.000000
2	1.105118	92.88874	3.624653	3.486603
3	1.259070	88.09384	5.049786	6.856373
4	1.347530	85.30358	5.599132	9.097284
5	1.397888	83.72305	5.835364	10.44158
6	1.426325	82.83517	5.947145	11.21769
7	1.442299	82.33809	6.003622	11.65829
8	1.451243	82.06030	6.033349	11.90636
9	1.456240	81.90521	6.049381	12.04541
10	1.459029	81.81868	6.058150	12.12317

At the first horizon, the variation of live animals export is explained by its own shock only. In the second quarter, shock to the volume of live animals export accounts for 93% variation of the fluctuation in live animals export (own shock) and the remaining 3.6 and 3.4 percent is explained by the volume of meat and leather exports, respectively. Then after shock to the volume of live animals, leather and meat export account approximately for 82%, 6% and 12% of the variability in the volume of live animals export, respectively.

5. Conclusions

In this empirical work, an attempt was made to apply multivariate time series analysis to model the co integration of Ethiopian livestock products export using quarterly data from 2002 to 2017. The data were tested for seasonality and results revealed that all of the series were not affected by periodicity. Moreover, unit root tests show that all four series were non-stationary in level, but stationary after first differencing.

Among all candidate VAR models, VAR (1) was found to be the best to describe the data. The co-integration test shows that there exists a long-run association between the volumes of Ethiopian live animals, leather and meat export.

The long-run equation shows that the volume of live animals export has a positive long-run relationship with the volume of meat export: for a one percent increase in the volume of meat export, the volume of live animals export is increased by 3.9 percent, in the long-run. One naturally expects an inverse relationship between the two. But the

finding that the two series drift upward together may support the fact that the Ethiopian livestock resource is underutilized.

From the fitted short-run models, 3.7% of the short-run disequilibria in the volume of leather export are adjusted each quarter. The volume of meat export is significantly affected by lagged volume of live animals export in the short-run. On the other hand, the volume of live animals export is significantly affected by its own lagged values in the short-run.

It is recommended for concerned bodies to properly utilize the Ethiopian livestock resource. It is also recommended to include more exogenous factors (such as fuel oil price and domestic price of meat).

Reference

- Aklilu, Y., (2002). *An audit of the livestock marketing status in Kenya, Ethiopia and Sudan*. 2 volumes. Prepared for AU-IBAR and PACE. Aklilu_market_vol2.pdf)
- Berihu, A., (2014). Constraints of Livestock Development in Eastern Zone of Tigray; the case of “Ganta Afeshum Woreda”, *Agricultural Science Research*, 2(1), pp. 1–9.
- Central Statistical Agency, (2015). *Central Statistics Agency Annual Report 2015/16*. Addis Ababa
- Cho, G., Sheldon, I. M. and McCorriston, S., (2002). Exchange Rate Uncertainty and Agricultural Trade. *American Journal of Agricultural Economics*, 84 (4), pp. 931–942.
- Dickey, D. A., W. A. Fuller, (1979). Distribution of the estimators for autoregressive time series with a unit-root. *Journal of the American Statistical Association*, 74, pp. 427–431.
- DeGrauwe, P., B. de Bellefoid, (1986). *Long-run Exchange Rate Variability and International Trade*. Cambridge, MA: MIT Press.
- Gebergziaher, G., Solomon, A., and Afework, H., (2019). *The impact of live animal export on meat and leather products export in Ethiopia*, unpublished.
- Granger, C. W. J., (1969). Investigating Causal Relations by Econometric Models and Cross Spectral Methods, *Econometrica*, 37, pp. 424–438.
- Johansen, S., (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, 59, pp. 1551–1580.
- Lothian, J., Morry, M., (1978). A Set of Quality Control Statistics for the X-11-ARIMA Seasonal Adjustment Method, *Research Paper*, Statistics Canada.

- MacKinnon, J. G., (1996). Numerical distribution functions for unit root and Cointegration tests. *Journal of Applied Econometrics*, 11, pp. 601–18.
- Ministry of Agriculture and Rural Development, (2015). *Major challenges and Achievements in Ethiopian Livestock production*.
- National Bank of Ethiopia, (2016). *National Bank Annual Report 2016/17*. Addis Ababa.
- Phillips, P. C. B., Perron, P., (1988). Testing for a unit-root in time series regression. *Biometrika*, 75, pp. 335–346.
- Reinsel, D. E., (1993). Vector Auto regressions and Reality. *Journal of Business and Economic Statistics*, 3, pp. 23–30.
- Sintayehu, G., Samuel, A., Derek, B. and Ayele, S., (2010). *Diagnostic study of live cattle and beef production and marketing: Constraints and opportunities for enhancing the system*. ILRI & IFPRI, Addis Ababa, Ethiopia.

On some efficient classes of estimators using auxiliary attribute

Shashi Bhushan¹, Anoop Kumar²

ABSTRACT

This paper considers some efficient classes of estimators for the estimation of population mean using known population proportion. The usual mean estimator, classical ratio, and regression estimators suggested by Naik and Gupta (1996) and Abd-Elfattah *et al.* (2010) estimators are identified as the members of the suggested class of estimators. The expressions of bias and mean square errors are derived up to first-order approximation. The proposed estimators were put to test against various other competing estimators till date. It has been found both theoretically and empirically that the suggested classes of estimators dominate the existing estimators.

Key words: Bias, Mean square error, Efficiency, Auxiliary attribute.

1. Introduction

In sample survey methodology, it is well known that the information on the auxiliary variable helps to meliorate the efficiency of the estimators. Literature comprises several improved and modified ratio, regression, product and exponential type estimators in this dimension. Some contemporary relevant studies in this direction, namely, Zaman and Kadilar (2020), Bhushan and Kumar (2020, 2022), Bhushan *et al.* (2020a, b, c, d, e), Bhushan *et al.* (2021), Zaman and Kadilar (2021a, b) can be viewed. However, many times, in real life situations, the variable of interest may be associated with some qualitative auxiliary characteristics that might be easily available. For example:

- (i). The height of the person (y) may rely on sex ϕ i.e. the person is male or female.
- (ii). The amount of yield of paddy crop (y) may rely on a certain variety of paddy (ϕ).
- (iii). The amount of milk produce (y) may depend on a certain breed of buffalo (ϕ).
- (iv). The use of drugs (y) may depend on the sex (ϕ).

Furthermore, if measuring a quantitative variable is expensive, then such an auxiliary attribute may be considered, which can be constructed from the auxiliary variable and is highly associated with the variable of interest. For example:

- (i). The yield of crop (y) may depend on large/small land holdings (ϕ).

¹Department of Statistics, University of Lucknow, Lucknow, India, 226007. E-mail: bhushan_s@lkouniv.ac.in, ORCID: <https://orcid.org/0000-0002-3931-888X>.

²Corresponding author. Department of Statistics, Amity University, Lucknow, India, 226028. E-mail: anoop.asy@gmail.com, ORCID: <https://orcid.org/0000-0003-2775-6548>.

- (ii). The tax paid by a company (y) may depend on its turn over (ϕ) which can be converted into large/small company.
- (iii). The family expenditure (y) may depend on the household size (ϕ) which can be classified large/small household.

Thus, by considering the advantage of bi-serial correlation (ρ), Naik and Gupta (1996) suggested the classical ratio, product and regression estimators using the knowledge on auxiliary attribute. Later on, Jhajj *et al.* (2006) considered a general class of estimators for population mean using auxiliary attribute. Singh *et al.* (2007) developed exponential type ratio and product estimators of population mean using auxiliary attribute. Abd-Elfattah *et al.* (2010) introduced some modified estimators of population mean based on attribute. Grover and Kaur (2011) mooted an exponential type estimators of population mean using auxiliary attribute. Singh and Solanki (2012) investigated a general class of estimators of population mean using auxiliary attribute. Koyuncu (2012) suggested an efficient estimators of population mean based on auxiliary attribute. Zaman and Toksoy (2019) suggested ratio type estimators of population mean utilizing bivariate auxiliary attribute. Following Bahl and Tuteja (1991), Zaman and Kadilar (2019) suggested attribute based exponential ratio estimator of population mean. Yadav and Zaman (2020) extended the work of Zaman and Kadilar (2019) utilizing auxiliary attribute. Zaman (2019a) suggested improved ratio estimator of population mean utilizing skewness as an auxiliary attribute. Zaman (2019b) investigated an efficient class of estimator in stratified sampling based on attribute whereas following Ozel (2016), Zaman (2020) suggested a ratio cum exponential ratio type estimator using attribute. Further, Bhushan and Gupta (2020) developed logarithmic type estimators of population mean based on attribute. Recently, Zaman (2021) addressed an efficient exponential estimator of population mean in stratified random sampling.

This paper addresses the problem of estimating the population mean using information on an auxiliary attribute. The rest of the paper is drafted in the following sections. In Section 2, we have discussed the existing estimators along with their properties. In Section 3, we have suggested some efficient classes of estimators with their properties. The theoretical conditions are derived in Section 4 whereas an empirical study is conducted in Section 5. The final conclusion is made in Section 6.

2. Existing estimators

Consider a finite population $U = (U_1, U_2, \dots, U_N)$ based on N units from which a sample s of size n is measured using the simple random sampling without replacement (SRSWOR) scheme. Let y_i and ϕ_i be the total number of units of the study variable y and the auxiliary attribute ϕ for unit i of the population U . It is to be noted that $\phi_i=1$ if the unit i owns the attribute ϕ and $\phi_i = 0$, otherwise. Let $X = \sum_{i=1}^N \phi_i$ and $x = \sum_{i=1}^n \phi_i$ be the total number of units in the population U and sample s , respectively, possessing attribute ϕ whereas $P = X/N$ and $p = x/n$, respectively, denote the population proportion and sample proportion having attribute ϕ . Let $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$ and $\bar{y} = n^{-1} \sum_{i=1}^n y_i$, respectively, be the population mean and sample mean of study variable y , $S_y^2 = (N-1)^{-1} \sum_{i=1}^N (y_i - \bar{Y})^2$ and $s_y^2 = (n-1)^{-1} \sum_{i=1}^n (y_i - \bar{y})^2$, respectively, be the population mean square and sample mean

square of study variable y , $S_\phi^2 = (N-1)^{-1} \sum_{i=1}^N (\phi_i - P)^2$ and $s_\phi^2 = (n-1)^{-1} \sum_{i=1}^n (\phi_i - p)^2$, respectively, be the population mean square and sample mean square of auxiliary attribute ϕ , $S_{y\phi} = (N-1)^{-1} (\sum_{i=1}^N y_i \phi_i - NP\bar{Y})$ and $s_{y\phi} = (n-1)^{-1} (\sum_{i=1}^n y_i \phi_i - np\bar{y})$ respectively be the population covariance and sample covariance between study variable y and attribute ϕ , $C_y = S_y/\bar{Y}$ and $C_\phi = S_\phi/P$, respectively, be the population coefficient of variation of study variable y and attribute ϕ and $\rho = S_{y\phi}/(S_y S_\phi)$ be the point biserial correlation coefficient between study variable y and attribute ϕ .

To obtain the bias and mean square error (MSE) of various estimators, let us assume that, $\bar{y} = \bar{Y}(1 + e_0)$ and $p = P(1 + e_1)$ such that $E(e_0) = E(e_1) = 0$ and $E(e_0^2) = \gamma C_y^2$, $E(e_1^2) = \gamma C_\phi^2$, $E(e_0 e_1) = \gamma \rho C_y C_\phi$, where $\gamma = (N-n)/Nn$.

The usual mean estimator is given by

$$t_m = \bar{y} \quad (2.1)$$

Naik and Gupta (1996) suggested the classical ratio and regression estimator for population mean \bar{Y} using information on the auxiliary attribute as

$$t_r = \bar{y} \left(\frac{P}{p} \right) \quad (2.2)$$

$$t_{lr} = \bar{y} + \beta_\phi (P - p) \quad (2.3)$$

where $\beta_\phi = S_{y\phi}/S_\phi^2$ is the regression coefficient of y on ϕ .

Following Hansen *et al.* (1954), Srivastava (1967), Walsh (1970) and Bhushan and Gupta (2019), one may introduce the following class of estimators based on the auxiliary attribute as

$$t_1 = \bar{y} + \theta(p^* - P^*) \quad (2.4)$$

$$t_2 = \bar{y} \left(\frac{P^*}{p^*} \right)^\Lambda \quad (2.5)$$

$$t_3 = \bar{y} \left\{ \frac{P^*}{P^* + \Delta(p^* - P^*)} \right\} \quad (2.6)$$

$$t_4 = \bar{y} \left\{ 1 + \log \left(\frac{P^*}{p^*} \right) \right\}^\beth \quad (2.7)$$

where θ , Λ , Δ and \beth are suitably chosen scalars to be determined later. Furthermore, $p^* = \eta p + \lambda$ and $P^* = \eta P + \lambda$ such that η and λ are any real values or function of some known parameters of the auxiliary attribute ϕ , namely, population standard deviation S_ϕ , population coefficient of variation C_ϕ , population coefficient of skewness $\beta_1(\phi)$, population coefficient of kurtosis $\beta_2(\phi)$ and population point biserial correlation coefficient ρ between study variable y and attribute ϕ .

Jhajj *et al.* (2006) considered the class of estimators for population mean \bar{Y} as

$$t_J = h(\bar{y}, u) \quad (2.8)$$

where $u = p/P$ and $h(\bar{y}, u)$ are the function of (\bar{y}, u) such that $h(\bar{y}, 1) = \bar{Y}$, $\forall \bar{Y}$ and the function $h(\bar{y}, u)$ satisfies certain regularity conditions as mentioned in Jhaji *et al.* (2006). On the lines of Bahl and Tuteja (1991), Singh *et al.* (2007) suggested ratio and product exponential type estimators as

$$t_{sr} = \bar{y} \exp \left(\frac{P-p}{P+p} \right) \quad (2.9)$$

$$t_{sp} = \bar{y} \exp \left(\frac{p-P}{p+P} \right) \quad (2.10)$$

Singh *et al.* (2008) envisaged a class of estimator using information on auxiliary attribute as

$$t_s = [\bar{y} + \beta_\phi(P-p)] \left(\frac{P^*}{p^*} \right) \quad (2.11)$$

Some members of the estimator t_s are given in Table 1 for ready reference.

Abd-Elfattah *et al.* (2010) also suggested the following classes of estimators for the population mean using information on auxiliary attribute as

$$t_{a1} = m_1 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{P}{p} \right) + m_2 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{P + \beta_2(\phi)}{p + \beta_2(\phi)} \right) \quad (2.12)$$

$$t_{a2} = m_1 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{P}{p} \right) + m_2 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{P + C_\phi}{p + C_\phi} \right) \quad (2.13)$$

$$t_{a3} = m_1 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{P}{p} \right) + m_2 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{\beta_2(\phi)P + C_\phi}{\beta_2(\phi)p + C_\phi} \right) \quad (2.14)$$

$$t_{a4} = m_1 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{P}{p} \right) + m_2 \{ \bar{y} + \beta_\phi(P-p) \} \left(\frac{C_\phi P + \beta_2(\phi)}{C_\phi p + \beta_2(\phi)} \right) \quad (2.15)$$

where m_1 and m_2 are weights such that $m_1 + m_2 = 1$.

Furthermore, Abd-Elfattah *et al.* (2010) envisaged the following class of estimators as

$$t_{a5} = \bar{y} \left(\frac{P + \beta_2(\phi)}{p + \beta_2(\phi)} \right) \quad (2.16)$$

$$t_{a6} = \bar{y} \left(\frac{P + C_\phi}{p + C_\phi} \right) \quad (2.17)$$

$$t_{a7} = \bar{y} \left(\frac{\beta_2(\phi)P + C_\phi}{\beta_2(\phi)p + C_\phi} \right) \quad (2.18)$$

$$t_{a8} = \bar{y} \left(\frac{C_\phi P + \beta_2(\phi)}{C_\phi p + \beta_2(\phi)} \right) \quad (2.19)$$

$$t_{a9} = \bar{y} \left(\frac{P + \rho}{p + \rho} \right) \quad (2.20)$$

Following Kadilar and Cingi (2006), one can define some more estimators like the ratio type estimators t_{a_i} , $i = 5, 6, \dots, 9$ as

$$t_{a_{10}} = \bar{y} \left(\frac{C_\phi P + \rho}{C_\phi p + \rho} \right) \quad (2.21)$$

$$t_{a_{11}} = \bar{y} \left(\frac{\rho P + C_\phi}{\rho p + C_\phi} \right) \quad (2.22)$$

$$t_{a_{12}} = \bar{y} \left(\frac{\beta_2(\phi)P + \rho}{\beta_2(\phi)p + \rho} \right) \quad (2.23)$$

$$t_{a_{13}} = \bar{y} \left(\frac{\rho P + \beta_2(\phi)}{\rho p + \beta_2(\phi)} \right) \quad (2.24)$$

$$t_{a_{14}} = \bar{y} \left(\frac{S_\phi P + \beta_2(\phi)}{S_\phi p + \beta_2(\phi)} \right) \quad (2.25)$$

Solanki and Singh (2013) suggested the exponential type estimator as

$$t_{ss} = \bar{y} \exp \left\{ \frac{\alpha(P - p)}{(P + p)} \right\} \quad (2.26)$$

where α is a suitably chosen scalar.

Zaman (2019a) introduced some improved general class of estimators based on the coefficient of skeness of auxiliary attribte as

$$t_{z_1} = \omega \bar{y} \left(\frac{P + \beta_1(\phi)}{p + \beta_1(\phi)} \right) + (1 - \omega) \bar{y} \left(\frac{\beta_2(\phi)P + \beta_1(\phi)}{\beta_2(\phi)p + \beta_1(\phi)} \right) \quad (2.27)$$

$$t_{z_2} = \omega \bar{y} \left(\frac{P + \beta_1(\phi)}{p + \beta_1(\phi)} \right) + (1 - \omega) \bar{y} \left(\frac{\beta_1(\phi)P + \beta_2(\phi)}{\beta_1(\phi)p + \beta_2(\phi)} \right) \quad (2.28)$$

$$t_{z_3} = \omega \bar{y} \left(\frac{P + \beta_1(\phi)}{p + \beta_1(\phi)} \right) + (1 - \omega) \bar{y} \left(\frac{C_\phi P + \beta_1(\phi)}{C_\phi p + \beta_1(\phi)} \right) \quad (2.29)$$

$$t_{z_4} = \omega \bar{y} \left(\frac{P + \beta_1(\phi)}{p + \beta_1(\phi)} \right) + (1 - \omega) \bar{y} \left(\frac{\beta_1(\phi)P + C_\phi}{\beta_1(\phi)p + C_\phi} \right) \quad (2.30)$$

where ω is a suitably chosen scalar.

Zaman and Kadilar (2019) suggested a family of ratio exponential estimator as

$$t_{zk} = \bar{y} \exp \left[\frac{(\eta P + \lambda) - (\eta p + \lambda)}{(\eta P + \lambda) + (\eta p + \lambda)} \right] \quad (2.31)$$

where η and λ are same as defined earlier.

Following Zaman and Kadilar (2019), Yadav and Zaman (2020) introduced a general class of estimators of population mean using the auxiliary attribute as

$$t_{yz} = k_1 \bar{y} + k_2 \bar{y} \exp \left[\frac{(\eta P + \lambda) - (\eta p + \lambda)}{(\eta P + \lambda) + (\eta p + \lambda)} \right] \quad (2.32)$$

where k_1 and k_2 are suitably chosen scalars such that $k_1 + k_2 = 1$. Furthermore, some members of the estimator t_{zk} and t_{yz} are given in Table 1 for ready reference.

Following Ozel (2016), Zaman (2020) suggested exponential ratio type estimator using auxiliary attribute as

$$t_{z5} = \bar{y} \left(\frac{p}{P} \right)^\theta \exp \left[\frac{(\eta P + \lambda) - (\eta p + \lambda)}{(\eta P + \lambda) + (\eta p + \lambda)} \right] \quad (2.33)$$

where θ is a suitably chosen scalar and η and λ are the same as defined earlier.

Bhushan and Gupta (2020) suggested the following family of estimators for the estimation of population mean \bar{Y} as

$$t_{bg} = \left[w_1 \bar{y} + w_2 \left(\frac{p}{P} \right) \right] \left[1 + \alpha \log \left(\frac{p^*}{P^*} \right) \right] \quad (2.34)$$

where w_1 , w_2 and α are suitably chosen scalars. In addition, some members of the estimator t_{bg} are given in Table 1 for ready reference.

We would like to note that the minimum MSE of the estimators t_i , $i = 1, 2, 3, 4$ envisaged on the lines of Hansen *et al.* (1954), Srivastava (1967), Walsh (1970) and Bhushan and Gupta (2019), Jhajj *et al.* (2006) estimator t_J , Abd-Elfattah *et al.* (2010) estimators t_{a_i} , $i = 1, 2, 3, 4$, Solanki and Singh (2013) estimator t_{ss} , Zaman (2019a) estimators t_{z_i} , $i = 1, 2, 3, 4$, Yadav and Zaman (2020) estimators t_{yz} and Zaman (2020) estimator t_z attain the minimum MSE of the classical regression estimator t_{lr} .

The MSE of the above estimators are given in Appendix A for quick reference.

3. Proposed Estimators

Adapting the procedure of Kadilar and Cingi (2006), we introduce the following classes of estimators by combining the difference estimator given in (2.4) with the log type, Srivastava and Walsh type estimator given in (2.7), (2.5) and (2.6) and Srivastava and Walsh type estimators given in (2.5) and (2.6).

$$t_{k1} = \zeta_1 \{ \bar{y} + \theta(p^* - P^*) \} + \psi_1 \bar{y} \left\{ 1 + \log \left(\frac{p^*}{P^*} \right) \right\}^{\beth} \quad (3.1)$$

$$t_{k2} = \zeta_2 \{ \bar{y} + \theta(p^* - P^*) \} + \psi_2 \bar{y} \left(\frac{P^*}{p^*} \right)^\Lambda \quad (3.2)$$

$$t_{k3} = \zeta_3 \{ \bar{y} + \theta(p^* - P^*) \} + \psi_3 \bar{y} \left(\frac{P^*}{P^* + \Delta(p^* - P^*)} \right) \quad (3.3)$$

$$t_{k4} = \zeta_4 \bar{y} \left(\frac{P^*}{p^*} \right)^\Lambda + \psi_4 \bar{y} \left(\frac{P^*}{P^* + \Delta(p^* - P^*)} \right) \quad (3.4)$$

where ζ_i , ψ_i , $i = 1, 2, 3, 4$, θ , Λ , Δ and \beth are suitably chosen scalars. The proposed classes of estimators are contemporary and novel in nature as they proposed to utilize the relationship

of a quantitative variable with qualitative characteristics i.e., attribute. Moreover, the proposed classes explore various linear and non-linear relationships including exponential and logarithmic relationships. Furthermore, these proposed classes of estimators are general in nature and possess various prominent classes of estimators as their special cases like linear regression estimator, logarithmic type, Srivastava, and Walsh type estimators and their linear combinations. Further, the classes of estimators t_{k_i} , $i = 1, 2, 3, 4$ are reduced into some existing estimators for different values of scalars as:

- (i). for $(\zeta_1, \psi_1, \theta, \mathfrak{I})=(1, 0, 0, 0)$; $t_{k_1} \rightarrow t_m$
- (ii). for $(\zeta_2, \psi_2, \theta, \Lambda, \eta, \lambda)=(1, 0, -\beta_\phi, 0, 1, 0)$; $t_{k_2} \rightarrow t_{lr}$
- (iii). for $(\zeta_2, \psi_2, \theta, \Lambda, \eta, \lambda)=(0, 1, 0, 1, 1, 0)$; $t_{k_2} \rightarrow t_r$
- (iv). for $(\zeta_4, \psi_4, \Lambda, \Delta, \eta, \lambda)=(1, 1, 0, 0, \eta, \lambda)$; $t_{k_4} \rightarrow t_{a_i}$, $i = 5, 6, \dots, 13$

Several other estimators can be generated for different values of scalars. Some members of the proposed classes of estimators are given in Table 1.

