Volume 22, Number 1, March 2021 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

EDITORIAL OFFICE **ISSN 1234-7655**

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

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

CONTENTS

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1, pp. III–VII

From the Editor

A set of twelve papers that make up this issue is conventionally divided into three sections: research articles, other articles, and research communicates. They cover a wide spectrum of problems while representing also a rich set of approaches.

The first section, *research articles*, starts with **Andrzej Młodak's** article ***An application of a complex measure to model–based imputation in business statistics.*** Imputation is typically used in order to fill the gaps created by missing data and to reduce bias in aggregated estimates as their consequence. This paper presents research on the efficiency of model–based imputation in business statistics, where the explanatory variable is a complex measure constructed by taxonomic methods. The proposed approach involves selecting explanatory variables that fit best in terms of variation and correlation from a set of possible explanatory variables for imputed information, and then replacing them with a single complex measure (meta–feature) exploiting their whole informational potential. The paper also presents five types of similar techniques: ratio imputation, regression imputation, regression imputation with iteration, predictive mean matching and the propensity score method. The results show that models with a strong dependence on functional form assumptions can be improved by using a complex measure to summarize the predictor variables rather than the variables themselves (raw or normalized).

In the paper ***Modelling and forecasting monthly Brent crude oil prices: a long memory and volatility approach*** by **Remal Shaher Al-Gounmeein** and **Mohd Tahir Ismail** the Standard Generalised Autoregressive Conditionally Heteroskedastic (sGARCH) model and the Functional Generalised Autoregressive Conditionally Heteroskedastic (fGARCH) model were applied to study the volatility of the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model. The other goal of this paper is to examine long memory and volatilities simultaneously by using the ARFIMA-sGARCH hybrid model and comparing it against the ARFIMA-fGARCH hybrid model. Consequently, the hybrid models were configured with the