Theorem 3.1. *The bias and MSE of the suggested classes of estimators t_{k_i} , $i = 1, 2, 3, 4$ are given as*

$$Bias(t_{k_i}) = \bar{Y} \left[\zeta_i I_i + \psi_i J_i - 1 \right]$$

(3.5)

$$MSE(t_{k_i}) = \bar{Y}^2 \left[1 + \zeta_i^2 F_i + \psi_i^2 G_i + 2\zeta_i \psi_i H_i - 2\zeta_i I_i - 2\psi_i J_i \right]$$

(3.6)

Proof. Refer to Appendix B for the outline of the derivation and definitions of parametric functions ζ_i , ψ_i , F_i , G_i , H_i , I_i and J_i . □

Corollary 3.1. *The minimum MSE of the suggested classes of estimators t_{k_i} , $i = 1, 2, 3, 4$ are given as*

$$minMSE(t_{k_i}) = \bar{Y}^2 \left[1 - \frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \right]$$

(3.7)

Proof. Refer to Appendix B. □

Table 1: Some members of the existing and proposed classes of estimators

Value of		Class of estimators				
η	λ	Singh <i>et al.</i> (2008) estimators t_{s_j}	Zaman and Kadilar (2019) estimators t_{zk_j}	Yadav and Zaman (2020) estimators t_{yz_j}	Bhushan and Gupta (2020) estimators t_{bg_j}	Proposed estimators $t_{k(i)}$, $i = 1, 2, 3, 4$
1	$\beta_2(\phi)$	t_{s1}	t_{zk1}	t_{yz1}	t_{bg1}	$t_{k(1)}$
1	C_ϕ	t_{s2}	t_{zk2}	t_{yz2}	t_{bg2}	$t_{k(2)}$
$\beta_2(\phi)$	C_ϕ	t_{s3}	t_{zk3}	t_{yz3}	t_{bg3}	$t_{k(3)}$
C_ϕ	$\beta_2(\phi)$	t_{s4}	t_{zk4}	t_{yz4}	t_{bg4}	$t_{k(4)}$
1	ρ	t_{s5}	t_{zk5}	t_{yz5}	t_{bg5}	$t_{k(5)}$
C_ϕ	ρ	t_{s6}	t_{zk6}	t_{yz6}	t_{bg6}	$t_{k(6)}$
ρ	C_ϕ	t_{s7}	t_{zk7}	t_{yz7}	t_{bg7}	$t_{k(7)}$
$\beta_2(\phi)$	ρ	t_{s8}	t_{zk8}	t_{yz8}	t_{bg8}	$t_{k(8)}$
ρ	$\beta_2(\phi)$	t_{s9}	t_{zk9}	t_{yz9}	t_{bg9}	$t_{k(9)}$
S_ϕ	$\beta_2(\phi)$	t_{s10}	t_{zk10}	t_{yz10}	t_{bg10}	$t_{k(10)}$

4. Efficiency Conditions

On comparing the minimum MSE of the proposed classes of estimators t_{k_i} , $i = 1, 2, 3, 4$ given in (3.7) with the minimum MSE of the of existing estimators given in (A.1), (A.2), (A.3), (A.4), (A.5), (A.6), (A.7), (A.9) and (A.10), we get the following conditions.

$$MSE(t_m) \geq MSE(t_{k_i})$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - \gamma C_y^2 \quad (4.1)$$

$$MSE(t_r) \geq MSE(t_{k_i})$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - \gamma [C_y^2 + C_\phi^2 - 2\rho C_y C_\phi] \quad (4.2)$$

$$MSE(t) \geq MSE(t_{k_i}); t = t_{lr}, t_J, t_i, t_{a_i}, i = 1, 2, 3, 4, t_{ss}, t_{z_i}, i = 1, 2, \dots, 5 \text{ and } t_{yz}$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - [\gamma C_y^2 (1 - \rho^2)] \quad (4.3)$$

$$MSE(t_{sr}) \geq MSE(t_{k_i})$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - \gamma \left[C_y^2 + \frac{C_\phi^2}{4} - \rho C_y C_\phi \right] \quad (4.4)$$

$$MSE(t_{sp}) \geq MSE(t_{k_i})$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - \gamma \left[C_y^2 + \frac{C_\phi^2}{4} + \rho C_y C_\phi \right] \quad (4.5)$$

$$MSE(t_s) \geq MSE(t_{k_i})$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - \gamma [v^2 C_\phi^2 + C_y^2 (1 - \rho^2)] \quad (4.6)$$

$$MSE(t_{a_i}) \geq MSE(t_{k_i}), i = 1, 2, 3, 4$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - \gamma [C_y^2 + v^2 C_\phi^2 - 2v\rho C_y C_\phi] \quad (4.7)$$

$$MSE(t_{zc}) \geq MSE(t_{k_i})$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq 1 - \gamma [\lambda^2 C_\phi^2 + C_y^2 - 2\lambda\rho C_y C_\phi] \quad (4.8)$$

$$MSE(t_{bg}) \geq MSE(t_{k_i})$$

$$\frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \geq \frac{(AG^2 + BD^2 - 2DFG)}{(4AB - F^2)} \quad (4.9)$$

Under the conditions of (4.1) to (4.9), the proposed classes of estimators t_{k_i} , $i = 1, 2, 3, 4$ dominate the usual mean estimator, classical ratio, product and regression estimator, Jhaji *et al.* (2006) estimator, Singh *et al.* (2007) estimator, Singh *et al.* (2008) estimator, Abd-Elfattah *et al.* (2010) estimators, Solanki and Singh (2013) estimator, Zaman and Kadilar

(2019) estimator, Zaman (2019a, 2020) estimators, Yadav and Zaman (2020) estimators and Bhushan and Gupta (2020) estimator. Further, these conditions are supported with an empirical study using three different populations.

5. Empirical Study

To have clear idea about the properties of the proposed estimators, we consider an empirical study over two real populations. The description of the populations is given below:

Population 1. (Source: Sukhatme and Sukhatme (1970), pp. 256)

y: Number of villages in the circles, ϕ : A circle consisting of more than five villages, $N=89$, $n=23$, $\bar{Y}=3.36$, $P=0.124$, $C_y=0.601$, $C_\phi=2.678$, $\rho=0.766$ and $\beta_2(\phi)=3.492$.

Population 2. (Source: Sukhatme and Sukhatme (1970), pp. 256)

y: Area (in acres) under the wheat crop within the circles, ϕ : A circle consisting of more than five villages, $N=89$, $n=23$, $\bar{Y}=1102$, $P=0.124$, $C_y=0.65$, $C_\phi=2.678$, $\rho=0.624$ and $\beta_2(\phi)=3.492$.

Population 3. (Source: Zaman *et al.* (2014))

y: The number of teachers, ϕ : The number of teachers is more than 60, $N=111$, $n=30$, $\bar{Y}=29.279$, $P=0.117$, $C_y=0.872$, $C_\phi=2.758$, $\rho=0.797$ and $\beta_2(\phi)=3.898$.

For the above populations, we have calculated the percent relative efficiency (*PRE*) of various estimators T with respect to the usual mean estimator using the following expression.

$$PRE = \frac{MSE(T)}{MSE(t_m)} \times 100 \quad (5.1)$$

The results of the numerical study for the above populations are disclosed in Table 2. It has been seen from Table 2 that the members $t_{k_{i(j)}}$, $i=1,2,3,4$; $j=1,2,\dots,10$, of the suggested classes of estimators t_{k_i} are superior than:

- (i). the usual mean estimator t_m , classical ratio estimator t_r and regression estimator t_{lr} envisaged by Naik and Gupta (1996), Jhajj (2006) estimator t_J , Abd-Elfattah *et al.* (2010) estimators t_{a_i} , $i = 1, 2, 3, 9$, ratio type estimators t_{a_i} , $i = 10, 11, 12, 13, 14$ defined on the lines of Kadilar and Cingi (2006), ratio and product exponential estimators t_{sr} , t_{sp} suggested by Singh *et al.* (2007), Solanki and Singh (2013) estimator t_{ss} , Zaman (2019a) estimators t_{z_i} , $i = 1, 2, 3, 4$ and Zaman (2020) estimator t_z .
- (ii). the members t_{s_j} , $j=1$ to 10; of the class of estimators t_s suggested by Singh *et al.* (2008).
- (iii). the members t_{zk_j} , $j=1$ to 10; of the class of estimators t_{zk} introduced by Zaman and Kadilar (2019).
- (iv). the members t_{yz_j} , $j=1$ to 10; of the class of estimators t_{yz} introduced by Yadav and Zaman (2020).
- (v). the members t_{bg_j} , $j=1$ to 10; of the class of estimators t_{bg} investigated by Bhushan and Gupta (2020).

On comparing the findings of Table 2, we have seen that the *PRE* of the members $t_{k_{i(j)}}$, $i=1,2,3,4$; $j=1,2,\dots,10$, of the suggested classes of estimators dominate the estimators discussed in the earlier section. Moreover, we have also observed that the estimator t_{k_1} consisting of the information $(\beta_2(\phi), \rho)$ is the most efficient among the suggested classes of estimators in each population.

6. Conclusion

In this paper, we have introduced various novel classes of estimators by combining difference estimator with Srivastava, Walsh and Log type estimators and Srivastava type estimator with Walsh type estimator for estimating the population mean of study variable utilizing the information on an auxiliary attribute and compared them with the relevant contemporary estimators till date. It is important to consider various classes of estimators in a single study so that their relative efficiencies can be compared to get a better understanding regarding the performance of such estimators with the existing estimators. The bias and *MSE* of these estimators are derived up to the first order of approximation. The efficiency conditions have been obtained under which the proposed estimators dominate various estimators available till date. These efficiency conditions are further verified by an empirical study using three real populations. The empirical results show the dominance of the proposed classes of estimators over the usual mean estimator, classical ratio, regression and difference estimators, Srivastava (1967) type estimator, Walsh (1970) type estimator, Jhaji *et al.* (2006) estimator, Singh *et al.* (2007) estimator, Singh *et al.* (2008) estimator, Abd-Elfattah *et al.* (2010) estimator, Solanki and Singh (2013) estimator, Log type estimator envisaged on the lines of Bhushan and Gupta (2016), Zaman and Kadilar (2019) estimator, Yadav and Zaman (2020) estimators, Zaman (2019a, 2020) estimators and Bhushan and Gupta (2020) estimator. The empirical results also show that the estimator t_{k_1} based on the information $(\beta_2(\phi), \rho)$ is found to be the most efficient among the proposed classes of estimators in each population. Thus, the proposed classes of estimators are enthusiastically recommended for the estimation of population mean when information is available in the form of auxiliary attribute.

Moreover, the proposed classes of estimators can also be developed in stratified sampling based on auxiliary attribute and it is the authors research work in forthcoming studies.

Acknowledgments

The authors are thankful to the reviewers and the editor for their valuable comments and suggestions that led to improving the article.

Table 2: *PRE* of different estimators with respect to t_m

Populations				Populations			
Estimators	1	2	3	Estimators	1	2	3
t_r	7.100	7.792	16.772	t_{bg9}	144.303	116.249	146.009
t	241.987	163.766	274.129	t_{bg10}	158.350	120.130	156.516
t_{sr}	39.207	37.415	102.029	$t_{k1(1)}$	242.987	165.043	275.675
t_{sp}	10.664	12.796	16.606	$t_{k1(2)}$	242.959	165.030	275.590
t_{s1}	229.046	158.582	267.891	$t_{k1(3)}$	243.084	165.125	275.596
t_{s2}	221.175	155.311	262.220	$t_{k1(4)}$	242.986	165.053	275.574
t_{s3}	125.330	69.105	106.443	$t_{k1(5)}$	243.057	165.218	275.406
t_{s4}	177.630	135.387	236.259	$t_{k1(6)}$	244.550	166.421	276.697
t_{s5}	125.208	92.841	189.138	$t_{k1(7)}$	242.964	165.033	275.610
t_{s6}	44.861	37.616	83.637	$t_{k1(8)}$	245.630	167.153	278.057
t_{s7}	229.074	160.253	266.318	$t_{k1(9)}$	242.995	165.050	275.695
t_{s8}	33.302	28.908	59.173	$t_{k1(10)}$	243.018	165.058	275.751
t_{s9}	234.100	161.655	270.086	$t_{k2(1)}$	243.377	165.243	276.367
t_{s10}	240.418	163.150	273.439	$t_{k2(2)}$	243.437	165.273	276.528
t_{a5}	114.539	110.475	116.065	$t_{k2(3)}$	243.598	165.334	277.041
t_{a6}	135.724	124.115	123.244	$t_{k2(4)}$	243.497	165.296	276.642
t_{a7}	215.966	143.873	205.746	$t_{k2(5)}$	243.965	165.524	277.833
t_{a8}	142.278	127.965	148.582	$t_{k2(6)}$	243.603	165.155	277.953
t_{a9}	230.244	162.838	192.843	$t_{k2(7)}$	243.412	165.254	276.480
t_{a10}	133.081	79.247	264.617	$t_{k2(8)}$	243.386	164.980	277.779
t_{a11}	126.667	115.075	118.228	$t_{k2(9)}$	243.356	165.226	276.332
t_{a12}	39.332	30.183	203.993	$t_{k2(10)}$	243.313	165.211	276.245
t_{a13}	110.994	106.513	112.654	$t_{k3(1)}$	242.950	165.064	275.642
t_{a14}	104.650	103.453	104.980	$t_{k3(2)}$	242.896	165.020	275.527
t_{zk1}	112.5124	109.071	107.681	$t_{k3(3)}$	242.713	164.883	275.150
t_{zk2}	116.458	111.789	110.919	$t_{k3(4)}$	242.829	164.952	275.425
t_{zk3}	161.007	138.049	144.370	$t_{k3(5)}$	242.558	164.748	274.924
t_{zk4}	134.946	123.647	121.670	$t_{k3(6)}$	242.449	164.692	274.709
t_{zk5}	161.081	144.836	139.218	$t_{k3(7)}$	242.920	165.013	275.564
t_{zk6}	237.106	161.551	207.797	$t_{k3(8)}$	242.433	164.688	274.663
t_{zk7}	112.497	107.393	108.668	$t_{k3(9)}$	242.970	165.038	275.671
t_{zk8}	241.118	151.750	241.650	$t_{k3(10)}$	243.009	165.052	275.743
t_{zk9}	109.517	105.678	106.104	$t_{k4(1)}$	242.123	163.990	275.007
t_{zk10}	104.069	103.027	102.453	$t_{k4(2)}$	242.104	164.003	274.944
t_{bg1}	138.1014	111.875	142.173	$t_{k4(3)}$	242.000	164.136	274.500
t_{bg2}	131.138	108.767	135.331	$t_{k4(4)}$	242.043	164.061	274.765
t_{bg3}	93.983	89.283	102.100	$t_{k4(5)}$	241.999	164.176	274.550
t_{bg4}	109.671	98.167	119.360	$t_{k4(6)}$	242.044	164.462	274.173
t_{bg5}	93.951	85.960	104.900	$t_{k4(7)}$	242.122	163.982	274.987
t_{bg6}	74.142	75.212	87.121	$t_{k4(8)}$	242.110	164.559	274.129
t_{bg7}	138.130	113.963	139.952	$t_{k4(9)}$	242.138	163.974	275.040
t_{bg8}	71.899	74.593	85.370	$t_{k4(10)}$	242.169	163.961	275.121

where $t = t_{lr}, t_i, t_{a_i}, i = 1, 2, 3, 4, t_j, t_{ss}, t_{z_i}, i = 1, 2, \dots, 5$ and $t_{y_{z_i}}, i = 1, 2, \dots, 10$

References

- Abd-Elfattah, A.M., El-Sherpieny, E.A., Mohamed, S.M., Abdou, O.F., (2010). Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute. *Applied Mathematics and Computation*, 215, pp. 4198–4202.
- Bahl, S., Tuteja, R.K., (1991). Ratio and product type exponential estimator. *Information and Optimization Sciences*, 12, pp. 159–163.
- Bhushan, S., Gupta, R., (2019). Some log-type classes of estimators using auxiliary attribute. *Advances in Computational Sciences and Technology*, 12(2), pp. 99–108.
- Bhushan, S., Gupta, R., (2020). An improved log-type family of estimators using attribute. *Journal of Statistics and Management Systems*, 23(3), pp. 593–602.
- Bhushan, S., Kumar, A., (2020). Log type estimators of population mean under ranked set sampling. *Predictive Analytics using Statistics and Big Data: Concepts and Modeling*, 28, pp. 47–74.
- Bhushan, S., Kumar, A., (2022). On optimal classes of estimators under ranked set sampling. *Communications in Statistics - Theory and Methods*, 51(8), pp. 2610–2639.
- Bhushan, S., Gupta, R., Singh, S., Kumar, A., (2020a). A modified class of log-type estimators for population mean using auxiliary information on variables. *International Journal of Applied Engineering Research*, 15(6), pp. 612–627.
- Bhushan, S., Gupta, R., Singh, S., Kumar, A., (2020b). Some improved classes of estimators using auxiliary information. *International Journal for Research in Applied Science & Engineering Technology*, 8(VI), pp. 1088–1098.
- Bhushan, S., Gupta, R., Singh, S., Kumar, A., (2020c). A new efficient log-type class of estimators using auxiliary variable. *International Journal of Statistics and Systems*, 15(1), pp. 19–28.
- Bhushan, S., Gupta, R., Singh, S., Kumar, A., (2020d). Some new improved classes of estimators using multiple auxiliary information. *Global Journal of Pure and Applied Mathematics*, 16(3), pp. 515–528.
- Bhushan, S., Gupta, R., Singh, S., Kumar, A., (2020e). Some log-type classes of estimators using multiple auxiliary information. *International Journal of Scientific Engineering and Research*, 8(6), pp. 12–17.
- Bhushan, S., Kumar, A., Singh, S., (2021). Some efficient classes of estimators under stratified sampling. *Communications in Statistics - Theory and Methods*, pp. 1–30. DOI:<https://doi.org/10.1080/03610926.2021.1939052>.
- Grover, L. K., Kaur, P., (2011). An improved exponential estimator of finite population mean in simple random sampling using an auxiliary attribute. *Applied Mathematics and Computation*, 218(7), pp. 3093–3099.
- Hansen, M. H. Hurwitz, W. N., Madow, W. G., (1953). *Sample Survey Methods and Theory*. John Wiley and Sons, New York, U.S.A.
- Jhaji, H. S., Sharma, M. K., Grover, L. K., (2006). A family of estimators of population mean using information on auxiliary attribute. *Pakistan Journal of Statistics*, 22, pp. 43–50.
- Kadilar, C., Cingi, H., (2006). Improvement in estimating the population mean in simple random sampling. *Applied Mathematics Letters*, 19, pp. 75–79.

- Koyuncu, N., (2012). Efficient estimators of population mean using auxiliary attributes. *Applied Mathematics and Computation*, 218(22), pp. 10900–10905.
- Naik, V. D., Gupta, P.C., (1996). A note on estimation of mean with known population proportion of an auxiliary character. *Journal of Indian Society of Agricultural Statistics*, 48(2), pp. 151–158.
- Ozel, K. G., (2016). A new exponential type estimator for the population mean in simple random sampling. *Journal of Modern Applied Statistical Methods*, 15(2), pp. 207–214.
- Solanki, R. S., Singh, H. P., (2013). Improved estimation of population mean using population proportion of an auxiliary character. *Chilean Journal of Statistics*, 4(1), pp. 3–17.
- Singh, R., Chouhan, P., Sawan, N., Smarandache, F., (2007). Ratio-product type exponential estimator for estimating finite population mean using information on auxiliary attribute. *Renaissance High Press, USA*. pp. 18–32.
- Singh, R., Chauhan, P., Sawan, N., Smarandache, F., (2008). Ratio estimators in simple random sampling using information on auxiliary attribute. *Pakistan Journal of Statistics and Operation Research*, IV(1), pp. 47–53.
- Singh, H. P., Solanki, R. S., (2012). Improved estimation of population mean in simple random sampling using information on auxiliary attribute. *Applied Mathematics and Computation*, 218(15), pp. 7798–7812.
- Srivastava, S. K., (1967). An estimator using auxiliary information. *Calcutta Statistical Association Bulletin*, 16, pp. 121–132.
- Sukhatme, P. V., Sukhatme, B. V., (1970). *Sampling Theory of Surveys with Applications*. Iowa State University Press, Ames, U.S.A.
- Walsh, J. E., (1970). Generalization of ratio estimator for population total. *Sankhya, A*, 32, pp. 99–106.
- Yadav, S. K., Zaman, T., (2020). A general exponential family of estimators for population mean using auxiliary attribute. *Journal of Science and Arts*, 20(1), pp. 25–34.
- Zaman, T., Saglam, V., Sagir, M., Yucesoy, E., Zobu, M., (2014). Investigation of some estimators via Taylor series approach and an application. *American Journal of Theoretical and Applied Statistics*, 3(5), pp. 141–147.
- Zaman, T., Kadilar, C., (2019). Novel family of exponential estimators using information of auxiliary attribute. *Journal of Statistics and Management Systems*, 22(8), pp. 1499–1509.
- Zaman, T., Toksoy, E., (2019). Improvement in estimating the population mean in simple random sampling using information on two auxiliary attributes and numerical application in agricultural engineering. *Fresenius Environmental Bulletin*, 28(6), pp. 4584–4590.
- Zaman, T., (2019a). Improved estimators using coefficient of skewness of auxiliary attribute. *Journal of Reliability and Statistical Studies*, pp. 177–186.
- Zaman, T., (2019b). Efficient estimators of population mean using auxiliary attribute in stratified random sampling. *Advances and Applications in Statistics*, 56(2), pp. 153–171.

Zaman, T., (2020). Generalized exponential estimators for the finite population mean. *Statistics in Transition - new series*, 21(1), pp. 159–168.

Zaman, T., Kadilar, C., (2020). On estimating the population mean using auxiliary character in stratified random sampling. *Journal of Statistics and Management Systems*, 23(8), pp. 1415–1426.

Zaman, T., (2021). An efficient exponential estimator of the mean under stratified random sampling. *Mathematical Population Studies*, 28(2), pp. 104–121.

Zaman, T., Kadilar, C., (2021a). Exponential ratio and product type estimators of the mean in stratified two-phase sampling. *AIMS Mathematics*, 6(5), pp. 4265–4279.

Zaman, T., Kadilar, C., (2021b). New class of exponential estimators for finite population mean in two-phase sampling. *Communications in Statistics-Theory and Methods*, 50(4), pp. 874–889.

Appendix A

To the first degree of approximation, the *MSE* of the estimators reviewed in Section 2 are respectively given below.

$$MSE(t_m) = \gamma \bar{Y}^2 C_y^2 \quad (\text{A.1})$$

$$MSE(t_r) = \gamma \bar{Y}^2 [C_y^2 + C_\phi^2 - 2\rho C_y C_\phi] \quad (\text{A.2})$$

$$\min MSE(t) = \bar{Y}^2 \gamma C_y^2 (1 - \rho^2); t = t_{lr}, t_i, t_{a_i}, i = 1, 2, 3, 4, t_J, t_{ss}, t_{z_i}, i = 1, 2, \dots, 5, t_{yz} \quad (\text{A.3})$$

$$MSE(t_{sr}) = \gamma \bar{Y}^2 \left[C_y^2 + \frac{C_\phi^2}{4} - \rho C_y C_\phi \right] \quad (\text{A.4})$$

$$MSE(t_{sp}) = \gamma \bar{Y}^2 \left[C_y^2 + \frac{C_\phi^2}{4} + \rho C_y C_\phi \right] \quad (\text{A.5})$$

$$MSE(t_s) = \gamma \bar{Y}^2 [v^2 C_\phi^2 + C_y^2 (1 - \rho^2)] \quad (\text{A.6})$$

$$MSE(t_{a_i}) = \gamma \bar{Y}^2 [C_y^2 + v^2 C_\phi^2 - 2v\rho C_y C_\phi], i = 5, 6, \dots, 13 \quad (\text{A.7})$$

$$MSE(t_{ss}) = \gamma \bar{Y}^2 \left[C_y^2 + \frac{\alpha C_\phi^2}{4} \left(\alpha - 4\rho \frac{C_y}{C_\phi} \right) \right] \quad (\text{A.8})$$

$$MSE(t_{zk}) = \gamma \bar{Y}^2 [\lambda^2 C_\phi^2 + C_y^2 - 2\lambda\rho C_y C_\phi] \quad (\text{A.9})$$

$$\min MSE(t_{bg}) = \bar{Y}^2 \left[1 - \frac{(AG^2 + BD^2 - 2DFG)}{(4AB - F^2)} \right] \quad (\text{A.10})$$

The optimum values of scalars involved in the estimators t_{lr} , t_i , $i = 1, 2, 3, 4$, t_{ss} , t_{zi} , $i = 1, 2, \dots, 5$, t_{yz} and t_{bg} are tabulated below:

$$\beta_{\phi(opt)} = \rho \frac{C_y}{C_\phi} \quad (\text{A.11})$$

$$\theta_{(opt)} = -\rho \frac{C_y}{C_\phi} = \Lambda_{(opt)} = \Delta_{(opt)} = \mathfrak{I}_{(opt)} \quad (\text{A.12})$$

$$\alpha_{(opt)} = 2\rho \frac{C_y}{C_\phi} \quad (\text{A.13})$$

$$\omega_{(opt)} = \frac{(\rho C_y - C_\phi \zeta_i)}{C_\phi (\zeta_1 - \zeta_i)}, \quad i = 1, 2, 3, 4 \quad (\text{A.14})$$

$$k_{1(opt)} = 1 - \rho \frac{C_y}{v C_\phi} \quad (\text{A.15})$$

$$k_{2(opt)} = \rho \frac{C_y}{v C_\phi} \quad (\text{A.16})$$

$$\theta_{(opt)} = v - \rho \frac{C_y}{C_\phi} \quad (\text{A.17})$$

$$w_{1(opt)} = \frac{(GF - 2BD)}{(4AB - F^2)} \quad (\text{A.18})$$

$$w_{2(opt)} = \bar{Y} \frac{(DF - 2GA)}{(4AB - F^2)} \quad (\text{A.19})$$

where $v = \eta P / (\eta P + \lambda)$, $\zeta_1 = P / (P + \beta_1(\phi))$, $\zeta_2 = \beta_2(\phi)P / (\beta_2(\phi)P + \beta_1(\phi))$, $\zeta_3 = C_\phi P / (C_\phi P + \beta_1(\phi))$, $\zeta_4 = \beta_1(\phi)P / (\beta_1(\phi)P + C_\phi)$, $A = 1 + \gamma(C_y^2 + \alpha^2 v^2 C_\phi^2 + 4\alpha v \rho C_y C_\phi - \alpha v C_\phi^2)$, $B = 1 + \gamma(C_\phi^2 + \alpha^2 v^2 C_\phi^2 - \alpha v^2 C_\phi^2 + 4\alpha v C_\phi^2)$, $D = \gamma(\alpha v^2 C_\phi^2 - 2\alpha v \rho C_y C_\phi) - 2$, $G = \gamma(\alpha v^2 C_\phi^2 - 2\alpha v C_\phi^2) - 2$, $F = 2 + 2\gamma(2\alpha v C_\phi^2 + 2\alpha v \rho C_y C_\phi + \rho C_y C_\phi - \alpha v^2 C_\phi^2 + \alpha^2 v^2 C_\phi^2)$.

Appendix B

Consider the first estimator

$$t_{k_1} = \zeta_1 \{\bar{y} + \theta(p^* - P^*)\} + \psi_1 \bar{y} \left\{ 1 + \log \left(\frac{P^*}{P^*} \right) \right\}^{\mathfrak{I}} \quad (\text{B.20})$$

Now, using the notations defined in the earlier section, we express this estimator in terms of e' 's as

$$t_{k_1} - \bar{Y} = \bar{Y} \left[\zeta_1 \left(1 + e_0 + \frac{\theta}{R} \eta e_1 \right) + \psi_1 \left(1 + e_0 + \mathfrak{I} v e_1 - \mathfrak{I} v^2 e_1^2 + \frac{\mathfrak{I}^2}{2} v^2 e_1^2 + \mathfrak{I} v e_0 e_1 \right) \right] \quad (\text{B.21})$$

Taking expectation both sides of Equation (B.21), we get the bias of the estimator t_{k_1} up to the first order of approximation as

$$\text{Bias}(t_{k_1}) = \bar{Y} [\zeta_1 I_1 + \psi_1 J_1 - 1] \quad (\text{B.22})$$

Now, squaring and taking expectation both sides of Equation (B.21), we get the MSE of the estimator up to the first order of approximation as

$$MSE(t_{k_1}) = \bar{Y}^2 [1 + \zeta_1^2 F_1 + \psi_1^2 G_1 + 2\zeta_1 \psi_1 H_1 - 2\zeta_1 I_1 - 2\psi_1 J_1] \quad (\text{B.23})$$

The MSE of the estimator t_{k_1} is minimized by

$$\zeta_{1(opt)} = \frac{(G_1 I_1 - H_1 J_1)}{(F_1 G_1 - H_1^2)} \quad \text{and} \quad \psi_{1(opt)} = \frac{(F_1 J_1 - H_1 I_1)}{(F_1 G_1 - H_1^2)} \quad (\text{B.24})$$

The minimum MSE at $\zeta_{1(opt)}$ and $\psi_{1(opt)}$ is given by

$$\min MSE(t_{k_1}) = \bar{Y}^2 \left[1 - \frac{(F_1 J_1^2 + G_1 I_1^2 - 2H_1 I_1 J_1)}{(F_1 G_1 - H_1^2)} \right] \quad (\text{B.25})$$

Similarly, the MSE of the other estimators can be obtained. In general, we can write

$$MSE(t_{k_i}) = \bar{Y}^2 [1 + \zeta_i^2 F_i + \psi_i^2 G_i + 2\zeta_i \psi_i H_i - 2\zeta_i I_i - 2\psi_i J_i] \quad (\text{B.26})$$

The $MSE(t_{k_i})$ is minimized for

$$\zeta_{i(opt)} = \frac{(G_i I_i - H_i J_i)}{(F_i G_i - H_i^2)} \quad \text{and} \quad \psi_{i(opt)} = \frac{(F_i J_i - H_i I_i)}{(F_i G_i - H_i^2)} \quad (\text{B.27})$$

The minimum MSE at the optimum values of the above scalars is given as

$$\min MSE(t_{k_i}) = \bar{Y}^2 \left[1 - \frac{(F_i J_i^2 + G_i I_i^2 - 2H_i I_i J_i)}{(F_i G_i - H_i^2)} \right] \quad (\text{B.28})$$

where

$$\begin{aligned} F_1 &= \left[1 + \gamma \left\{ C_y^2 + \left(\frac{\theta}{R} \right)^2 \eta^2 C_\phi^2 + 2 \left(\frac{\theta}{R} \right) \eta \rho C_y C_\phi \right\} \right] \\ G_1 &= \left[1 + \gamma \{ C_y^2 + (2\mathfrak{I}^2 - 2\mathfrak{I}) v^2 C_\phi^2 + 4\mathfrak{I} v \rho C_y C_\phi \} \right] \\ H_1 &= \left[1 + \gamma \left\{ C_y^2 + \left(\frac{\mathfrak{I}^2}{2} v^2 - \mathfrak{I} v^2 + \frac{\theta \mathfrak{I}}{R} \eta v \right) C_\phi^2 + \left(\frac{\theta}{R} \eta + 2\mathfrak{I} v \right) \rho C_y C_\phi \right\} \right] \\ I_1 &= 1 \\ J_1 &= \left[1 + \gamma \left\{ \left(\frac{\mathfrak{I}^2}{2} - \mathfrak{I} \right) v^2 C_\phi^2 + \mathfrak{I} v \rho C_y C_\phi \right\} \right] \end{aligned}$$

$$F_2 = \left[1 + \gamma \left\{ C_y^2 + \left(\frac{\theta}{R} \right)^2 \eta^2 C_\phi^2 + 2 \left(\frac{\theta}{R} \right) \eta \rho C_y C_\phi \right\} \right]$$

$$G_2 = [1 + \gamma \{ C_y^2 + (2\Lambda^2 v^2 + \Lambda v^2) C_\phi^2 - 4\Lambda v \rho C_y C_\phi \}]$$

$$H_2 = \left[1 + \gamma \left\{ C_y^2 + \left(\frac{\Lambda(\Lambda+1)}{2} v^2 - \frac{\theta\Lambda}{R} \eta v \right) C_\phi^2 + \left(\frac{\theta}{R} \eta - 2\Lambda v \right) \rho C_y C_\phi \right\} \right]$$

$$I_2 = 1$$

$$J_2 = \left[1 + \gamma \left\{ \frac{\Lambda(\Lambda+1)}{2} v^2 C_\phi^2 - \Lambda v \rho C_y C_\phi \right\} \right]$$

$$F_3 = \left[1 + \gamma \left\{ C_y^2 + \left(\frac{\theta}{R} \right)^2 \eta^2 C_\phi^2 + 2 \left(\frac{\theta}{R} \right) \eta \rho C_y C_\phi \right\} \right]$$

$$G_3 = [1 + \gamma \{ C_y^2 + 3\Delta^2 v^2 C_\phi^2 - 4\Delta v \rho C_y C_\phi \}]$$

$$H_3 = \left[1 + \gamma \left\{ C_y^2 + \left(\Delta^2 v^2 - \frac{\theta\Delta}{R} \eta v \right) C_\phi^2 + \left(\frac{\theta}{R} \eta - 2\Delta v \right) \rho C_y C_\phi \right\} \right]$$

$$I_3 = 1$$

$$J_3 = [1 - \gamma \{ \Delta v \rho C_y C_\phi - \Delta^2 v^2 C_\phi^2 \}]$$

$$F_4 = [1 + \gamma \{ C_y^2 + (2\Lambda^2 v^2 + \Lambda v^2) C_\phi^2 - 4\Lambda v \rho C_y C_\phi \}]$$

$$G_4 = [1 + \gamma \{ C_y^2 + 3\Delta^2 v^2 C_\phi^2 - 4\Delta v \rho C_y C_\phi \}]$$

$$H_4 = \left[1 + \gamma \left\{ C_y^2 + \left(\Delta^2 v^2 + \Lambda \Delta v^2 + \frac{\Lambda(\Lambda+1)}{2} v^2 \right) C_\phi^2 - 2v(\Lambda + \Delta) \rho C_y C_\phi \right\} \right]$$

$$I_4 = \left[1 + \gamma \left\{ \frac{\Lambda(\Lambda+1)}{2} v^2 C_\phi^2 - \Lambda v \rho C_y C_\phi \right\} \right]$$

$$J_4 = [1 + \gamma \{ \Delta^2 v^2 C_\phi^2 - \Delta v \rho C_y C_\phi \}]$$

Proposal of a causal model measuring the impact of an ISO 9001 certified Quality Management System on financial performance of Moroccan service-based companies

El Moury Ibtissam¹, Mohamed Hadini², Adil Chebir³
Ben Ali Mohamed⁴, Echchelh Adil⁵

Abstract

Implemented by an increasing number of organisations worldwide, the ISO 9001 standard for quality management received considerable attention in the existing literature. Researchers worldwide have found positive, negative and even mixed effects of ISO 9001 certification on firms' performance, while in Morocco this issue has been rarely examined. It is the combination of these observations that led to this study.

The aim of this paper is to test and validate a causal model designed to measure the performance of an ISO 9001 certified Quality Management System (QMS) and its impact on a company's financial performance. By means of this causal analysis/model, the study examines the relationship between:

- QMS and the financial performance of 41 companies based in Morocco;
- the management responsibility process and all the QMS processes;
- the management resources process and all the QMS processes;
- the organisational and financial performance of the studied companies.

All of the considered firms are part of the service industry and range from medium-sized to large companies.

The data gathered in this study have been instrumental in devising actionable insights. The Statistical Package for the Social Sciences (SPSS) was the statistical software platform that enabled the use of a linear regression analysis to prove the positive correlation between the above-mentioned elements.

Key words: financial performance, organisational performance, Quality Management System (QMS), ISO 9001 certification.

¹ Electronic Systems, Information Processing, Mechanics and Energetics Laboratory, Faculty of Science, Ibn Tofail University, Morocco. E-mail: ibtissamelmoury@gmail.com. ORCID: <https://orcid.org/0009-0000-6175-0380>.

² Mechanics, Production and Industrial Engineering Laboratory, Higher School of Technology, Hassan II University, Casablanca, Morocco. E-mail: mohammed.hadini1@gmail.com.

³ Electronic Systems, Information Processing, Mechanics and Energetics Laboratory, Faculty of Science, Ibn Tofail University, Morocco. E-mail: achebir@gmail.com.

⁴ Mechanics, Production and Industrial Engineering Laboratory, Higher School of Technology, Hassan II University, Casablanca, Morocco. E-mail: benali8mohamed@gmail.com

⁵ Electronic Systems, Information Processing, Mechanics and Energetics Laboratory, Faculty of Science, Ibn Tofail University, Morocco. E-mail: adilechel@gmail.com. ORCID: <https://orcid.org/0000-0002-5302-4255>.



1. Introduction

The service industry is one of the most important components in Morocco's economic performance. It provides jobs and valuable services to the economy and hence supplies a substantial contribution to the national GDP (Gross Domestic Product). It is, therefore, a support sector contributing to national growth, witnessed by the positive correlation between the evolution of overall economic activity and the growth of the sector.

However, this sector remains relatively less developed because of its fragmented structure, its high cost and the shortcomings recorded in terms of organization and management, especially in quality management.

In fact, the Moroccan firm services evolve in an environment characterized by a competitive offer more and more strong, a requirement of competitiveness more and more acute and high customers' expectations. To become and remain competitive in its market, the firm service must establish its brand image and strengthen the reputation of its services. Flexibility, speed, and adaptability are the imperatives that the company should meet, in all circumstances. Therefore, the service provider must master and ensure an efficient quality management and customer satisfaction.

In this context, ISO 9001 certification remains one of the most suitable tools for harmonizing practices and establishing a dynamic of continuous improvement, covering the quality of services of all activities of any organization.

However, the question that is arising is to what extent a quality management system 'QMS' certified ISO 9001 has a positive impact on the performance of the Moroccan firm service, mainly the financial performance, what is the relationship between the latter and organizational performance? And how can this system allow the organization to reach its efficient objectives?

2. Literature Review

2.1. Organizational and Financial Performance

For a long time, performance has been a one-dimensional concept, measured by profit alone, mainly because of the weight of owners in the decision-making process. From this perspective, performance measurement focuses mainly on creating value for shareholders. It is therefore not surprising that corporate management is focused on this value creation and the way to manage it. Recent studies show that 200 companies listed by Fortune magazine currently use an indicator based on the value created for shareholders to evaluate performance (Patrick Jaulient (2012)).

Despite this observation, it should be noted that at this stage this purely financial logic is strongly criticized in the existing literature (Dohou and Berland (2007)), because it does not integrate the various actors who participate in the development of the firm (managers, employees, customers, etc.).

According to Serhan (2019), firm performance entails three areas of the company including product market, shareholder return, and financial performance. The improvement of this performance includes business re-engineering activities, processes for continuously improving the business and the quality of services or products offered. To ensure that the organization is efficient it is necessary to analyse the main performance indicators (Barna and Roxana (2021)).

According to Rashid et al. (2018), firm performance could be categorized into two broad categories, financial and nonfinancial measures. Some researchers used different terms, such as financial and operational performance measures, finance and efficiency and short and long-term measures. Short-term measures are normally based on the financial returns, while long-term measures are normally based on the non-financial returns. In general, financial performance indicators are a set of variables which usually can show the firm's capability in making profits, while non-financial indicators are a set of variables that are not measured by financial systems (Al-Mamar et al. (2020)).

For Zehir et al. (2018), there are two performance classifications which are qualitative and quantitative performance. Qualitative performance is largely related to the culture, environment, human resources, and abstract outputs within the organization and includes criteria such as employee satisfaction, customer satisfaction, quality and innovation performance. Quantitative performance includes criteria such as turnover increase, market share increases and profitability increase, which are partly influenced by qualitative factors and moreover based on marketing and financial management success. Bartoli and Blatrix (2015) believed that the definition of performance should be achieved through items such as piloting, evaluation, efficiency, effectiveness, and quality.

According to Rafoi (2016), a company's performance indicators can be classified as follows:

- Strategic indicators: market share, turnover, customer satisfaction, return (profit).
- Managerial indicators: availability of resources, costs, budget.
- Operational indicators: individual performance, processes performance, products, efficiency.

For Moulai Ali (2012), organizational performance "deals with how the firm is organized to achieve its goals and mainly how to realize them in a good way".

Organizational performance determines the ability of the firm to implement effective processes to reach its operational and strategic projections. The pillars of this efficiency can only be:

- The development and respect of a 'Process' approach.
- The relationships between the pilots of the different departments of the organization.
- The quality of the information flow.
- And the degree of flexibility of the organization.

Regarding the measurement of organizational performance, Berberoglu (2018) suggests that we can measure it by evaluating numerical data, which includes objective and timely information about how the organization is doing. However, performance measurement is not always necessarily based on objective data.

Hadini et al. (2020) concluded that: the theoretical difficulty in defining the concept of performance and that of organizational efficiency means that the use of indicators to measure one or the other of these concepts, depending on the objectives set by the firm, remains the only way to assess the functioning of a firm.

Regarding financial performance, Farrukh et al. (2016) consider it as the extent to which a company financial health over a period is measured. In other words, financial performance is a composite of an organization's financial health, its ability and willingness to meet its long-term financial obligations and its commitments to provide services in the foreseeable future. In a broader sense, financial performance refers to the degree to which financial objectives are accomplished (Ganyam and Ivungu (2019)). Cost-related performance is measured by quantitative indicators such as return on investment and sales, profitability, productivity, return on assets, efficiency, etc.

In the literature, the financial performance is measured based on a variety of indicators, or data issued from financial statements, balance sheets, income statements, statement of cash flows, etc., but can also refer to market data (e.g. market value of the shares). It can be defined by a variety of indicators, such as turnover (sales), return on assets (ROA), return on sales (ROS), return on equity (ROE), return on investment (ROI), earnings per share (EPS), earnings before interest (EBI), tax depreciation and amortization (Matradi and Mounir (2022)).

For a long time, this financial aspect of performance has remained the reference in terms of company performance and evaluation. Even if it facilitates a simple reading of the company's management, this financial dimension alone no longer ensures the company's competitiveness.

2.2. ISO 9001 standard and certification

Every firm must meet the requirements of its stakeholders. The set of political, policies and procedures that allow to satisfy these requirements form what is commonly called a 'Quality Management System (QMS)'.

QMS leads to the control of the processes and the quality of products (or services) which, in turn, allows the satisfaction of the customer and the achievement of the economic objectives that the organization has predicted to be reached (El Moury et al. (2020)).

We should say, to succeed in the business world, building and retaining a customer base is compulsory. In other words, it is necessary to respond in an optimal way to the present expectations of the market as well as to the futures ones, which inevitably passes by the offer of quality services.

A sharp expertise in Quality Management is the tool to succeed in this mission and ISO 9001 is the preferred way to standardize and make a QMS reliable. Thus, alignment with customer and regulatory requirements becomes possible (El Moury et al. (2020)).

Certification also has a word to say in this context. This certification, which translates into the delivery of a written assurance by an external and independent organization, gives the necessary glow to any QMS respecting the ISO 9001 standard.

Indeed, and beyond the fact that it brings a contribution of experienced experts, the certification guarantees to the company a solid reputation and a more preponderant influence. In addition to this, the organization's continuous improvement and better chances to gain market share are guaranteed.

According to Echour and Nbigui (2021), the adoption of ISO 9001 is a voluntary approach, certification is a fashion effect for companies that want to be recognized for their quality and resist in an increasingly competitive market.

For Isuf et al. (2016), ISO 9001 standards do not refer to the compliance with a given goal or result. In other words, they are not performance standards measuring the quality of a firm's products or services or a firm's environmental results, rather, they are standards setting out the need to systematize and formalize many corporate processes within a set of procedures, and to document such implementation.

International Organization for Standardization 'ISO' is an independent, non-governmental international organization, bringing together experts to share knowledge to develop consensus-based, market-relevant, voluntary international standards, which support innovation and provide solutions to global challenges.

In its ISO/IEC guide, ISO defines a standard as "a document established by consensus and approved by a recognized body, which provides, for common and

repeated use, rules, guidelines or characteristics for activities or their results that ensure an optimum level of order in a given context".

ISO defines certification as follows: 'Procedure by which a third party gives written assurance that a product, process or service conforms to the requirements specified in a standard'. Thus, the ISO 9001 certification certifies that an organization has a management system that complies with the ISO 9001 standard. Note that the other standards of the 9000 series: vocabulary (ISO 9000), guidelines (ISO 9004), do not contain requirements and cannot be used as a basis for certification.

3. Research Methodology

The objective of this research work is to test and validate a conceptual model (causal) allowing to measure the impact of:

- The processes of quality management system an 'ISO 9001 certified' on financial performance.
- Organizational performance on financial performance.
- Management Responsibility Processes on these three processes: Service Realization Process 'SRP', Measurement, Analysis, and Improvement Process 'MAIP' and Resource Management Process 'RMP'.
- The Resource Management process on the Service Realization process 'SRP' and the Measurement, Analysis, and Improvement process 'MAIP'.

Using the method of linear regression by SPSS software on a sample of 41 questionnaires administered face to face to directors and quality managers of Moroccan firms, certified ISO 9001, in the service sector (banking, transport, trade ...)

3.1. Research model constructs definitions

Our model is composed of two major constructs:

- The processes of an ISO 9001 certified quality management system.
- Organizational and financial performance.

3.1.1. First research construct: Processes of an ISO 9001 certified QMS

A quality management system (QMS) is the set of activities by which the organization defines, implements, and reviews its quality policy and objectives in accordance with its strategy. An organization's QMS is made up of interrelated and interactive processes that use resources to achieve intended results and deliver value (product, service, etc.). The QMS processes form our first research construct:

- Management Responsibility Process: MRP;

- Service Realization Process: SRP;
- Measurement, Analysis, and Improvement Process: MAIP;
- Resource Management Process: RMP.

3.1.2 Second research construct: Performance of the firm

- Organizational performance: OP;
- Financial performance: FP.

3.2. Presentation of the research model

Our model is based on 6 criteria, which are divided into 2 families: 4 criteria refer to the means (QMS process), the other criteria refer to the results (financial and organizational performance) (Figure 1).

We assume that there is a causal relationship between the criteria of means and the criteria of results. In other words, the means in place are the causes of the given results. It should be noted that for each causal relationship a hypothesis has been formulated. Since the proposed conceptual model has 10 causal relationships, 10 hypotheses have been formulated.

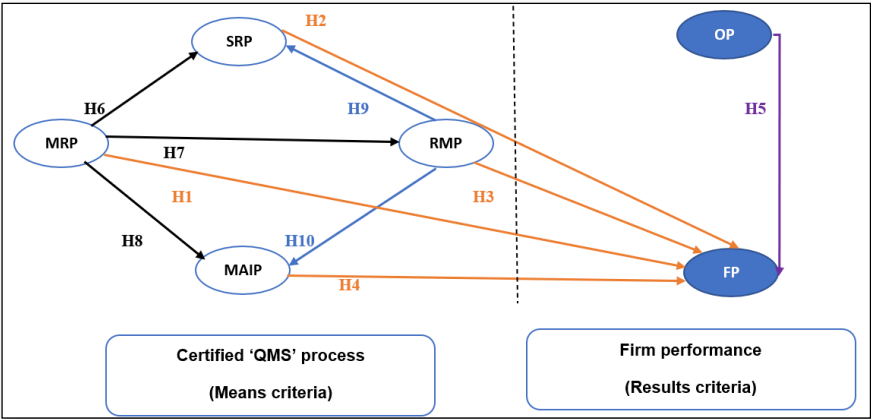


Figure 1: The Proposed Model

Table 1: Showing the codes used in the causal model

Model construct proposed	Code	Title
Certified 'QMS' process (Means criteria)	MRP	Management Responsibility Process
	SRP	Service Realization Process
	MAIP	Measurement, Analysis, and Improvement Process
	RMP	Resource Management Process
Firm performance (Results criteria)	OP	Organizational performance
	FP	Financial performance

3.3. Formulated hypotheses

We intend through our study to validate or invalidate the Ten Hypotheses below:

Table 2: List of hypotheses

Hypothesis Number	Causal Relationship	Hypothesis Formulated
H1	MRP→FP	We suppose that MRP has a strong impact on FP
H2	SRP→FP	We suppose that SRP has a strong impact on FP
H3	RMP→FP	We suppose that RMP has a strong impact on FP
H4	MAIP→FP	We suppose that MAIP has a strong impact on FP
H5	OP→FP	We suppose that OP has a strong impact on FP
H6	MRP→SRP	We suppose that MRP has a strong impact on SRP
H7	MRP→RMP	We suppose that MRP has a strong impact on RMP
H8	MRP→MAIP	We suppose that MRP has a strong impact on MAIP
H9	RMP→SRP	We suppose that RMP has a strong impact on SRP
H10	RMP→MAIP	We suppose that RMP has a strong impact on MAIP

3.4. Qualitative and quantitative study

Before writing our questionnaires, we collected primary data from a qualitative study. The purpose of this phase was to identify the main benefits of certification for the company. Once collected, the material – entirely transcribed – was used for a thematic content analysis.

We chose to conduct a study of managers' perception of the benefits of certification, by administering a questionnaire. The questionnaire was structured in distinct questions operationalizing the different themes emerging from the qualitative study.

3.5. Data Collection

Our study is empirical, and the collection of information is conducted by a questionnaire. This latter is divided into four parts:

- **Part One**, including information such as the name, the firm size and the type of service provided by the latter, in addition to the motivation for the implementation of a QMS certified ISO 9001.
- **Part Two** is used to collect information allowing to test the causal relationship between the various processes of the Certified ISO 9001 QMS.
- **Parts Three and Four** are used to collect information allowing to test the strength and the sense of the various causal relationship, which can exist between the impact of the ISO 9001 certification and the axes relating to the organizational and financial performance.

To implement our measurement instrument, we use Churchill's [1979] best-known paradigm. This paradigm is mostly used by most researchers to develop their own measurement scales rather than using the instruments that already exist, Legardinier (2013).

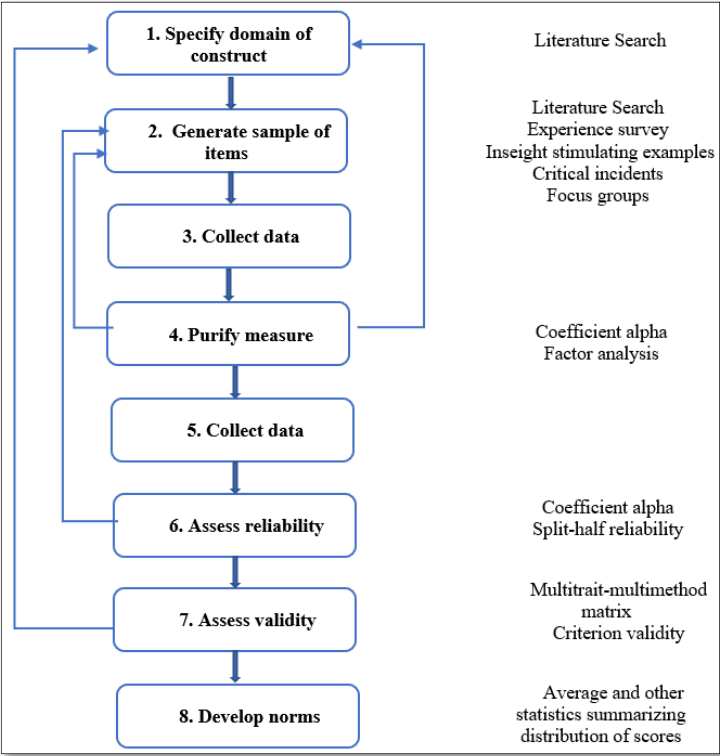


Figure 2: The Churchill Paradigm Approach [1979]

3.6. Established measurement scales

To provide a measure for all the criteria of the designed model, we opted for the Likert scale, as it is considered the most known, the most appropriate for opinion studies and for sure the most used (Evrard et al. (2003)).

The questionnaire items are collected on the 5 point Likert scales: (Strongly disagree, somewhat disagree, moderately agree, somewhat agree, strongly agree). Note that the total number of items is 50 (items/questions).

3.7. Reliability test

The reliability of measurements is concerned with the reduction of the random part of the measurement error: if the same phenomenon is measured several times by the same measuring instruments, the results should be as close as possible.

For this reason, we will use Cronbach's alpha index. This index allows us to study the internal consistency between the set of items for each latent variable, that is, it allows us to measure the reliability of the measurements of a set of questions (or items) intended to measure a specific phenomenon (the answers to questions on the same subject must be correlated so that all the interviewees or respondents must have the same understanding of each separate question).

3.8. Testing the reliability of the two research constructs

From the results in Table 3, we can conclude that all values exceed 0.7. This shows that the items composing the two constructs have a good internal consistency.

Table 3: Reliability analysis of the two research constructs

Criteria	Variables	Code	Alpha value Cronbach	Number of items
Means criteria	Management Responsibility Process	MRP	0.939	9
	Service Realization Process	SRP	0.829	6
	Measurement, Analysis, and Improvement Process	MAIP	0.920	11
	Resource Management Process	RMP	0.849	6
Results criteria	Organizational performance	OP	0.923	9
	Financial performance	FP	0.842	6

4. Results of the empirical research

In this part we present the results of our empirical study

4.8. Measuring the relationship between QMS criteria factors and financial performance

4.8.1. Overall Model: QMS Processes (MRP, SRP, MAIP, RMP) – Financial Performance 'FP'

According to Table 4, we observe that the strength of the relationship between the QMS criteria factors and Financial Performance is quite strong in a positive way ($R=0.615$). The QMS factors explain 26.5% of the financial performance (adjusted $R\text{-squared}=0.265$). Overall, the model is valid ($\text{significance}=0.028<5\%$, Table 5).

Table 4: Summary of overall model (QMS process-FP)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
1	0.615	0.378	0.265	0.83993530

- a. Predicted values: MRP, SRP, MAIP, RMP
- b. For regression at the origin (model without constant). R-squared measures the proportion of variability in the dependent variable around the origin determined by regression. This cannot compare to R-squared for models that include a constant.
- c. Dependent variable: FP
- d. Linear regression at the origin

Table 5: Analysis of variance (QMS process-FP)

Model	Sum of squares	ddl	Average square	D	Sig
Regression	9.423	4	2.356	3.339	0.028
Residual	15.521	22	0.705		
Total	24.943	26			

- a. Dependent variable : FP
- b. Linear regression at the origin
- c. Predicted values: MRP, SRP, MAIP; RMP
- d. This total of squares is not corrected for the constant because the constant is zero for the regression at the origin.

4.8.2. Regression Equation: QMS Process - Financial Performance

Based on Table 6, the linear regression equation can be written as follows:

FP = 0.662 MRP- 0.260SRP + 0.297 RMP - 0.114 MAIP

Note that the significances for SRP, RMP, MAIP are invalid (sig.>5%). But this is not the case for MRP. Therefore, the relationship between FP and MRP is significant.

Table 6 : Criteria Coefficients (QMS-FP)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig.(P-value)
	A	Standard error	Beta		
MRP	0.662	0.296	0.663	2.233	0.036
SRP	-0.260	0.326	-0.261	-0.796	0.435
RMP	0.297	0.401	0.275	0.741	0.467
MAIP	-0.114	0.359	-0.097	-0.317	0.754

4.9. Measuring the relationship between organizational performance and financial performance

4.9.1. Overall Model: (OP-FP)

Generally, the strength of the relationship between organizational performance and financial performance is quite high ($R=0.704$). The variable to be explained (PE) explains 49.5% of the predicted variable PO ($R\text{-squared}=0.495$). Also, the significance value of the model is lower than 5%. Therefore, the model is globally valid.

Table 7: Summary of global model (OP-FP)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
	0.704	0.495	0.483	0.71032858

- a. Predicted values: OP
- b. For regression at the origin (model without constant). R-squared measures the proportion of variability in the dependent variable around the origin determined by regression. This cannot compare to R-squared for models that include a constant.
- c. Dependent variable: FP
- d. Linear regression at the origin

Table 8: Analysis of variance (OP-FP)

Model	Sum of squares	ddl	Average square	D	Sig
Regression	19.817	1	19.817	39.276	0.000
Residual	20.183	40	0.505		
Total	40.000	41			

- a. Dependent variable : FP
- b. Linear regression at the origin
- c. Predicted value: OP
- d. This total of squares is not corrected for the constant because the constant is zero for the regression at the origin.

4.9.2. Regression equation (OP-FP)

Based on Table 9, the linear regression equation can be written as follows:

FP= 0.704 OP

The significance of the OP-FP causal relationship is significant ($Sig=0 < 5\%$), we can conclude that OP strongly impacts FP.

Table 9: Criteria Coefficients (OP-FP)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
OP	0.704	0.112	0.704	6.267	0.00

4.10. Measuring the relationship between the Management Responsibility Process 'MRP' and the Service Realization Process 'SRP'

4.10.1. Global Model (MRP-SRP)

There is thus a quite strong relationship between SRP and MRP ($R=0.783$). The dependent variable SRP explains 61.4% of the relative independent variable MRP. (Table 10). The significance is less than 5%. This proves the validity of the overall model. (Table 11).

Table 10: Summary of global model MRP-SRP

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
	0.783	0.614	0.598	0.62409402

- Predicted value: MRP
- For regression at the origin (model without constant). R-squared measures the proportion of variability in the dependent variable around the origin determined by regression. This cannot compare to R-squared for models that include a constant.
- Dependent variable: SRP
- Linear regression at the origin.

Table 11: Analysis of variance MRP-SRP

Model	Sum of squares	ddl	Average square	D	Sig
Regression	15.476	1	15.476	39.734	0.000
Residual	9.737	25	0.389		
Total	25.213	26			

- Dependent variable : MRP
- Linear regression at the origin
- Predicted value: SRP
- This total of squares is not corrected for the constant because the constant is zero for the regression at the origin.

4.10.2. Regression equation (MRP-SRP)

The linear regression equation can be written as follows: **SRP= 0.787 MRP**. Also, the causal relationship is significant (<5%).

Table 12: Criteria Coefficients (MRP-SRP)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig.(P-value)
	A	Standard error	Beta		
MRP	0.787	0.125	0.783	6.303	0.00

- a. Dependent variable: SRP
- b. Linear regression at the origin

4.11. Measuring the relationship between the Management Responsibility Process and the Resource Management Process (MRP--->RMP)

4.11.1. Global Model (MRP-RMP)

There is thus a quite strong relationship between RMP and MRP (R=0.757). The significance is less than 5%. This proves the validity of the overall model.

Table 13: Summary of global model (MRP-RMP)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
	0.757	0.573	0.556	0.60426602

- a. Predicted value: MRP
- b. For regression at the origin (model without constant). R-squared measures the proportion of variability in the dependent variable around the origin determined by regression. This cannot compare to R-squared for models that include a constant.
- c. Dependent variable: RMP
- d. Linear regression at the origin.

Table 14: Analysis of variance (MRP-RMP)

Model	Sum of squares	ddl	Average square	D	Sig
Regression	12.249	1	12.249	33.547	0.000
Residual	9.128	25	0.365		
Total	21.378	26			

- a. Dependent variable : RMP
- b. Linear regression at the origin
- c. Predicted value: MRP
- d. This total of squares is not corrected for the constant because the constant is zero for the regression at the origin.

4.11.2. Regression Equation (MRP-RMP):

The linear regression equation can be written as follows: **RMP= 0.7 MRP**.

Also, the causal relationship is significant (<5%).

Table 15: Criteria Coefficients (MRP-RMP)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
MRP	0.700	0.121	0.757	5.792	0.00

4.12. Measuring the Relationship between the Management Responsibility Process and the Measurement, Analysis and Improvement Process

4.12.1. Global Model (MRP-MAIP):

There is thus a quite strong relationship between MAIP and MRP ($R=0.725$). The dependent variable MAIP explains 52.6% of the relative independent variable MRP. The significance is less than 5%. This proves the validity of the overall model.

Table 16: Summary of global model (MRP-MAIP)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
	0.725	0.526	0.507	0.58705996

- Predicted value: MRP
- For regression at the origin (model without constant). R-squared measures the proportion of variability in the dependent variable around the origin determined by regression. This cannot compare to R-squared for models that include a constant.
- Dependent variable: MAIP
- Linear regression at the origin

Table 17: Analysis of variance (MRP-MAIP)

Model	Sum of squares	ddl	Average square	D	Sig
Regression	9.553	1	9.553	27.720	0.000
Residual	8.616	25	0.345		
Total	18.169	26			

- Dependent variable : MAIP
- Linear regression at the origin
- Predicted value: MRP
- This total of squares is not corrected for the constant because the constant is zero for the regression at the origin.

4.12.2. Regression equation (MRP-MAIP)

The linear regression equation can be written as follows: MAIP= 0.618 MRP.

Also, the causal relationship is significant (<5%).

Table 18: Criteria Coefficients (MRP-MAIP)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
MRP	0.618	0.117	0.725	5.265	0.00

4.13. Measuring the relationship between the Resource Management Process and the Service Realization Process (RMP--->SRP)

4.13.1. Overall Model RMP--->SRP

There is thus a quite strong relationship between SRP and RMP ($R=0.834$). The dependent variable SRP explains 69.5% of the independent variable RMP. The significance is less than 5%. This proves the validity of the overall model.

Table 19: Summary of global model (RMP-SRP)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
	0.834	0.695	0.688	0.55200219

- Predicted value: SRP
- For regression at the origin (model without constant). R-squared measures the proportion of variability in the dependent variable around the origin determined by regression. This cannot compare to R-squared for models that include a constant.
- Dependent variable: RMP
- Linear regression at the origin.

Table 20: Analysis of variance (RMP-SRP)

Model	Sum of squares	ddl	Average square	D	Sig
Regression	27.812	1	27.812	91.274	0.000
Residual	12.188	40	0.305		
Total	40.000	41			

- Dependent variable : RMP
- Linear regression at the origin
- Predicted values: SRP
- This total of squares is not corrected for the constant because the constant is zero for the regression at the origin.

4.13.2. Regression equation (RMP-SRP)

Referring to Table 21, the linear regression equation can be written as follows: **SRP = 0.834 RMP**. Also, the causal relationship is significant (<5%).

Table 21: Criteria Coefficients (RMP-SRP)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig.(P-value)
	A	Standard error	Beta		
SRP	0.834	0.087	0.834	9.554	0.00

4.14. Measuring the Relationship between the Resource Management Process and the Measurement, Analysis, and Improvement Process (RMP--->MAIP)

4.14.1. Global Model: RMP-MAIP

Similarly, overall, there is a strong relationship between RMP and MAIP ($R=0.872$). The dependent variable MAIP explains 76.1% of the relative independent variable RMP. The significance is less than 5%. This proves the validity of the overall model.

Table 22: Summary of global model (RMP-MAIP)

Model	R	R-squared	Adjusted R-squared	Standard error of estimation
	0.872	0.761	0.755	0.48903625

- Predicted value: RMP
- For regression at the origin (model without constant). R-squared measures the proportion of variability in the dependent variable around the origin determined by regression. This cannot compare to R-squared for models that include a constant.
- Dependent variable: MAIP
- Linear regression at the origin.**

Table 23: Analysis of variance (RMP-MAIP)

Model	Sum of squares	ddl	Average square	D	Sig
Regression	30.434	1	30.434	127.255	0.000
Residual	9.566	40	0.239		
Total	40.000	41			

- Dependent variable : MAIP
- Linear regression at the origin
- Predicted values: RMP
- This total of squares is not corrected for the constant because the constant is zero for the regression at the origin.

4.14.2. Regression Equation RMP-MAIP

The linear regression equation can be written as follows: $MAIP = 0.872 RMP$. Also, the causal relationship is significant ($Sig < 5\%$).

Table 24: Criteria Coefficients (RMP-MAIP)

Model	Unstandardized coefficients		Standardized coefficients	t-student	Sig. (P-value)
	A	Standard error	Beta		
RMP	0.872	0.077	0.872	11.281	0.00

4.15. Global model with the results and hypothesis testing results

The global model with the results and hypothesis testing results are presented as follows:

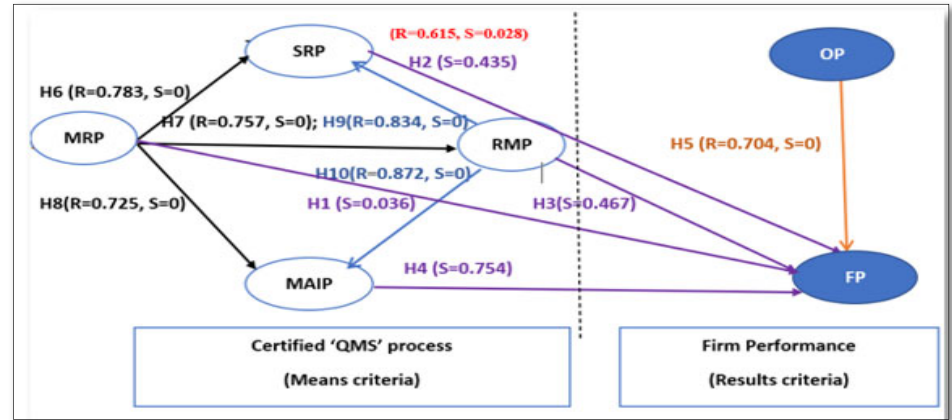


Figure 3: Global model with results

Table 25: Hypothesis Testing Results

Hypotheses	Results
H1: MRP→FP	Valid
H2: SRP→FP	Invalid
H3 : RMP→FP	Invalid
H4 : MAIP→FP	Invalid
H5 : OP→FP	Valid
H6 : MRP→SRP	Valid
H7 : MRP→RMP	Valid
H8 : MRP→MAIP	Valid
H9 : RMP→SRP	Valid
H10 : RMP→MAIP	Valid

5. Discussion and conclusion

Firms put in mind their concern for maximizing yield, so it is very important to analyse the impact of an ISO 9001 certified QMS on the firm's financial results.

Otieno and Kithae (2021) assessed the effect of the implementation of ISO 9001 quality management system on the financial performance of SMEs in Kenya. The upshot of these studies concluded that the implementation of this system has statistically no significant effect on the financial performance of these firms. Also, Islam et al. (2015) and Zondo (2018) confirm that there is no significant relationship between these two elements.

In a business environment where an economic downturn and financial crisis dominates (Greece to be precise). Evangelos and Dimitrios (2014) concluded that the ISO 9001 certified manufacturing companies significantly outperform the non-certified regarding product quality, customer satisfaction, operational, market and financial performance. Also, Matheus et al. (2021) proved that ISO 9001 certification has a positive impact on the financial performance of several Brazilian companies. Furthermore, an in-depth analysis of 92 studies shows that the ISO 9001 certification helps companies to increase their income and financial performance (Basak et al. (2018)). Always in the same context the results of several research show that there is a positive relationship between the quality management system implementation and the financial performance (Astrini, 2021; Ionaşcu et al., 2017; Jalil et al., 2017).

In Iceland, the study of Hróbjartsson (2012) was in a comparative perspective between ISO 9001 certified firms and those that are not. The result of this work justified the remarkable difference in financial performance on behalf of companies holding ISO 9001 certification. Also, Valmohammadi & Kalantari (2015) confirm the same finding: the ISO 9001 certified companies have a more relevant performance than companies not certified.

The effect of ISO 9001 standards on financial performance continued to attract the attention of Nguyen et al. (2016) in Vietnam, whose results confirmed the causal relationship between these two concepts. In the United States of America, the research of Awoku (2012) argued the feasibility of this version of the ISO 9001 international standard as a determinant of financial performance and supplier selection.

In Pakistan, the findings of the research work of Faryal et al. (2019) suggest that QMS certified ISO 9001 implementation has a significant and positive role in improving innovation and financial performance of the manufacturing organizations.

With regard to the kingdom of Morocco, and from an economic environment perspective, Ben Ali (2016) measured the impact of an ISO 9001 certified quality approach (case: young manufacturing companies in growth phase, located in the north of Morocco) on the firm's overall performance, including financial performance, the results showed that there is a positive relationship between these two elements.

Also, Belkasseh (2019) affirm the positive effect of ISO registration on the financial performance. However, Hadini (2020) proved that the practices of an ISO 9001 certified quality approach (case: multinational firm located in Morocco) have a weak impact on the axes of financial performance.

Based on a large sample of certified firms (21,482 ISO 9000 certifications issued in the United States), Corbett et al. (2005) reported a positive influence of certification on financial performance. Specifically, the firms that experienced a deterioration in financial performance were those that did not seek certification. While Sharma (2005) confirms the same influence, Martynez-Costa and Martynez-Lorente (2007) display an opposite opinion by showing that certification produces a negative effect on financial performance.

According to Coffey et al. (2011), the deployment of the quality management principles of the ISO 9001 international standard has a positive and significant impact on product and service quality, increased sales and market share, profitability, product sustainability and employee satisfaction.

Matradi and Mounir (2022), conducted a review of the literature regarding the effects of an ISO 9001 certified QMS on financial performance, the results of their study show that: some works attempted to show a positive effect (56%), when others showed a negative effect (10%). Some authors report neutral or mediated impact (15%). However, many authors do not confirm it (19%).

According to our exploratory study, the results show that:

- The strength of the relationship between the factors QMS criteria, ISO 9001 certified and financial performance is positive and quite strong.
- The criterion 'Management responsibility' positively influences the financial performance.
- The relationship between the 'Service realization' criterion and financial performance is invalid: this result highlights the fact that although ISO 9001 standards have been integrated, their manifestation in the service realization process does not seem to be visible to managers.
- The relationship between the criterion 'Resource management' and financial performance is invalid, we can interpret this result as follows: An ISO 9001 certification can only be beneficial if the leadership of the organization invests in the quality of resources (human resources in particular) with the mission of managing the pre and post certification.
- The relationship between the criterion 'Measurement, analysis and improvement' and financial performance is invalid, this result is fairly revealing: the audit of a certified QMS is a pillar of success. Indeed, if the leadership of the organization does not give much importance to the measurement and continuous improvement of its certified QMS, there can be little economic benefit from an ISO 9001 certification.

- Organizational performance has a positive and fairly strong influence on financial performance.

According to Ataseven et al. (2015), managers must invest their efforts in the proper integration of the requirements of the ISO 9001 international standard and not in obtaining certification. And Jang and Lin (2008) confirm that thorough QMS implementation significantly determines financial performance.

The efficiency of the management responsibility process within a firm results from:

- A clear dissemination of the principles of its mission, vision, values, and ethics.
- Disseminating to its team the knowledge to be put into practice to improve the implementation and the quality of the services rendered to clients and other interested parties.

The criterion of Management Responsibility aims at assessing the action of the leaders and the excellence of their behaviour, in the accomplishment of their mission, as well as their vision of the organization through the implementation of values and systems necessary for sustainable success. In accordance with the research of Calvo-Mora et al. (2005), this criterion has effects on the criteria of resource and process management.

Thus, the results of our study prove that this criterion positively influences in a fairly strong way the criteria: 'Service realization', 'Resource management' and 'Measurement, analysis and improvement'.

In consideration of the analysis of the observations made by the authors mentioned above and the results of our study, we realize that the commitment of the leaders to manage the processes of the ISO 9001 certified QMS in an efficient way must be visible, permanent, "proactive" and exist at all levels of management, and to enhance firm's performance it is important for management to understand and find different sources of leadership that will lead to improved organizational performance (Uhl-Bien et al. (2014)).

"ISO 9001 standards consider that a management system consists of interacting processes. All processes need, to function, planning, information, resources, evaluation and monitoring, improvements, and decisions internal to the process or provided by other processes such as support processes, management processes and implementation processes. These different processes are linked and influence each other". According to our empirical study we have proven that the process 'Resource Management' does indeed influence positively and strongly all processes of the QMS.

Finally, it should be noted that any research work such as ours has certain limitations for different reasons. For example, the method of data collection by means of a questionnaire is not free of limitations. It only allows for the collection of subjective data and information (perceptions of managers).

References:

- Ataseven, C., Nair, A. and Prajogo, I., (2015). ISO 9000 Internalization and Organizational Commitment – Implications for Process Improvement and Operational Performance. *IEEE Transactions on Engineering Management*, vol. 6, no. 1, pp. 5–17.
- Awoku, R. A., (2012). *An empirical study on quality management practices, organization performance and supplier's selection in Southern Minnesota manufacturing firms*. Master of Science, Faculty of The Graduate School of Minnesota State University, Mankato, Minnesota: USA.
- Astrini, N., (2021). ISO 9001 and performance: a method review. *Total Quality Management & Business Excellence*, vol 32, (1-2), pp 5–32.
- Barna, L., Roxana, D., (2021). The influence of the implementation of ERP systems on the performance of an organization, *Proceedings of the 15th International Conference on Business Excellence*, vol.15, Num. 1, pp. 268–279.
- Bartoli, A., Blatrix, C., (2015). *Management dans les Organisations publiques – 4ème édition*. Dunod, Paris.
- Basak Manders, Henk de Vries and Knut Blind, (2018). *The Relationship Between ISO 9001 and Financial Performance: a Meta-analysis*, Academy of Management, <https://doi.org/10.5465/ambpp.2013.12255abstract>
- Berberoglu, A., (2018). Impact of organizational climate on organizational commitment and perceived organizational performance: empirical evidence from public hospitals. *BMC Health Serv Res* 18, 399, <https://doi.org/10.1186/s12913-018-3149-z>
- Ben Ali, (2016). Thèse de Doctorat en Génie Industriel : Proposition d'un modèle causal mesurant les impacts de la Qualité sur la Performance globale et la Responsabilité Sociétale des Entreprises (RSE): Cas des jeunes entreprises manufacturières en phase de croissance installées au Nord du Maroc. Université Hassan 2, Maroc.
- Belkasseh, M., (2019). The Relationship between Total Quality Management and Financial Performance: Evidence from Morocco. *Archives of Business Research*, 7(5), pp. 28–47.
- Calvo-Mora, A., Leal, A. and Roldan, J., (2005). Relationships between the EFQM Model Criteria: A Study in Spanish Universities. *Total Quality Management*, vol. 16, no.° 6, pp. 741–770.

- Coffey, V., Trigunarsyah, B. and Willar, D., (2011). *Quality Management System and Construction Performance*. Australia: School of Urban Development, Queenslan University of Technology.
- Corbett C. J., Montes-Sancho M. J. and Kirsch D. A., (2005). The Financial Impact of ISO 9000 Certification in the US: An Empirical Analysis. *Management science*, vol. 51, no°7, pp. 1046–1059.
- Dohou, R., Berland N., (2007). *Mesure de la performance globale des entreprises*, HALSHS.
- Echour Said; Taibi Nbigui, (2021). *Motivations related to the quality management system and benefits of its implementation in the company: state of the art*, IEEE 13th International Colloquium of Logistics and Supply Chain Management (LOGISTIQUA), 2–4 Dec. 2020, Fez, Morocco, HST (EST) – Sidi Mohamed Ben Abdellah University. [Viewed 01 April 2021]. Available from: <https://ieeexplore.ieee.org/abstract/document/9353877/metrics#metrics>
- El Moury, I., Hadini, M., Chebir, A., Echchelh, A., (2020). Impact of ISO 9001 Certification on Organizational Performance: State of the Art. *International Journal of Innovation and Applied Studies*, vol. 31, no. 3, pp. 648–654.
- Evrard, Y., Pras, B. and Roux, E., (2003). *Market Etudes et recherches en Marketing*. Edition n°3, Dunod, Paris.
- Evangelos, P., Dimitrios, K., (2014). Performance measures of ISO 9001 certified and non-certified manufacturing companies. *Benchmarking: An International Journal*, vol. 21, no. 5, pp. 756–774.
- Farrukh, I. Farah, N., Faizan, N., (2016). Financial Performance of Firms: Evidence from Pakistan Cement Industry. *Journal of Teaching and Education*, vol. 5, no. 1, pp. 81–94.
- Faryal Jalil, Shafiq M. and Wasim ul Rehman, (2019). Effect of QMS on innovation and financial performance a developing country perspective. *Business & Economic Review*, vol. 21, no. 3, pp. 56–72.
- Ganyam, A. L., Ivungu, J. A., (2019). Effect of Accounting Information System on Financial Performance of Firms: A Review of Literature. *Journal of Business and Management*, vol.21., no. 5, pp. 39–49.
- Hadini, M., Ben Ali, M., Rifai, S., Bouksour, O., Adri, A., (2020). Le Concept de la Performance Industrielle : Etat de l'Art. *International Journal of Innovation and Applied Studies*, 28(3), 726–739.

- Hróbjartsson, A., (2012). *Financial benefits of an ISO 9001 certification*. Iceland: Reykjavík University.
- Hadini Mohammed, (2020). Thèse de Doctorat en Génie Industriel : La Conduite du Changement par la Démarche Qualité, Santé-Sécurité & Environnement 'QSSE' en tant que levier de pilotage de la performance industrielle. Cas d'une entreprise multinationale implantée au Maroc. Université Hassan 2, Maroc.
- ISO, (2022). A propos de l'ISO, ISO, March 2022, [online] Available: <https://www.iso.org/fr/about-us.html>.
- Isuf, L., Mane, A., Ilir, K., Remzi, K., (2016). A Literature Review On Iso 9001 Standards, *European Journal of Business, Economics and Accountancy*, vol. 4, no. 2, pp. 81–85.
- Islam, M. M., Karim, M. A. and Habes, E. M., (2015). Relationship between quality certification and financial & non-financial performance of organizations. *The Journal of Developing Areas*, 49(6), pp. 119–132.
- Ionașcu, M., Ionașcu, I., Săcărin, M., & Minu, M., (2017). Exploring the impact of ISO 9001, ISO 14001 and OHSAS 18001 certification on financial performance: The case of companies listed on the Bucharest Stock Exchange. *Amfiteatru Economic Journal*, 19(44), pp. 166–180.
- Jang, W. Y., Lin, Ch. I., (2008). An integrated framework for ISO 9000 motivations, depth of ISO implementation and firm performance: The case of Taiwan. *Journal of Manufacturing Technology Management*, vol. 19, no. 2, pp. 194–216.
- Jalil, F., Shafiq, M., Rehman, W. and Akram, M. W., (2017). Assessing the Mediating Role of manufacturing competitive strategies in the relationship of Quality Management System and Financial Performance. *Journal of Managerial Sciences*, 11(3), pp. 415–432.
- Legardinier, A., (2013). Comment limiter les biais liés au choix des échelles de mesure dans les études marketing?, *Gestion et management*.
- Matheus, B. C, Fabiane, L. L, de Toledo, J. C., (2021). The Impact of ISO 9001 Certification on Brazilian Firms' Performance: Insights from Multiple Case Studies. *International Journal of Economics and Management Engineering*, vol. 15, no. 8, pp. 677–683.
- Matradi, S., Mounir, Y., (2022). The Effect of ISO 9001 Certification on Financial Performance: A Systematic Review. *International Journal of Accounting, Finance, Auditing, Management and Economics*, 3(2-1), pp. 83–99, <https://doi.org/10.5281/zenodo.6399991>.

- Martinez-Costa, M., Martinez-Lorente, A. R., (2007). A Triple Analysis of ISO 9000 Effects on Company Performance. *International journal of productivity and performance management*, vol. 56, no. 5/6, pp. 484–499.
- Moulai Ali, M., (2012). PhD thesis in economics: *The performance of the national cement industry in the light of contractual theories of organizations*, University of Oran, Algeria.
- Nguyen, A. D., Pham, C. H. and Pham, L., (2016). Total Quality Management and Financial Performance of Construction Companies in Ha Noi. *International Journal of Financial Research*, vol. 7, pp. 41–53.
- Otieno Walter Ochieng, Kithae P., (2021). Effect of ISO 9001 Quality Management System on Financial Performance of Small and Medium Enterprises in Kenya: A Survey of Top 100 SMES In Kenya. *International Journal of Management and Leadership Studies*, vol. 3, no. 2, pp.137–155.
- Patrick Jaulient, (2012). *How do you measure the performance of a company?*, Les Echos, viewed 24 May 2021: <http://archives.lesechos.fr/archives/cercle/2012/12/26/cercle_61804.htm>
- Rashid, N., Ismail, W. N. S. W., Abd Rahman, M. S. and Afthanorhan, A., (2018). Conceptual Analysis on Performance Measurement Used in SMEs Research: The Effectiveness of Firm's Overall Performance. *International Journal of Academic Research in Business and Social Sciences*, 8(11), pp.1401–1412
- Rafoi, A., (2016). *Top 5 KPIs for Distribution*, Bit software, viewed 30 March 2022: <<https://info.bitsoftware.eu/blog/bitsoftware-ro/5-indicatori-de-performanta-importanti-pentru-industria-de-distributie>>
- Serhan, A., El Hajj, W., (2019). Impact of ERPS on Organizations' Financial Performance. *Proceedings of the 13th International Conference on Business Excellence*, vol. 13. no. 1, pp.361–372.
- Sharma, D. S., (2005). The Association Between ISO 9000 And Financial Performance. *The International Journal of Accounting*, vol.40, no. 2, pp. 151–72.
- Uhl-Bien M, Schermerhon J. R. and Osborn R. N., (2014). Organizational behavior: experience, grow, Contribute. John Wiley and Sons inc.
- Valmohammadi, C., Kalantari, M., (2015). The moderating effect of motivations on the relationship between obtaining ISO 9001 certification and organizational performance, *The TQM Journal*, vol. 27, no. 5, pp. 503–518.

- Yaser Hasan Al-Mamar, Mohammed A. Alwaheeb, Naif Ghazi M. Alshammari, Mohammed Abdulrab, Hamad Balhareth, and Hela Ben Soltane, (2020). The effect of entrepreneurial orientation on financial and non-financial performance in Saudi Smes: A review, *Journal of critical Reviews*, vol. 7, no. 14, pp 200–208.
- Zehir, Cemal., Can, Esin., Urfa, A. Nerve, (2018). *Strategic Entrepreneurial Posture, Entrepreneurial Orientation and Firm Performance Relationship in Family Businesses*, The European Proceedings of Social & Behavioural Sciences.
- Zondo, R. W., (2018). Assessing the financial implications of quality management system accreditation on small training providers in KwaZulu-Natal. *South African Journal of Economic and Management Sciences*, 21(1), pp. 1–8.

Dynamics of survey responses before and during the pandemic: entropy and dissimilarity measures applied to business tendency survey data

Emilia Tomczyk¹

Abstract

This article is set within the framework of studies focusing on the impact of the SARS-CoV-2 virus on the dynamics of economic activity. For the purposes of the analysis of the expectations expressed in business tendency surveys, the paper aims to verify whether the pandemic of 2020-2022 can be seen as just another contraction phase. Entropy and dissimilarity measures are employed to study the characteristics of the expectations and assessments expressed in the business tendency survey of Polish manufacturing companies. The empirical results show that the dynamics of the manufacturing sector data, particularly as far as general economic conditions are concerned, set the pandemic period apart. The economic consequences of the COVID-19 pandemic expressed in business tendency surveys tend to be unfavourable, but the statistical properties or the degree of the concentration of respondents' answers do not correspond closely either to the expansion or contraction phases of the business cycle.

Key words: business cycles, survey data, expectations, manufacturing industry, COVID-19 pandemic.

1. Introduction

The COVID-19 pandemic is an ongoing global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of April 2022, there have been more than 504 million registered cases and 6.2 million deaths due to the virus (Worldometer, 2022). There is a consensus that the potential economic consequences of the recent world-wide pandemic will be significant. The World Bank estimates that the world economy has shrunk by 4.3% in 2020 (Boianovsky, Erreygers 2021). Substantial downward revisions in enterprises and households' economic sentiments have been noted in recent literature (see Bartik et al. 2020; van der Wielen, Barrios 2021; Meyer et al. 2022). However, the scale and range of

¹ SGH Warsaw School of Economics, Poland. E-mail: emilia.tomczyk@sgh.waw.pl.
ORCID: <https://orcid.org/0000-0002-4565-0352>.



negative effects of COVID-19 pandemic are far from obvious or uniform across countries and economic variables. Long-term consequences were found to depend on the region; for example, Teresiene et al. (2021) note that in the case of the Eurozone, the spread of COVID-19 pandemic did not affect the consumer-confidence index as much as in the US and China.

Similar ambiguities are noted with respect to response of the Polish economy to the pandemic. The World Bank points out that Poland has survived the pandemic relatively unscathed and may attain 3.3 percent growth in 2021 (The World Bank, 2021). Results of the RIED (Research Institute for Economic Development of SGH Warsaw School of Economics) survey in manufacturing sector show that in February 2022, enterprises evaluate their future prospects as favourable: with respect to 2021, manufacturing activity and industrial confidence indicators increase, inflation slows down, and the main survey balances (in production, orders, employment, and financial situation) reflect optimism of respondents (Adamowicz, Walczyk 2022). Generally, manufacturing sector enterprises express confidence as far as their own prospects are concerned even though their assessment of the general economic situation in Poland remains pessimistic. Also, the official aggregated statistics of Statistics Poland paint a darker picture. In March 2022, monthly general business climate indicator in manufacturing remains lower as compared to the corresponding month of the previous year: down 6.6 points for non-seasonally adjusted indicator and 7.2 points for seasonally adjusted one (GUS 2022, p. 9). To conclude, the jury is still out as far as the size of the pandemic's negative effects for the economy, as well as its long-term consequences, are concerned.

For these reasons, up-to-date analysis of the dynamics of economic phenomena during recent turbulent times poses a very current important research problem for applied economists. One of the key topics concerns behaviour of expectations which, in turn, substantially affects decisions of economic agents. Yet, results of tests performed so far on aggregated macroeconomic data proved to be inconclusive and in high degree dependent on many factors, including the phase of a business cycle. Observed changes reported by respondents constitute a unique in its timelines data source on the current state of the economy. Additionally, the use of entropy and dissimilarity measures in the field of expectation analyses has been relatively rare so far. These two factors combined – unexpected arrival of the pandemic, and lack of unequivocal results on behaviour of economic expectations in critical times – has motivated this study. It aims to verify whether behaviour of business survey expectations and observed changes allows for classification of the pandemic phase as another contraction phase, similar in this respect to other downturns in Polish economy.

The remaining part of the paper is organized as follows. In Section 2, the dataset (i.e. the RIED database on business tendency surveys in manufacturing) is described as well as the empirical methods employed to analyse the dynamics of assessments and expectations across business cycle phases. Section 3 provides a description of the expansion, contraction and pandemic phases of 2009 – 2022 on the basis of descriptive statistics of observed and expected changes in five fields of economic activity, and Section 4 – on the basis of entropy and dissimilarity measures. Section 5 presents a summary of the empirical results and their interpretation in terms of the goals of the study, as well as conclusions and limitations.

2. Data and methods

The data on assessment and expectations concerning major economic variables has been obtained from the monthly business tendency surveys in manufacturing conducted by the Research Institute for Economic Development of SGH Warsaw School of Economics (henceforth RIED) since March 1997. The scope of the survey and variants of the answers are presented in Table 1.

Table 1: Monthly RIED questionnaire in the manufacturing industry

Code	Category	Observed within the last month	Expected for the next 3-4 months
q01	Level of production	up unchanged down	will increase will remain unchanged will decrease
q02	Level of orders	up normal down	will increase will remain normal will decrease
q03	Level of export orders	up normal down not applicable	will increase will remain normal will decrease not applicable
q04	Stocks of finished goods	up unchanged down	will increase will remain unchanged will decrease
q05	Prices of goods produced	up unchanged down	will increase will remain unchanged will decrease
q06	Level of employment	up unchanged down	will increase will remain unchanged will decrease
q07	Financial standing	improved unchanged deteriorated	will improve will remain unchanged will deteriorate
q08	General situation of the economy	improved unchanged deteriorated	will improve will remain unchanged will deteriorate

Source: RIED database.

Eight fields of economic activity are evaluated by the respondents with respect to changes they observe and expect for the next 3-4 months.² On the basis of individual qualitative responses, balance statistics (i.e. differences between the number of optimists – those who report or expect improvement – and pessimists), are calculated and presented in percentage points. Aggregated results and comments are regularly published in the RIED Bulletins (see Adamowicz, Walczyk 2022).

The starting point for empirical analysis is October 2009, when the contraction phase associated with the financial crisis of 2008–09 came to an end. Following Tomczyk (2022), the following phases of business cycle are identified:

- expansion phase of October 2009 – June 2012,
- contraction phase of July 2012 – December 2012,
- expansion phase of January 2013 – February 2020.

In Tomczyk (2022), the last phase ended in December 2019 as it was the final point of database then available. For the purpose of current analysis, expansion phase has been extended until February 2020. Even though in January and February 2020 the first signs of deterioration of the macroeconomic situation and business sentiment emerged, pandemic-related restriction have not been yet introduced in Poland. The first confirmed case of COVID-19 in Europe occurred in France on January 24, 2020, and in Poland – on March 4, 2020. Officially, the state of pandemic has been declared on March 14, 2020 (Regulation of the Minister of Health on the declaration of an epidemic threat in the territory of the Republic of Poland, Journal of Laws of 2020, item 433). It has not yet been revoked, but most of the restrictions, including the obligation to wear masks and of home isolation, border quarantine, and home quarantine for family members, were lifted on March 28, 2022. Consequently, the pandemic phase has been defined as starting in March 2020 and continuing until the end of RIED sample available (February 2022), that is:

- pandemic phase: March 2020 – February 2022.

The variables selected from the RIED questionnaire in manufacturing (see Table 1) are those than can be compared with aggregated Statistics Poland data to quantify survey expectations data for further analysis: q01 (level of production), q05 (prices of goods produced), q06 (level of employment), q07 (financial standing), and q08 (general situation of the economy).

Two sets of methods of empirical analysis of business survey data are employed in this paper. First, averages, medians and standard deviations for both observed and expected changes in balance statistics for selected fields of economic activity surveyed by RIED are calculated in order to measure typical levels and volatility of expectations

² For the purposes of empirical analysis, the 3-month forecast horizon has been selected on the basis of previous studies of the RIED business survey data (Tomczyk 2011).

and assessments of the current situation by manufacturing enterprises during expansion, contraction, and pandemic phases. These results are presented and analysed in Section 3.

Second, entropy and dissimilarity measures are used to evaluate similarities between *a priori* information supplied by business tendency surveys (i.e. expectations), and *a posteriori* information (i.e. realizations). Following Wędrowska (2010), let us define structure S^n as a vector $S^n = [s_1, s_2, \dots, s_n]^T \in R^n$ whose elements s_i ($i = 1, 2, \dots, n$) fulfil two conditions:

$$0 \leq s_i \leq 1, \quad (1)$$

$$\sum_{i=1}^n s_i = 1. \quad (2)$$

Structure S^n is, therefore, fully described by a vector of fractions (structure elements) summing to a total of 1.

The amount of information provided by a message (i.e. its information content) is defined in information theory in relation to the probability that a given message is received from the set of all possible messages: the less probable the message, the more information it carries. On the basis of the elements of S^n it is now possible to define the empirical measure of entropy introduced by C. E. Shannon in his classic 1948 paper *A mathematical theory of communication* as

$$H(S^n) = \sum_{i=1}^n s_i \log_2 \frac{1}{s_i}. \quad (3)$$

It is worth noting that the value of $H(S^n)$ depends only on characteristics of the structure analyzed, i.e. its elements s_i .

An important property of $H(S^n)$ as a measure of entropy is that it reaches its maximum value of $H_{max} = \log_2 n$ if all structure elements s_i are equal (i.e. $s_1 = s_2 = \dots = s_n$). As $H(S^n)$ approaches its maximum value, differences between structure elements decrease, and for $H(S^n) = H_{max}$, the distribution of structure elements becomes uniform. Also, $H(S^n) = H_{min} = 0$ if one of the elements s_i ($i = 1, 2, \dots, n$) is equal to 1, and all the remaining structure elements are equal to 0 (i.e. distribution is concentrated in one element of structure only). The value of $H(S^n)$ can be, therefore, interpreted as the measure of concentration of elements s_i of structure S^n , and can be used in empirical setting to evaluate information content of a structure. When several structures ordered in time are available, it is also possible to analyse their dynamics. Empirical values and dynamics of entropy measure $H(S^n)$ for expectations and realizations expressed in the RIED business tendency surveys are presented in the next section.

In practice, however, not only the degree of uncertainty associated with *a priori* and *a posteriori* structures may be economically interesting but also the extent of changes detected between assumed (*a priori*) and observed (*a posteriori*) structures. In order to

analyse the size of change between *a priori* structure S_p^n and *a posteriori* structure S_q^n , relative entropy (or Kullback-Leibler divergence; see Zhang, Jiang 2008) is calculated:

$$I(S_q^n: S_p^n) = \sum_{i=1}^n q_i \log \frac{q_i}{p_i}. \quad (4)$$

Relative entropy is also known as information gain; it measures expected amount of “new” information provided by *a posteriori* structure. One of the properties of (4) states that it takes its minimum value of zero if both structures are identical (i.e. $S_p^n = S_q^n$), and increases with the size of differences between the structures to infinity (see Wędrowska 2010). $I(S_q^n: S_p^n)$ can be interpreted as the degree of change between assumed (*a priori*) and observed (*a posteriori*) structures, and therefore serve as measure of dissimilarity of structures: the larger it is, the less similar the structures are.

In empirical setting, it is more convenient to apply a standardized coefficient defined on interval $[0, 1]$ to facilitate interpretations and comparisons. Chomątowski and Sokołowski (1978) introduce a similarity measure to classify data into comparable phases, and employ it to define clusters of industrial production in Poland. They also provide a related dissimilarity measure that can be used to evaluate extent of change from *a priori* to *a posteriori* structure:

$$P(S_q^n: S_p^n) = 1 - \sum_{i=1}^n \min(q_i, p_i). \quad (5)$$

From the properties of the structure defined by (1) and (2) it follows that $P(S_q^n: S_p^n) \in [0, 1]$. The lower limit is attained when analysed structures are identical, i.e. $S_p^n = S_q^n$. As dissimilarities between structures increase, value of $P(S_q^n: S_p^n)$ increases towards the upper limit of 1.

Empirical values and dynamics of dissimilarity measure $P(S_q^n: S_p^n)$ employed to evaluate similarities between expectations and realizations expressed in business tendency surveys are presented in Section 4. They are introduced to supplement results obtained on the basis of the entropy measure as both methods reflect structure change from its *a priori* to *a posteriori* state.

3. Statistical properties of observed and expected balance statistics

In Table 2, means, medians and standard deviations for both observed and expected changes in balance statistics for selected fields of economic activity surveyed by RIED are presented, aggregated into expansion and contraction phases. Means and medians of balance statistics measure average level of optimism in each phase across assessments and expectations; standard deviation – its volatility. Also, the average percentages of “no change observed / expected” answers are calculated in each case in order to evaluate the dynamics of no-change (i.e. “everything remains/will remain stable”) responses.

Table 2: Descriptive statistics for observed and expected balance statistics

Variable	Measure	Expansion 2009.10 – 2012.06	Contraction 2012.07 – 2012.12	Expansion 2013.01 – 2020.02	Pandemic 2020.03 – 2022.02
Production Observed	mean	2.97	-9.47	-0.68	-9.85
	median	2.40	-6.20	0.40	-6.75
	std dev	11.58	9.47	9.51	15.07
	avg.unch	47.82	47.67	50.91	47.12
Production Expected	mean	6.33	-11.67	3.21	-5.37
	median	9.70	-12.40	4.95	-2.85
	std dev	11.40	9.02	9.16	18.61
	avg.unch	50.80	49.17	54.00	46.14
Prices Observed	mean	7.10	-3.45	1.45	23.29
	median	6.60	-4.35	-1.80	27.95
	std dev	9.53	3.30	8.46	23.67
	avg.unch	73.16	79.48	77.79	63.38
Prices Expected	mean	8.48	-0.77	5.41	29.78
	median	8.40	-1.30	2.60	31.95
	std dev	7.80	2.43	10.02	24.55
	avg.unch	73.95	77.82	76.70	55.53
Employment Observed	mean	-8.09	-12.80	-1.47	-3.29
	median	-6.90	-11.45	0.00	-0.95
	std dev	6.55	4.36	6.35	6.18
	avg.unch	66.54	67.67	70.22	76.13
Employment Expected	mean	-12.55	-21.53	-2.85	-2.36
	median	-12.00	-22.45	-2.15	-0.60
	std dev	5.79	5.46	5.62	9.22
	avg.unch	70.39	66.37	72.45	71.83
Finances Observed	mean	-7.94	-15.22	-8.74	-20.35
	median	-8.10	-14.40	-8.30	-17.80
	std dev	6.51	2.63	6.20	12.28
	avg.unch	63.48	65.57	64.17	59.45
Finances Expected	mean	-3.75	-19.17	-7.74	-17.23
	median	-3.30	-19.55	-6.90	-11.80
	std dev	5.74	3.95	6.60	16.96
	avg.unch	64.56	63.22	64.70	55.46
General Observed	mean	-21.95	-49.38	-12.88	-52.10
	median	-22.50	-48.85	-11.20	-54.80
	std dev	9.19	5.38	16.04	18.49
	avg.unch	62.12	48.02	62.90	35.13
General Expected	mean	-20.52	-53.30	-16.47	-45.93
	median	-18.10	-53.40	-15.00	-48.60
	std dev	13.19	5.05	13.86	19.76
	avg.unch	57.77	42.77	59.06	33.76

Notation: see Table 1; avg.unch – average percentage of “no change observed/expected” answers.

Source: own calculations on the basis of RIED data.

As far as average level of optimism is concerned, it is much lower as measured by means and medians in the pandemic phase than in the preceding expansion phase. (Let us keep in mind that in the case of prices (q05), higher mean and median signify higher prices observed or expected, which is generally not good news for the economy; higher values are therefore consistent with the interpretation of less optimism in the pandemic phase.) This finding, of course, is hardly surprising and is consistent with results already noted in the literature. For example, Teresiene *et al.* (2021) show widespread pessimism among manufacturing and service sectors enterprises, both locally (by Eurozone countries) and globally. In particular, they document negative and significant impact of COVID-19 infections and fatalities on business sentiment indicators.

In the pandemic phase there are several instances of sizable differences between the average and median, which may suggest that more of the data values are clustered towards one end of their range or a few extreme values are observed. This may suggest more uncertainty during the pandemic. Also, much higher volatility (as measured by standard deviation) in the pandemic phase as compared to the preceding expansion phase of January 2013 – February 2020 is observed, particularly in the case of production (q01), prices (q05), and financial standing of companies (q07) in the manufacturing sector. This stands in contrast with the previous analysis of expectations expressed in Polish business tendency surveys, where more volatility was noted during expansion phases, more often for observed changes than for forecasts, and lower uncertainty was observed in contraction phases (see Tomczyk 2022). It is perhaps a first empirical indication that the pandemic phase cannot be straightforwardly interpreted as another contraction phase. Instead, it seems to be a separate phenomenon not to be confused with previous slumps in economic activity. The ambiguous behaviour of price expectations during pandemic has been noted previously. For example, Meyer *et al.* (2022) note that in the United States, enterprises expect lower selling prices in the short term and lower inflation in contrast to rising household inflation expectations. Generally, behaviour of price expectations during pandemics requires further detailed analysis as it is not typical either for expansion or contraction phases of a business cycle.

There are exceptions to the “higher volatility during pandemics” rule though: standard deviations of observed and expected changes in employment (q06) and general situation of the economy (q08) remain stable across expansion and pandemic phases.

Fractions of “no change” responses are generally lower for both observed and expected changes in the pandemic phase than in the expansion phases directly preceding. This result suggests that survey respondents found it easier to express a specific (and generally pessimistic, judging by means and medians) opinion about all the economic variables. This effect is particularly strong in the case of a general situation of the economy (q08), where percentages of “no change” answers fell from 63–59 percent to 34–35 percent between the expansion phase of January 2013 – February 2020 and pandemic phase of March 2020 – February 2022).

4. Results of application of entropy and dissimilarity measures

As the next step in analysis of assessment and expectations of enterprises across business cycle phases, Shannon's entropy was used for investigating the level of concentration of the structures, defined as percentages of "up – no change – down" answers. Table 3 shows the summary statistics for entropy measure $H(S^n)$ given by formula (3), calculated for five variables selected from the RIED business tendency survey in manufacturing, separately for expectations and observed changes, across business cycle phases. Since mean and median values were very similar, only mean is reported for purposes of clarity.

Table 3: Summary statistics for entropy measures: observed and expected changes

Variable	Measure	Expansion 2009.10 – 2012.06	Contraction 2012.07 – 2012.12	Expansion 2013.01 – 2020.02	Pandemic 2020.03 – 2022.02
Production Observed	spread	0.1143	0.0646	0.1929	0.2927
	mean	1.4978	1.4973	1.4738	1.4717
	std dev	0.0317	0.0260	0.0388	0.0630
Production Expected	spread	0.1188	0.0432	0.1832	0.2899
	mean	1.4654	1.4755	1.4376	1.4750
	std dev	0.0307	0.0169	0.0417	0.0681
Prices Observed	spread	0.3588	0.1088	0.5196	0.5196
	mean	1.0657	0.9291	0.9562	1.0475
	std dev	0.0818	0.0438	0.1054	0.1495
Prices Expected	spread	0.3297	0.2502	0.6480	0.3808
	mean	1.0478	0.9803	0.9652	1.1466
	std dev	0.0772	0.0842	0.1373	0.0837
Employment Observed	spread	0.2317	0.0304	0.3085	0.2990
	mean	1.2280	1.1899	1.1633	1.0129
	std dev	0.0601	0.0125	0.0667	0.0821
Employment Expected	spread	0.2326	0.1594	0.3440	0.2336
	mean	1.1226	1.1395	1.1113	1.1139
	std dev	0.0534	0.0700	0.0721	0.0680
Finances Observed	spread	0.2481	0.1354	0.2495	0.2995
	mean	1.2890	1.2198	1.2737	1.2655
	std dev	0.0552	0.0524	0.0522	0.0742
Finances Expected	spread	0.1809	0.0249	0.2692	0.3472
	mean	1.2813	1.2380	1.2657	1.3305
	std dev	0.0383	0.0098	0.0520	0.0754
General Observed	spread	0.1806	0.1428	0.3681	0.9566
	mean	1.2183	1.0767	1.2305	1.1643
	std dev	0.0487	0.0496	0.0833	0.2386
General Expected	spread	0.2646	0.1631	0.3329	0.8692
	mean	1.2940	1.1021	1.2925	1.2617
	std dev	0.0668	0.0591	0.0892	0.2119

Notation: see Table 1; obs – observed changes, exp – forecasted (expected) changes; spread = maximum – minimum value. Source: own calculations on the basis of RIED data.

High mean values of entropy obtained in the case of production (q01), both in comparison to other variables and in absolute terms, seem to be the most striking result. The maximum value of measure of entropy is $H_{max} = \log_2 n = \log_2 3 = 1.5850$; the closer empirical entropy of a structure to its maximum value, the more uniform the structure is, and therefore the less informative *a priori* structure becomes in relation to *a posteriori* structure. Mean values obtained for production across the business cycle phases (for example, 1.4750 for expectations and 1.4717 for realizations during the pandemic phase) are considerably higher than mean entropy of prices, employment, financial standing, or general business conditions. In the case of production, therefore, distribution of increase / no change / decrease fractions is relatively uniform, leading to high entropy and providing little information. On the other hand, entropy is equal to zero if one of the elements of a structure is equal to 1, i.e. there is no uncertainty associated with distribution of outcomes. The value of zero is not attained for any of the variables analysed, and the lowest values (slightly above or below 1) are observed for prices (q05). Since entropy allows to evaluate degree of concentration, in the case of prices fractions of survey answers seems to be particularly cantered on one of the three options provided in the questionnaire. In theory, answers might be cantered on either of the three options (increase / no change / decrease) and vary from one questionnaire to another. In practice, however, they are heavily biased towards the “no change” category (see Table 2) for all the variables in the RIED business tendency survey.

Compared to any other phase of the business cycle since 2009, the highest spread of entropy (i.e. difference between the maximum and minimum values) is observed in the pandemic phase, with a single exception of employment (q06), for which the highest spread, for both observed and expected changes, was noted during the expansion phase of January 2013 – February 2020. The largest increase as compared with the last expansion phase, the largest spread is noted for general condition of the economy (q08), where spread increases from 0.3881 to 0.9566 for observed and from 0.3329 to 0.8692 for expected changes. Also, in the case of the general situation of the economy, there is the most dramatic increase in variability of entropy as measured by standard deviation: from 0.0833 to 0.2386 for observed and from 0.0892 to 0.2119 for expected changes, confirming that the general situation of the economy is subject to the most volatile changes in information content of surveys from one month to another.

Finally, dissimilarity measure $P(S_q^n: S_p^n)$ given by equation (5) is used to quantify the divergence between the *a priori* and *a posteriori* structures; i.e. expectations and observed changes. Since expectations have to be matched with observed realizations to calculate the measure of dissimilarity, the length of time series is reduced by three observations. The final phase (pandemic) is therefore reported for March 2020 – November 2021 since the last three expectations data points (for December 2021, and January and February 2022) do not have matching observed changes to calculate the dissimilarity statistics. Statistical details, i.e. mean, median and standard variation across business cycle phases, are reported in Table 4.

Table 4: Summary statistics for dissimilarity measure (5)

Variable	Measure	Expansion 2009.10 – 2012.06	Contraction 2012.07 – 2012.12	Expansion 2013.01 – 2020.02	Pandemic 2020.03 – 2021.11
Production	mean	0.0763	0.0693	0.0800	0.0765
	median	0.0710	0.0655	0.0715	0.0690
	min	0.0240	0.0300	0.0140	0.0110
	max	0.1280	0.1420	0.3140	0.2610
	std dev	0.0331	0.0411	0.0474	0.0584
Prices	mean	0.0490	0.0278	0.0495	0.0723
	median	0.0470	0.0285	0.0385	0.0670
	min	0.0070	0.0080	0.0040	0.0010
	max	0.1410	0.0540	0.2620	0.1660
	std dev	0.0299	0.0180	0.0432	0.0480
Employment	mean	0.0571	0.0457	0.0487	0.0520
	median	0.0590	0.0470	0.0490	0.0430
	min	0.0060	0.0260	0.0020	0.0150
	max	0.1180	0.0730	0.1310	0.1090
	std dev	0.0275	0.0174	0.0271	0.0293
Finances	mean	0.0430	0.0308	0.0475	0.0941
	median	0.0440	0.0340	0.0380	0.0770
	min	0.0030	0.0140	0.0020	0.0310
	max	0.0910	0.0400	0.2480	0.2570
	std dev	0.0215	0.0098	0.0386	0.0507
General	mean	0.0753	0.0532	0.0765	0.1156
	median	0.0710	0.0430	0.0670	0.1000
	min	0.0190	0.0160	0.0070	0.0380
	max	0.1680	0.0980	0.3620	0.2390
	std dev	0.0413	0.0321	0.0587	0.0589

Notation: see Table 1. Source: own calculations on the basis of the RIED data. Min – minimum value, max – maximum value.

The highest mean value of the dissimilarity measure is observed in the divergence between the *a priori* and *a posteriori* structures of the expectations and assessments of general business conditions during the pandemic: 0.1156. It is the global maximum across all the business cycle phases and all variables. The majority of the values of the dissimilarity measure indicate only minor divergences between the analysed structures. The lowest means and medians are generally observed for prices (q05) with the global minimum of 0.0278 during the short contraction phase of July 2012 – December 2012, but, during the pandemic phase, the lowest value (0.0520) is associated with employment (q06). Volatility of dissimilarity of structures is consistently higher during pandemic than in any other business cycle phase with the maximum of 0.0589 for the general economic situation (q08) and a close second high of 0.0584 for production

(q01). These results confirm that for general situation of the economy, *a posteriori* structures of responses differ significantly from their *a priori* counterparts and, with their high variability, reflect a lot of uncertainty among the respondents as far as overall economic conditions are concerned.

Employment (q06) stands out as the only variable which exhibits constant sizable variation across the entire sample without the peak at the beginning of 2020. This result can be interpreted as structures (i.e. percentages of increase/ no change/ decrease answers) for employment expectations and assessments being relatively unaffected by the onset of the pandemic. Previous studies (for example, Bartik et al. 2020) show that businesses' expectations about the longer-term impact of COVID-19 on employment strongly depend on sector's familiarity with pandemic-relief programs and government assistance procedures. Lack of this type of data in case of Polish business surveys may explain the relatively uniform behaviour of the dissimilarity measure in the case of employment.

5. Summary and conclusions

COVID-19 pandemics is the second major event of this type in current (economic) memory, the first being the Spanish flu pandemic of 1918-1920, the most severe pandemic in modern history. However, recent abundance of research articles on economic consequences of COVID-19 outbreak is unprecedented in the post-pandemic economic literature. The Spanish flu pandemic, caused by an H1N1 virus of avian origin, is estimated by the Center for Disease Control and Prevention to have affected about 500 million people (one-third of the world's population) and caused at least 50 million deaths worldwide (CDC, 2021). As for COVID-19, there are more than 504 million registered cases and 6.2 million deaths as of April 17, 2022 (Worldometer, 2022). Boianovsky and Erreygers (2021) note that despite of the huge scale of the Spanish flu pandemic, none of the major economics journals published an article on the pandemic in the period of 1918–1921. As possible reasons they cite factors related to the organization of the economic profession, lack of nation-wide anti-pandemic government measures such as lockdowns, and generally low degree of visibility of the economic characteristics of the pandemic; let us note it took place right after the World War I when academic and publishing efforts were not given priority. The interest taken by 20th century economists in analysing economic consequences of the COVID-19 pandemic is fully warranted by remarkable irregularities of business tendency survey data dynamics in comparison to other phases of a business cycle. This is particularly visible for the general business conditions and much less so for individual variables such as production, prices, employment and financial standing of manufacturing sector companies.

On the basis of empirical results presented in Sections 3 and 4, the following overall conclusions can be drawn. General business conditions (q08) consistently stand out as the variable associated with the highest volatility and evident difficulties with predicting *a posteriori* structures of responses on the basis of *a priori* information. This category exhibits the highest discrepancies across characteristics of expectations and assessments, ranging from non-informative to relatively informative structures. Also, the largest global values of the dissimilarity measure are observed in the divergence between the *a priori* and *a posteriori* structures of general business conditions during the pandemic. Fractions of expected and observed percentages differ markedly, reflecting respondents' insecurity with respect to general economic conditions.

During COVID-19, higher volatilities of expected and observed changes in responses are noted, which in previous analyses (see Tomczyk 2022) were associated rather with expansion phases of the business cycle. Combined with results of the application of entropy and dissimilarity measures, this finding strongly suggests that the pandemic phase cannot be straightforwardly interpreted as another contraction in the business cycle. Instead, it seems to be a separate phenomenon not to be confused with previous slumps in economic activity.

Analysis of expectations of entrepreneurs during recent pandemic should be further extended to include tests of rationality of expectations with respect to the cycle phase (including the pandemic phase separately from typical upturns or downturns) or correlation of sentiments expressed in tendency surveys with other aggregated measures of economic activity. Special attention should be paid to general business situation as this variable exhibits the most traits separating the pandemic from other phases of the business cycle. However, since we have only two year's worth of pandemic-related data (and hopefully no more), the limited number of observations will make quantification procedures statistically dubious. What is more, the 2022 Russian invasion of Ukraine and the resulting war that continues at the time of writing of this article will likely further distort empirical results.

An additional research problem worth considering is whether current situation of an enterprise (particularly its financial standing reported in question q07) systematically influences its expectations, and consequently the degree of concentration of answers on a particular option.

Still another approach, significantly extended with respect to the current research project centered on properties of business survey data in separate business cycle phases, would focus on entropy and dissimilarity measures disaggregated by years or quarters. It would allow to analyse their variability in more detail, and seems to be particularly suitable during expansion phases which tend to be relatively long as compared to contractions. This research question remains as one of the promising directions of further study of statistical and information content properties of business tendency survey data.

References

- Adamowicz, E., Walczyk, K., (2022). Koniunktura w przemyśle. Luty 2022, *Badanie okresowe, nr 401*, Instytut Rozwoju Gospodarczego, Szkoła Główna Handlowa, Warszawa.
- Bartik, A. W., Bertrand, M., Cullen, Z., Glaeser, E. L., Luca, M., Stanton, C., (2020). The impact of COVID-19 on small business outcomes and expectations, *PNAS*, vol. 117 (30), pp. 17656–17666.
- Boianovsky, M., Erreygers, G., (2021). How Economists Ignored the Spanish Flu Pandemic in 1918–1920, *Erasmus Journal for Philosophy and Economics*, vol. 14, no. 1.
- CDC, (2021). 1918 Pandemic (H1N1 virus), <https://www.cdc.gov/flu/pandemic-resources/1918-pandemic-h1n1.html> [access: April 17, 2022].
- Chomętowski, S., Sokołowski, A., (1978). Taksonomia struktur, *Przegląd Statystyczny*, vol. 2, pp. 217–226.
- GUS, (2022). *Business tendency in manufacturing, construction, trade and services 2000–2022*, GUS (Statistics Poland), Warszawa.
- Meyer, B. H., Prescott, B., Sheng, X. S., (2022). The impact of the COVID-19 pandemic on business expectations, *Journal of Forecasting*, vol. 38, pp. 529–544.
- Shannon, C. E., (1948). A mathematical theory of communication, *The Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656.
- Teresiene, D., Keliuotyte-Staniuleniene, G., Liao, Y., Kanapickiene, R., Pu, R., Hu, S., Yue, X.-G., (2021). The Impact of the COVID-19 Pandemic on Consumer and Business Confidence Indicators, *Journal of Risk and Financial Management*, vol. 14(4), p. 159.
- Theil, H., (1967). *Economics and Information Theory*, North-Holland Publishing Company, Amsterdam.
- Tomczyk, E., (2011). *Oczekiwania w ekonomii: idea, pomiar, analiza*, Szkoła Główna Handlowa, Warszawa.
- Tomczyk, E., (2022). Do Survey Responses in Manufacturing Fluctuate with Business Cycle? Evidence from Poland, w: Białowas S. (red.) *Economic tendency surveys and economic policy - measuring output gaps and growth potentials*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, pp. 71–84.

- Van Der Wielen, W., Barrios, S., (2021). Economic sentiment during the COVID pandemic: Evidence from search behaviour in the EU, *Journal of Economics and Business*, vol. 115.
- Wędrowska, E., (2010). Oczekiwana ilość informacji o zmianie struktur jako miara niepodobieństwa struktur, *Acta Universitatis Nicolai Copernici*, vol. 397, pp. 99–109.
- The World Bank, (2021). Polska gospodarka powraca na ścieżkę wzrostu, pomimo problemów z pandemią, Polish Economy Returns to Growth Amidst Pandemic-Related Setbacks (worldbank.org) [access: April 17, 2022].
- Worldometer, (2022). *COVID-19 Coronavirus Death Toll*, <https://www.worldometers.info/coronavirus/coronavirus-death-toll/> [access: April 17, 2022].
- Zhang, Q.-S., Jiang, Y.-J., (2008). A note on information entropy measures for vague sets and its applications, *Information Sciences*, vol. 178, pp. 4184–4191.

Breaking Benford’s law: a statistical analysis of COVID-19 data using the Euclidean distance statistic

Leonardo Campanelli¹

ABSTRACT

Using the Euclidean distance statistical test of Benford’s law, we analyse the COVID-19 weekly case counts by country. While 62% of the 100 countries and territories considered in the present study conforms to Benford’s law at a significant level of $\alpha = 0.05$ and 17% at a significant level of $0.01 \leq \alpha < 0.05$, the remaining 21% shows a deviation from it (p values smaller than 0.01). In particular, 5% of the countries ‘break’ Benford’s law with a p value smaller than 0.001.

Key words: Benford’s law, COVID-19 data.

1. Introduction

At the end of the 19th century, Newcomb (1881) noticed that the first-digit distribution of logarithms were not uniform, as one would expect, but rather followed the rule

$$P_B(d) = \log\left(1 + \frac{1}{d}\right), \quad (1)$$

where $P_B(d)$ is the probability of the first significant digit d . About 60 years later, Benford (1938) rediscovered Newcomb’s rule (hereafter Benford’s law), extended the law to arbitrary logarithmic bases and to multiple digits, and successfully tested the law against 20 very different data sets, like physical constants, deaths rates, populations of cities, length of rivers, etc.

Although it is now known that some distributions satisfy Benford’s law [see Morrow (2014) and references therein] and that particular principles lead to the emergence of the Benford phenomenon in data (Hill, 1995a, 1995b, and 1995c), no general criteria has been found that fully explain when and why Benford’s law holds for a generic set of data. Although much work is still needed to understand the theoretical basis of the law, the number of its applications has grown in the last few decades [for theoretical insights and general applications of Benford’s law, see Miller (2015)].

Probably, the most famous applications are to detecting tax (Nigrini, 1996), campaign finance (Cho and Gaines, 2007), and election (Roukema, 2013) frauds. Other interesting applications are in image processing (Pérez-González et al., 2007), where Benford’s law can be used to look for hidden messages in pictures as well as to test whether or not the

¹ All Saints University School of Medicine, Toronto, Canada. E-mail: leonardo.s.campanelli@gmail.com.
ORCID: <https://orcid.org/0000-0002-7200-9990>.

image has been compressed, and in natural sciences, where the law has been shown to hold for geophysical observables such as the length of time between geomagnetic reversals, depths of earthquakes, models of Earth's gravity, and geomagnetic and seismic structure (Sambridge et al., 2010).

In general, however, it is important to stress that the rejection/acceptance of tests on data whose underlying distribution is not known to follow Benford's law should not be used as a tool to uncover error or, more importantly, fraud. This is particularly true for COVID-19 data since there is no theoretical basis or sufficient empirical evidence that these data follow a Benford distribution.

The first application of Benford's law to the study of COVID-19 data, in particular to daily and cumulative case and death counts, is due to Sambridge and Jackson (2020), while the most recent work on the 'Benfordness' of COVID-19 data is by Farhadi (2021). Using different statistical tests, the authors of both studies conclude that, in general, COVID-19 data conform to a Benford's distribution and also indicate 'anomalies' in the data of some countries. The results of these and similar analyses, however, cannot be completely trusted for reasons discussed in Section 2. Here, we will describe the statistical approach used to test the compliance of COVID-19 data with Benford's law.

Our goal is, indeed, to show that, in general, the first-digit distribution of COVID-19 (weekly) case counts by country do conform to Benford's law. This opens the possibility of detecting, in a statistically robust way, anomalous deviations from the law by specific countries. Our results, presented in Section 3, will be discussed in Section 4. In Section 5, we draw our conclusions.

2. Data and Method

It is well known that the compliance of data sets with Benford's law improves as the range of the data increases. Daily confirmed cases and daily death cases are then not appropriate when checking for the compliance of COVID-19 first-digit distributions with Benford's law because they typically extend over very few orders of magnitude. Another possibility would be the use of cumulative data. The disadvantage of using this type of data is that as cumulative cases numbers begin to flatten (especially after a COVID-19 'wave' has passed), first digits tend to become all the same, thus distorting relative digit frequencies. In order to overcome the above problems for COVID-19 data, we will only analyse the data on weekly confirmed cases by country: they extend, at least for about 45.0% of the countries, over 4 order of magnitudes, and do not flatten.

Data are from the World Health Organization (WHO, 2021) and are updated to December 20, 2021 (two years from the start of the pandemic). Counts collected by WHO reflect laboratory-confirmed cases and include both domestic and repatriated cases. True cases are subject to a time-variable under/overestimation since case definition, case detection, testing strategies, and reporting practice differ among countries, territories, and areas. All counts are continuously verified by WHO and then may change based on retrospective updates necessary to reflect changes in case definition and/or reporting practices.

Of the 222 countries and territories affected by COVID 19, only 100 have COVID-19 weekly case counts with range spanning 4, or more, orders of magnitude. These countries

and territories are shown in Table 1 and, following the WHO's convention (WHO, 2021), are grouped in six different regions: Africa, Americas, Eastern Mediterranean, Europe, South-East Asia, and Western Pacific. Also shown in the table is the range of weekly cases, $[N_{\min}, N_{\max}]$, and the number of weeks, N .

The most common tests in use for testing whether an observed sample of size N satisfies Benford's law are the Pearson's χ^2 , Kolmogorov-Smirnov, and Kuiper tests. However, such tests are based on the null hypothesis of a continuous distribution, and are generally conservative for testing discrete distributions as the Benford's one (Noether, 1963). This problem can be overcome if one uses the results by Morrow (2014) who has recently found asymptotically valid test values for these statistics under the specific null hypothesis that Benford's law holds.

Other tests have been recently proposed, based on new statistics such as the 'max' statistic, m , introduced by Leemis et al. (2000), and the 'normalized Euclidean distance' statistic, d^* , introduced by Cho and Gaines (2007). At the moment of their introduction, however, the properties of the corresponding estimators were not well understood and no test values were reported. These problems were solved by Morrow (2014), who provided asymptotically test values for those statistics too.

Recently enough (Campanelli, 2021), we have found, by means of Monte Carlo simulations, the (empirical) cumulative distribution function (CDF) of the 'Euclidean distance' statistic, d_N^* , which is based on the statistic d^* and was introduced by Morrow (2014). It is defined as

$$d_N^* = \sqrt{N \sum_{d=1}^9 [P(d) - P_B(d)]^2}, \quad (2)$$

where $P(d)$ is the observed first-digit frequency distribution.

In the following, we will use this statistic to study the first-digit distribution of COVID-19 weekly case counts by country since this is the only statistic, among the ones discussed before and analysed by Morrow, with known distribution. In particular, we will use its CDF to evaluate p values as $p = 1 - \text{CDF}(d_N^*)$.

3. Results

Our results are presented in Table 1 where we show, for each country, the Euclidean distance score, d_N^* , and the corresponding p value. Notice that the CDF of d_N^* , and then the p values, are reliable up to the second decimal place if $0.28 < d_N^* < 1.85$ and up to the third decimal place otherwise (Campanelli, 2021). In the first case, the uncertainty on p is ± 0.001 , while in second case is ± 0.0001 . In Table 1, the last digits in parentheses refer to these errors. For example, $p = 0.27(4)$ stands for $p = 0.274 \pm 0.001$, while $p = 0.000(2)$ stands for $p = 0.0002 \pm 0.0001$.

As shown in Figure 1, while the great majority of the countries (79%) conform to Benford's law ($p \leq 0.01$), 5% of them show a large deviation from it, having p values smaller than 0.001.

In Figure 2, we show the observed first-digit frequency distributions of weekly case

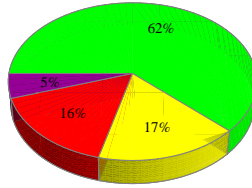


Figure 1: Percentages of the countries in a given range of p values of the Euclidean distance statistic for the first-digit distribution of COVID-19 weekly case counts by country: from top and clockwise, $p \geq 0.05$ (green), $0.01 \leq p < 0.05$ (yellow), $0.001 \leq p < 0.01$ (red), and $p < 0.001$ (purple).

counts for 15 selected countries superimposed to Benford's law. Represented countries are China (where the pandemic started), the United States of America (with the largest total number of cases), India (with the largest range of weekly case counts, N_{\max}/N_{\min}), Tanzania (with the smallest sample size N), Mauritius (with the smallest total number of cases), Algeria (with the smallest range of weekly case counts), Vietnam, Thailand, and Poland (the outliers in the first box plot of Figure 5 with the world largest p values), Honduras, Qatar, Belarus, Cuba, and Egypt (with the smallest p values), and Canada (with the smallest p value in the interval $0.001 \leq p < 0.01$). It is worth noticing that, although the first six countries in Figure 1 have very disparate statistical properties (such as sample size, total number of cases, and range of weekly cases), they all conform to Benford's law at a significant level of 0.01 (excluding Mauritius and Algeria, the other four countries conform to Benford's law at a significant level of 0.05). Moreover, the spatial distribution of p values is quite uniform, in the sense that countries with either large or small p values seem to be evenly scattered in the world, as shown in Figure 3. This suggests that there is no correlation between p value and geographical location of a country.

A better understanding of such a spatial variation of p values can be obtained by analysing the percentage of the countries in a given range of p values for each of the six regions of the world as defined by the WHO (2021). The result is shown in Figure 4. As it is clear from the pie charts, Africa conforms very well to Benford's law, all countries in this region having p values larger than 0.01. Also, South-East Asian and Western Pacific countries conform well to Benford's law, the only countries with a p value less than 0.01 being Maldives (for South-East Asia), and Philippines and Australia (for Western Pacific). Countries in Americas (Eastern Mediterranean), instead, show the largest deviation from Benford's law: only about 41% (53%) of them have p values bigger than 0.05, while about 12% (13%) have p values below 0.001.

To gain further insight into the 'global distribution' of p values, we present the box-and-whisker plots for the p values of all countries (the world) and the countries in the six WHO regions in Figure 5. All distributions are positively skewed, with medians well below 0.5. This indicates that the first-digit distribution of COVID-19 weekly case counts by country

deviates somehow from Benford's law on a 'global' scale.² Such a deviation is, however, to be expected for the reasons explained in Campanelli (2021). Indeed, Benford's law does not represent a true law of numbers: some distributions can be 'close' to but not exactly Benford's, and this regardless of data quality; also, Benford's law emerges in the limit of infinite range of the underlying distribution, condition which is never realized in practice.

4. Discussion

The results of our analysis show that the conformance to Benford's law cannot be rejected at a significant level of 0.01 for most of the countries (79%). This implies that (i) the first-digit distribution of COVID-19 weekly case counts by country follows Benford's law and, accordingly, (ii) it can be used to detect possible 'anomalies' in COVID-19 count data. Thus, in our case, data from Canada, Jordan, Puerto Rico, Greece, Philippines, Belgium, Tunisia, Latvia, Paraguay, Sweden, Guatemala, Pakistan, Kazakhstan, Maldives, Australia, and Russia show a possible anomalous behaviour ($0.001 \leq p < 0.01$), while anomalies are certainly present in the data of Honduras, Qatar, Belarus, Cuba, and Egypt ($p < 0.001$).³

Needless to say, the origin of such anomalies cannot be revealed by our statistical analysis, and further and specific investigations are needed to understand if the anomalous behaviour is the result of data manipulation or other factors.

One possibility in this direction is to look at a potential correlation between the Global Health Security Index (GHSI) and the level of Benfordness (Farhadi, 2021). The GHSI was introduced by the Johns Hopkins University (2021) and is an index of the capabilities of a country to respond to epidemics of potential international concern. Accordingly, one would expect that countries with a high GHSI score are prone to reporting reliable COVID-19 data, and then to conforming to Benford's law, while countries with a low GHSI are not.

However, our results shows that this is not the case. Indeed, countries with a very low GHSI, like the African countries, comply well with Benford' law, while advanced countries with very high GHSI scores, like Australia, Canada, and Sweden (ranked 4, 5, and 7, respectively) show a statistically significant deviation from it. Indeed, a global analysis of the GHSI scores versus the p values of the Euclidean distance statistic clearly indicates a lack of (positive) correlation between GHSI scores and p values. This, together with the fact the the great majority of the countries comply with Benford's law, suggests that non-

²Such a deviation can be quantified by a Kolmogorov-Smirnov (KS) statistical test for the distribution of p values, whose CDF is $\text{CDF}(p) = p$. The values (degrees of freedom) of the KS statistic for all countries and the ones in the six regions are 0.4295 (100), 0.4487 (13), 0.6165 (17), 0.6567 (15), 0.3379 (38), 0.5208 (8), and 0.3639 (9), respectively. Accordingly, conformance to Benford's law is rejected at a significant level of 0.001 (Facchinetti, 2009) in the case of all countries, and the countries in Americas, Eastern Mediterranean, and Europe. It is rejected at a significant level of 0.01 for the African countries. It is not rejected at a significant level of 0.01 for the South-East Asian countries, and it is not rejected at a significant level larger than 0.20 for the case of the Western Pacific countries.

³According to Sambridge et al. (2010) there is evidence that, in general, infection diseases conform to Benford's law. Indeed, they found that the total numbers of cases of 18 infectious diseases reported to the WHO by 193 countries worldwide in 2007 follow a Benford distribution. However, their result was not supported by a goodness-of-fit analysis. Using the Euclidean statistical test, we find that the null hypothesis of conformance to Benford's law cannot be rejected at a significant level of 0.01 (Campanelli, 2022). [In particular, the dynamic range of the data is $N_{\max}/N_{\min} = 10^6$, the number of data points is $N = 987$, and the Euclidean distance score is $d_N^* = 1.419$, corresponding to a p value of $p = 0.02(7)$.]

compliance might indicate a possible data manipulation. In Table 1, we report the GHSI score and rank for each country [data are from Farhadi (2021)], while in Figure 6 we show the p values of the Euclidean distance statistic versus the corresponding GHSI scores (the number of countries with known GHSI score is $N = 91$). The Pearson's coefficient and the corresponding p value are $r = 0.06$ and $p = 0.0166$, respectively, indicating that the very weak positive correlation between GHSI score and Benfordness is unlikely (at a significant level of 0.01).

Related works. – Our results about the Benfordness of COVID-19 data are in disagreement with those obtained by Sambridge and Jackson (2020) for some specific countries. This is not surprising since Sambridge and Jackson analysed cumulative counts which are expected to deviate from Benford's law after a COVID-19 wave has passed (see discussion above). Moreover, their analysis was not statistically robust being only based on the evaluation of the Pearson correlation coefficient r between observed and expected counts. In particular, Sambridge and Jackson found that out of the 53 analysed countries none had a very weak correlation with Benford's law ($|r| \leq 0.20$), and only China had weak correlation ($0.20 < |r| \leq 0.40$). In contrast, our analysis clearly shows that China conforms well to Benford's law, the p values for the Euclidean distance statistic being 0.81(4). Also, Qatar and Greece were found to have moderate ($0.40 < |r| \leq 0.60$) and strong correlation ($0.60 < |r| \leq 0.80$), respectively, while our results show that these two countries do not conform to Benford's law to a high significant level [their p values are 0.000(0) and 0.00(2), respectively]. Moreover, Egypt, Australia, Canada, Russia, Belgium, and Sweden presented very strong correlation ($0.80 < |r| \leq 1.00$) with Benford's law, while our result is that conformance to Benford's law for these countries can be rejected at a significant level less than 0.01.

Wei and Vellwock (2020) analysed cumulative and daily cases, and cumulative and daily deaths, from 20 countries. Their goodness-of-fit test was based on the normalized Euclidean distance estimator, d^* , defined as

$$d^* = \frac{1}{D} \sqrt{\sum_{d=1}^9 [P(d) - P_B(d)]^2}, \quad (3)$$

where $D = \sqrt{\sum_{d=1}^8 P_B^2(d) + [P(9) - 1]^2} \simeq 1.03631$ is a normalization factor that assures that the normalized Euclidean distance is bounded by 0 and 1. The measure of fit to check concordance with Benford's law was taken to be the one proposed by Goodman (2016), according to which compliance with Benford's law occurs when $d^* \leq 0.25$. However, such a rule of thumb has been shown to be statistically unfounded in Campanelli (2021) and, generally, gives untrustworthy results for a number of data points either much less or much bigger than 40 (in particular the rule has a very low statistical power for a number of data points $N \gg 40$.) In fact, in Wei and Vellwock (2020), cumulative cases for Italy, Spain, and U.K. gave high d^* scores (0.50, 0.45, and 0.32, respectively), while our result is that conformance to Benford's law for these countries cannot be rejected at a significant level of 0.05. Moreover, while our analysis rejects the null at a significant level less than 0.01 for Russia, Philippines, and South Africa, Wei and Vellwock found full conformance to Benford's law,

the d^* values for these countries being well below 0.25 (0.20, 0.11, and 0.10, respectively). Finally, daily cases for Philippines, Pakistan, South Africa, and Belgium, were found to have low d^* scores (0.12, 0.07, 0.06, and 0.05, respectively), while our analysis reject conformance to Benford's law for these countries at a significant level less than 0.01. Finally, it is worth noticing that, in our case, the use of d^* statistic together with Goodman's rule-of-thumb would give a highly questionable compliance with Benford's law for all countries excepted Honduras ($d^* = 0.260$) and Tanzania ($d^* = 0.251$).

Recently enough, Farhadi (2021) has performed a detailed analysis of COVID-19 data from 153 countries by using two standard statistical tests – the Kolmogorov-Smirnov and χ^2 tests (both based on a 0.05 significance level) – and the Goodman's rule of thumb for the normalized Euclidean statistic. In his analysis, he combined daily case counts, daily deaths, and daily 'new tests', the latter variable indicating the number of individuals identified for being contaminated with COVID 19. Farhadi found that 27% of the countries showed 'full conformance' to Benford's (the null hypothesis of compliance with Benford' law was not rejected by the three statistical tests), 69% a 'partial conformance' (the null was rejected by only one of the three statistical tests), while 4% of the countries did not conform to Benford's law (the null was rejected by all three statistical tests).⁴ Qualitatively speaking, then, Farhadi's findings agree with our conclusions that the first-digit distribution of COVID-19 counts follows Benford's law and can be used to flag anomalies. However, the method and data used by Farhadi (based on a combination of daily cases, deaths, and new tests) differ substantially from ours and a direct quantitative comparison with his results is not statistically feasible.

5. Conclusions

We have analysed the COVID-19 weekly case counts by country, as provided by the World Health Organization, updated to December 20, 2021. We worked under the null hypothesis that the first-digit distribution of those counts follows a Benford's distribution. The choice of weekly confirmed cases instead of daily ones came from the requirement of having counts that extended over many orders of magnitudes so to improve the compliance of the data sets with Benford's law. For the same reason we did not consider daily and weekly death counts. Also, cumulative cases were not considered as their numbers flatten (especially at the end of a 'wave'), thus distorting relative digit frequencies. Out of the 222 countries affected by COVID 19, we considered only the ones with weekly counts spanning at least 4 orders of magnitude. This choice reduced the study to the analysis of the data from 100 countries and territories. In order to test the null hypothesis, we used the Euclidean distance test introduced in Morrow (2014) and developed in Campanelli (2021), which avoids the specific problems introduced by other statistical tests.

Our analysis shows that the majority of the countries (62%) conforms to Benford's law at a significant level of 0.05. However, 5% of the countries (Honduras, Qatar, Belarus, Cuba, and Egypt) 'break' Benford's law with p values smaller than 0.001.

⁴As expected at the light of our discussion about the Goodman's rule of thumb, Farhadi found that, while the Kolmogorov-Smirnov and χ^2 tests produce similar results, the Goodman's rule of thumb was too conservative to signal anomalies in the distributions of the first digit.

References

- Benford, F., (1938). The Law of Anomalous Numbers. *Proceedings of the American Physical Society* 78, pp. 551–572.
- Campanelli, L., (2021). On the Euclidean Distance Statistic of Benford's Law. *Communications in Statistics - Theory and Methods*. DOI: 10.1080/03610926.2022.2082480.
- Campanelli, L., (2022). Testing Benford's Law: from small to very large data sets. Submitted to *Spanish Journal of Statistics*.
- Cho, W. K. T., Gaines, B. J., (2007). Breaking the (Benford) Law: Statistical Fraud Detection in Campaign Finance. *Am. Stat.* 61, pp. 218–223.
- Facchinetti, S., (2009). A procedure to find exact critical values of Kolmogorov-Smirnov test, *Ital. J. Appl. Stat.* 21, pp. 337–359.
- Farhadi, N., (2021). Can we rely on COVID-19 data? An assessment of data from over 200 countries worldwide. *Sci. Prog.* 104, pp. 1–19.
- Goodman, W., (2016). The promises and pitfalls of Benford's law. *Significance* 13, pp. 38–41.
- Hill, T. P., (1995a). The significant-digit phenomenon. *Am. Math. Mon.* 102, pp. 322–327.
- Hill, T. P., (1995b). Base-invariance implies Benford's law. *Proc. Am. Math. Soc.* 123, pp. 887–895.
- Hill, T. P., (1995c). A statistical derivation of the significant-digit law. *Stat. Sci.* 10, 354–363.
- International Health Regulations, (2005). The document can be observed at <https://www.who.int>.
- John Hopkin University, (2021). <https://www.ghsindex.org>.
- Leemis, L. M., Schmeiser, B. W., Evans, D. L., (2000). Survival Distributions Satisfying Benford's Law. *Am. Stat.* 54, pp. 236–241.
- Miller, S. J. (ed.), 2015. *Benford's Law: Theory and Applications*. Princeton. Princeton University Press.
- Morrow, J., (2014). *Benford's Law, Families of Distributions and a Test Basis*. London: Centre for Economic Performance.

- Newcomb, S., (1881). Note on the frequency of use of different digits in natural numbers. *Am. J. Math.* 4, pp. 39–40.
- Nigrini, M., (1996). A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association* 18, pp. 72–91.
- Noether, G. E., (1963). Note on the Kolmogorov statistic in the discrete case. *Metrika* 7, pp. 115–116.
- Pérez-González, F., Abdallah, C. T., Heileman, G. L., (2007). Benford's Law in Image Processing. IEEE International Conference on Image Processing, pp. 405–408.
- Roukema, B. F., (2013). A first-digit anomaly in the 2009 Iranian presidential election. *J. Appl. Stat.* 41:1, pp. 164–199.
- Sambridge, M., Tkalčić, N., Jackson, A., (2010). Benford's law in the natural sciences. *Geophys. Res. Lett.* 37, L22301.
- Sambridge, M., Jackson, A., (2020). National COVID numbers - Benford's law looks for errors. *Nature* 581, p. 384.
- Wei, A. and Vellwock, A. E., (2020). Is COVID-19 data reliable? A statistical analysis with Benford's law. DOI: 10.13140/RG.2.2.31321.75365/1.
- World Health Organization, (2021). <https://www.covid19.who.int>.

Appendices

Table 1: The Euclidean distance d_N^* in Eq. (1) and its corresponding p value for the first-digit distribution of Covid-19 weekly case counts for 100 countries. Also indicated are the range of cases, $[N_{\min}, N_{\max}]$, the number of weeks, N , and the GHSI score and rank. Counts are from WHO (2021) and are updated to December 20, 2021. (Digits in parentheses at the third and fourth decimal places indicate a statistical error on those digits of ± 1).

Country	Range	N	GHSI score	GHSI rank	d_N^*	p
Africa						
Algeria	[5, 10524]	96	23.6	173	1.4079	0.02(9)
Botswana	[1, 15884]	87	—	—	1.0374	0.24(3)
Ethiopia	[3, 19940]	94	40.6	84	0.8937	0.43(8)
Kenya	[3, 19023]	94	47.1	55	1.2771	0.06(7)
Mauritius	[1, 10258]	80	—	—	1.4535	0.02(1)
Mozambique	[2, 13268]	92	28.1	153	1.1051	0.17(5)
Namibia	[1, 12944]	89	35.6	104	1.1731	0.12(2)
Nigeria	[1, 12531]	95	37.8	96	0.6236	0.84(9)
South Africa	[7, 162987]	95	54.8	34	1.5317	0.01(2)
Tanzania	[4, 24307]	23	—	—	1.2457	0.08(0)
Uganda	[1, 22511]	90	—	—	0.7271	0.70(8)
Zambia	[1, 19058]	93	28.7	152	1.3057	0.05(7)
Zimbabwe	[1, 26671]	93	38.2	92	0.9221	0.39(5)
Americas						
Argentina	[16, 219910]	95	58.6	25	1.5532	0.01(1)
Brazil	[6, 533024]	96	59.7	22	1.2301	0.08(9)
Bolivia	[7, 19834]	94	35.8	102	1.0026	0.28(4)
Canada	[2, 60784]	100	75.3	5	1.8364	0.00(1)
Colombia	[5, 204556]	95	44.2	65	1.3949	0.03(2)
Costa Rica	[9, 17469]	95	45.1	62	1.3167	0.05(3)
Cuba	[8, 64196]	94	35.2	110	2.0674	0.000(1)
Dominican Republic	[4, 11168]	95	38.3	91	1.3509	0.04(3)
Ecuador	[5, 14597]	95	50.1	45	1.3332	0.04(8)
Guatemala	[5, 26678]	94	37.2	125	1.6470	0.00(5)
Honduras	[6, 10595]	94	27.6	156	2.6172	0.000(0)
Mexico	[5, 128779]	96	57.6	28	1.1353	0.15(0)
Paraguay	[5, 20955]	95	37.5	103	1.6844	0.00(4)
Peru	[9, 60739]	95	49.2	49	1.0690	0.20(9)
Puerto Rico	[7, 32162]	93	—	—	1.7721	0.00(2)
Uruguay	[6, 26378]	94	41.3	81	0.8801	0.46(0)
U.S.A.	[12, 1745361]	101	83.5	1	0.7242	0.71(2)

Table 1: continued

Country	Range	<i>N</i>	GHSI score	GHSI rank	d_N^*	<i>p</i>
Eastern Mediterranean						
Afghanistan	[3, 12314]	96	32.3	120	1.2214	0.09(3)
Egypt	[5, 10778]	96	39.9	87	1.9710	0.000(4)
Iran	[47, 269975]	97	37.7	97	1.3850	0.03(4)
Iraq	[2, 83098]	96	25.8	167	1.2867	0.06(4)
Jordan	[5, 57666]	95	42.1	80	1.7989	0.00(1)
Lebanon	[5, 33605]	97	43.1	73	1.1526	0.13(7)
Libya	[1, 19510]	92	25.7	168	1.4154	0.02(8)
Morocco	[6, 64784]	95	43.7	68	1.1256	0.15(8)
Oman	[6, 17783]	96	43.1	73	1.0093	0.27(6)
Pakistan	[2, 40287]	95	35.5	105	1.6157	0.00(6)
Palestinian Territories	[8, 17509]	96	–	–	1.0512	0.22(8)
Qatar	[7, 13049]	96	41.2	82	2.4137	0.000(0)
Saudi Arabia	[5, 30925]	95	49.3	47	1.2266	0.09(1)
Tunisia	[5, 52076]	95	33.7	122	1.7322	0.00(2)
U.A.E	[2, 26285]	100	46.7	56	0.9135	0.40(8)
Europe						
Armenia	[1, 14417]	95	50.2	44	0.7368	0.69(3)
Austria	[8, 96094]	96	58.5	26	0.8381	0.52(8)
Azerbaijan	[2, 29155]	96	34.2	117	0.6744	0.78(5)
Belarus	[1, 14213]	96	35.3	108	2.2927	0.000(0)
Belgium	[1, 125246]	96	61.0	19	1.7387	0.00(2)
Bosnia and Herzegovina	[2, 11122]	95	42.8	79	0.6642	0.79(9)
Bulgaria	[2, 32962]	95	45.6	61	1.4023	0.03(0)
Croatia	[1, 37433]	96	53.3	38	0.6771	0.78(1)
Czechia	[27, 127489]	95	52.0	42	0.9291	0.38(5)
Denmark	[3, 78981]	96	70.4	8	1.4868	0.01(7)
Estonia	[1, 11930]	96	57.0	29	1.4258	0.02(6)
Finland	[1, 16510]	98	68.7	10	0.7175	0.72(3)
France	[1, 504469]	100	68.2	11	0.8111	0.57(2)
Georgia	[3, 33665]	96	52.0	42	0.9460	0.36(0)
Germany	[2, 406754]	99	66.0	14	0.8982	0.43(1)
Greece	[7, 47411]	96	53.8	37	1.7715	0.00(2)
Hungary	[7, 70400]	95	54.0	35	1.1028	0.17(7)
Ireland	[1, 53846]	96	59.0	23	1.0170	0.26(7)
Israel	[1, 65917]	97	47.3	54	0.7419	0.68(5)
Italy	[3, 257579]	98	56.2	31	1.3064	0.05(6)
Kazakhstan	[6, 56120]	94	40.7	83	1.5923	0.00(8)

Table 1: continued

Country	Range	N	GHSI score	GHSI rank	d_N^*	p
Latvia	[3, 16957]	95	62.9	17	1.6877	0.00(4)
Lithuania	[1, 20730]	96	55.0	33	0.7973	0.59(5)
Moldova	[1, 11680]	95	42.9	78	1.4101	0.02(9)
Netherlands	[2, 156007]	96	75.6	3	1.0607	0.21(8)
Norway	[1, 33281]	97	64.6	16	1.0084	0.27(7)
Poland	[6, 192441]	95	55.4	32	0.4811	0.96(3)
Portugal	[2, 86549]	95	60.3	20	1.4460	0.02(3)
Romania	[3, 104668]	96	45.8	60	1.1152	0.16(6)
Russia	[5, 281305]	95	44.3	63	1.5750	0.00(9)
Serbia	[1, 49995]	95	—	—	0.9512	0.35(3)
Slovakia	[1, 61514]	95	47.9	52	1.2145	0.09(7)
Slovenia	[2, 22657]	95	67.2	12	1.0019	0.28(5)
Spain	[1, 245818]	99	65.9	15	0.6382	0.83(2)
Sweden	[1, 46511]	97	72.1	7	1.6545	0.00(5)
Turkey	[6, 414312]	94	52.4	40	1.4798	0.01(8)
U.K.	[1, 683874]	100	77.9	2	1.0711	0.20(7)
Ukraine	[1, 153131]	95	38.0	94	1.4855	0.01(7)
South-East Asia						
Bangladesh	[7, 99693]	94	35.0	113	1.3715	0.03(7)
India	[1, 2738957]	97	46.5	57	1.2935	0.06(1)
Indonesia	[10, 350273]	95	56.6	30	1.2031	0.10(4)
Maldives	[1, 11401]	94	33.8	121	1.5900	0.00(8)
Myanmar	[4, 40004]	92	—	—	0.8451	0.51(6)
Nepal	[4, 61814]	91	35.1	111	0.9980	0.29(0)
Sri Lanka	[5, 41519]	95	33.9	120	1.2816	0.06(6)
Thailand	[1, 150652]	102	73.2	6	0.4247	0.98(2)
Western Pacific						
Australia	[3, 45560]	100	75.5	4	1.5845	0.00(8)
China	[1, 31333]	104	48.2	51	0.6523	0.81(4)
Japan	[1, 156931]	101	59.8	21	1.1777	0.11(9)
Malaysia	[3, 150933]	100	62.2	18	0.8115	0.57(2)
Mongolia	[1, 36698]	91	—	—	1.0876	0.19(1)
Philippines	[1, 144991]	97	47.6	53	1.7711	0.00(2)
Singapore	[4, 25950]	101	58.7	24	0.8253	0.54(9)
South Korea	[3, 47825]	101	70.2	9	1.2551	0.07(7)
Vietnam	[1, 125955]	97	49.1	50	0.4202	0.98(4)

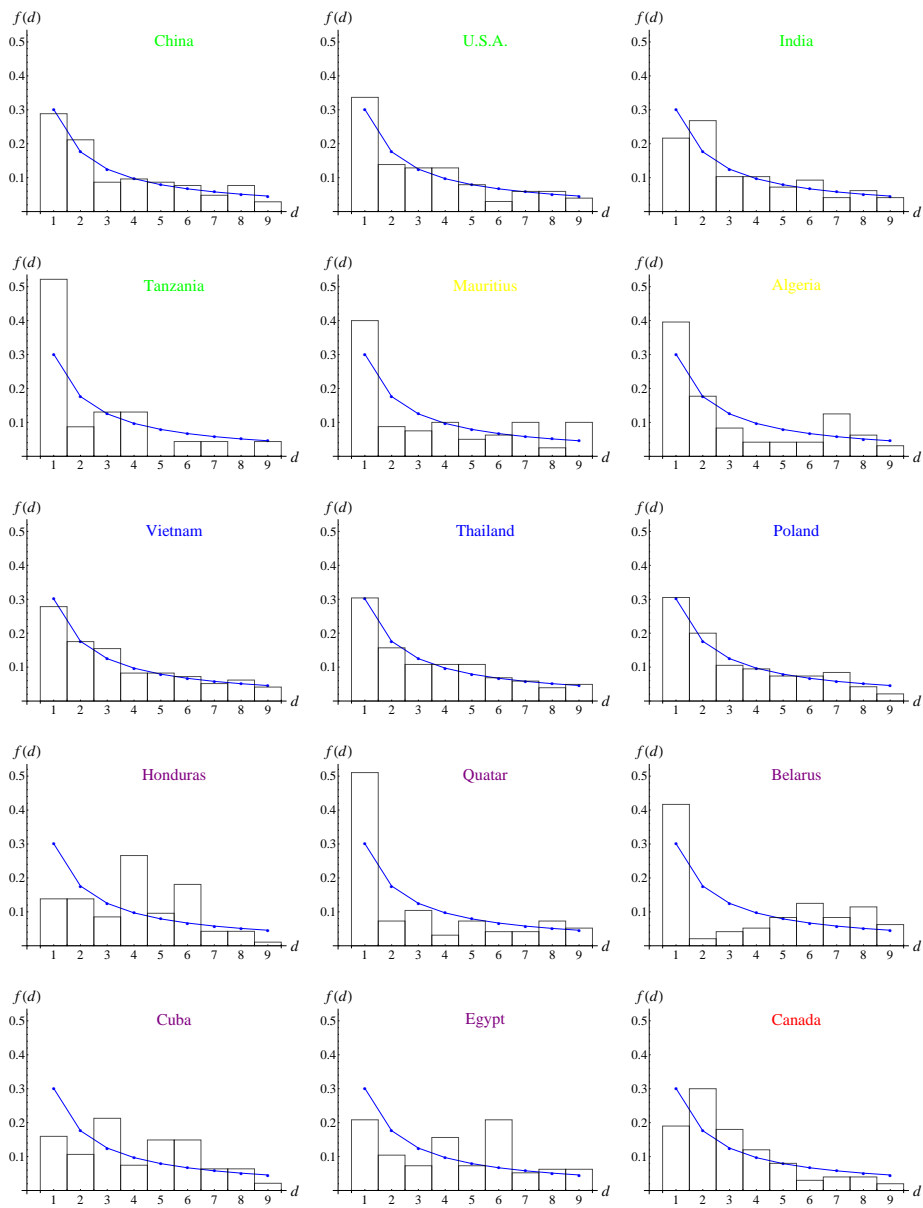


Figure 2: Observed first-digit frequencies of the Covid-19 weekly case counts for 15 selected countries: China (with the largest sample size N), USA (with the largest total number of cases), India (with the largest range of weekly case counts), Tanzania (with the smallest sample size N), Mauritius (with the smallest total number of cases), Algeria (with the smallest range of weekly case counts), Vietnam, Thailand, and Poland (the outliers in the first box plot of Fig. 5 with the world largest p values), Honduras, Qatar, Belarus, Cuba, and Egypt (with the smallest p values), and Canada (with the smallest p value in the interval $0.001 \leq p < 0.01$). The (blue) continuous lines represent Benford's law.

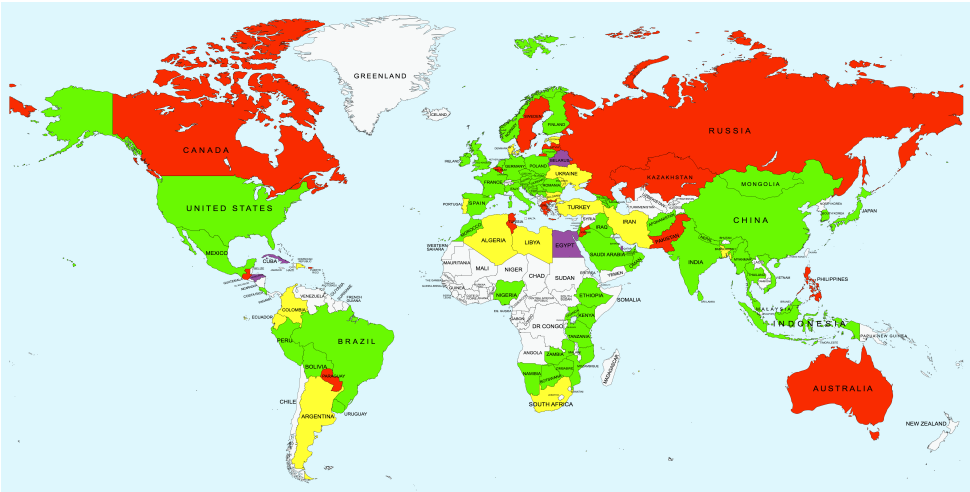


Figure 3: Spatial distribution of the p values of the Euclidean distance statistic for the first-digit distribution of Covid-19 weekly case counts by country. Ranges of p values are as follows: $p \geq 0.05$ (green), $0.01 \leq p < 0.05$ (yellow), $0.001 \leq p < 0.01$ (red), and $p < 0.001$ (purple). Light grey regions correspond to countries where Covid-19 weekly case counts have ranges below 4 orders of magnitude and then are excluded by our statistical analysis.

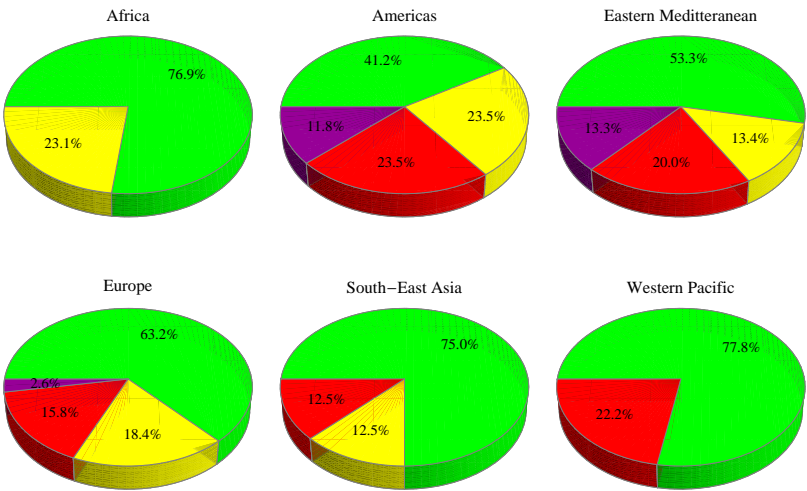


Figure 4: Percentages of countries in the six regions of the world (as defined by the World Health Organization) in a given range of p values of the Euclidean distance statistic for the first-digit distribution of Covid-19 weekly case counts by country. Ranges of p values in each pie chart are as follows: from top and clockwise, $p \geq 0.05$ (green), $0.01 \leq p < 0.05$ (yellow), $0.001 \leq p < 0.01$ (red), and $p < 0.001$ (purple).

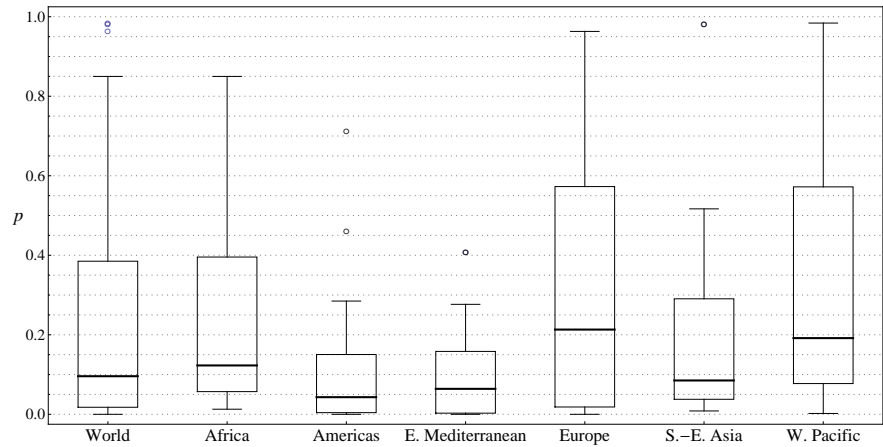


Figure 5: Box-and-whisker plots for the p values of the Euclidean distance statistic for the first-digit distribution of Covid-19 weekly case counts of all countries (the world) and countries in the six regions of the world, as defined by the World Health Organization.

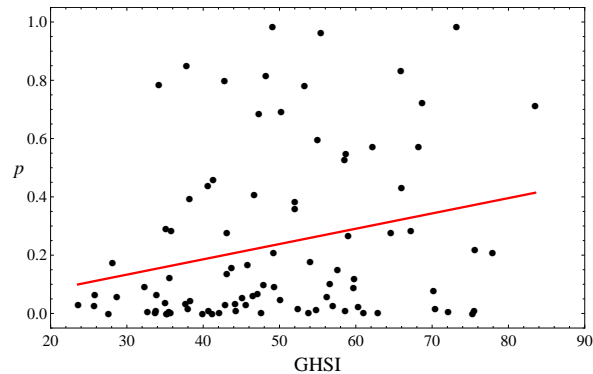


Figure 6: The p value of the Euclidean distance statistic for the first-digit distribution of Covid-19 weekly case counts by country as a function of the Global Health Security Index (GHSI) score. The (red) line is the regression line.

About the Authors

Adil Echchelh My career as a teacher-researcher was initiated at the University Louis Pasteur in Strasbourg (1992), the University of Limoges and finally the Ibn Tofail University, where I participated in the work of the establishment as a member elected to the University Council through the Pedagogical Commission, the Research Commission, and the management council as a member of the said commissions. In the case of international influence, I participated as a member in the work of the International Association of University Pedagogy, of the French Mechanical Society, of the French Society of Process Engineering, expert member of the CNRST project. Today, I focus on priority research areas such as water, the environment, health, energy, transport, road safety, artificial intelligence. On the educational level, I am responsible for three fields of study.

Bhushan Shashi is a Cadre Professor of Statistics at University of Lucknow, India. His research interests are applied statistics, sampling techniques, computational statistics and econometrics, among other things. Professor Bhushan has published more than 100 research papers in international/national journals. He has also published two books/monographs. Professor Bhushan is an active member of many scientific professional bodies.

Chebir Adil is an IRCA ISO 9001 Auditor. During his 14 years of experience in the field of quality management, compliance, and internal audit, he was able to conduct doctoral research on the issues of integration of ISO 9001, PCI-DSS and PCI-SLC standards within companies specializing in secure electronic transactions. His experiences are the subject of several articles published in international journals.

Chwila Adam is a PhD student at the Department of Statistics, Econometrics and Mathematics at the University of Economics in Katowice. He is employed as a senior market risk management specialist in the ING BSK. His research interests focus on small area estimation, machine learning regression models, and application of AI in banking domain.

Dwivedi Sada Nand is a former Senior Professor of Biostatistics, and Adjunct Faculty, Clinical Epidemiology Unit, All India Institute of Medical Sciences, New Delhi, India. Presently, he is a Senior Professor of Biostatistics at International Centre for Health Research, RD Gardi Medical College, Ujjain, Madhya Pradesh, India. His major specialisation is in randomized controlled clinical trials, systematic review and meta-analysis, epidemiological/ statistical model building, research methods including

program evaluation methods and sample surveys. As chief/co-guide, he has successfully guided 13 PhD students in biostatistics and more than 200 medical students regarding PhD/MD/DM/MS/MCh students. He has more than 350 published articles to his credit. He has been associated with large scale community based clinical studies providing translational/transformational results. He is founder Vice-President of the Society for Evidence Based Health Care, India; and President, Indian Society for Medical Statistics (2023–2024).

Eideh Abdulhakeem is an Associate Professor of Statistics at Al-Quds University, Palestine and an international expert and consultant in survey sampling. His research interests are small area estimation under informative sampling, statistical inferences from cross-sectional and longitudinal surveys, diagnostic and treatment of nonignorable nonresponse in surveys. Dr. Eideh has published more than 20 research papers in international journals. Besides, Dr. Eideh have obtained different memberships of statistical associations, including the ISI, IASS, and IAOS.

El Moury Ibtissam is pursuing a PHD degree in the field of Industrial Engineering. Researcher Associated at the Laboratory Electronic Systems, Information Processing, Mechanics and Energetics Laboratory, Faculty of Science, Ibn Tofail University. Her extensive enterprise-level experience brings a wealth of expertise in process optimization, project and quality management. Her dissertation examines the value an ISO 9001 certified Quality Management System adds to the overall performance of service-based companies. Her main areas of interest include quality management, logistics and econometric modelling.

Hadini Mohammed has a PhD in Industrial Engineering. Researcher Associated at the Laboratory Mechanics, Productivity and Industrial Engineering is an Industrial Engineering, ENSEM School of Engineering, Hassan II University, Casablanca (Morocco). He is a holder of an Engineering diploma in Computing & Automatics from ENSIAME, University of Valenciennes, France, and of a master's degree in management sciences from Business Administration Institute, France. In addition to that, he possesses certifications in the following areas: Risk management (ARM55, ARM54, ARM56), Six Sigma Black Belt, IRCA – Third-Party Audit. Furthermore, Mr. Hadini is holder of a post-graduate Expert Coach diploma and of a post-graduate Professional Coach diploma. With more than 20 years of experience in various fields, he is an experienced manager/VP with strong international credentials and experience in management of large and small organizations.

Keser İstem Köymen is an Associate Professor at the Department of Econometrics, Faculty of Economics and Administrative Sciences, Dokuz Eylul University. Her research interests are multivariate statistical analysis, functional data analysis, survey sampling, statistical inference, and data analysis in particular. She has many research papers on these topics.

Kocakoç Ipek Deveci is a full-time Professor at Dokuz Eylül University. She is an Industrial Engineer who completed her PhD in Dokuz Eylül University Department of Econometrics. She is also one of the founding faculty members of the Data Management and Analysis master's program. She works with master's and doctoral students on data science and artificial intelligence and provides consultancy in various companies. Her main topics of study are data visualization and analysis of health data, and machine learning.

Kumar Anoop is currently working as an Assistant Professor in the Department of Statistics, Amity University, Lucknow, India. He completed PhD in Applied Statistics from the Department of Mathematics and Statistics, Dr. Shakuntala Misra National Rehabilitation University, Lucknow, India. He also qualified UGC NET twice in population studies. His research area is sampling theory, missing data, measurement errors. Currently he is an academic editor in two journals: Plos One and Computational and Mathematical methods in Medicine.

Mohamed Ben Ali is a Researcher Professor at the Higher School of Technology in Casablanca. He received his PhD in Industrial Engineering. He is a researcher associated at the Laboratory Mechanics, Productivity and Industrial Engineering of High School of Technology, University Hassan II of Casablanca. He is a Professor of Bachelor of Quality Management and Productivity at the Faculty of Science and Techniques of Tangier. His research interests are primarily in the fields of food hygiene, modelling, entrepreneurship, productivity, quality, safety and environmental management and logistics.

Nath Dilip Chandra is currently serving as a Professor Emeritus in Royal Global University. His current research interests are in Biostatistics/Medical Statistics/Demography and Actuarial Statistics. He has guided successfully 30 PhD scholars. He has collaborated fruitfully with more than 60 national and international scholars and contributed more than two hundred research papers in reputed national/international journals.

Tomczyk Emilia is an Associate Professor at the Warsaw School of Economics in the Institute of Econometrics, head of the Department of Applied Econometrics and the chair of the Quantitative Methods in Economics and Information Science Board. She specializes in applied econometrics, theory and testing of rationality, quantification of qualitative data, analysis of survey data, applications of entropy measures in economics, and evolutionary game theory. She is a Fulbright scholar (at the University of Delaware, Newark, DE) and member of the editorial board of two major Polish economic journals. For 25 years has been teaching econometrics, mathematical programming, and game theory.

Ul Hassan Mahmood is a senior lecturer of statistics with a diverse range of research interests. His areas of expertise include applied statistics, item response theory, optimal experimental design, and analysis of extreme events, distribution theory, and data analysis. Hassan has published more than 25 research papers in both national and international journals.

Verma Vivek is an Assistant Professor at the Department of Statistical, Assam University, Silchar, India. Worked and is associated with many reputed organizations like Indian Statistical Institute, Kolkata, Rajendra Institute of Medical Sciences, Ranchi and All India Institute of Medical Sciences, New Delhi. His area of interest includes Bayesian analysis, epidemiology, demography, clinical trials and statistical mechanics. He has 35 research papers to his credit in various international journals.

Zielińska-Kolasińska Zofia's research interests include interval estimation of the odds ratio, Bayesian methods in actuarial science, forecasting the unemployment and employment rates.

Zieliński Wojciech is a Full Professor in the Department of Econometrics and Statistics of the University of Life Sciences in Warsaw. His main research area is classical inference and its applications, especially problems of interval estimation and testing statistical hypotheses as well as robust inference. He has published more than 80 articles and textbooks.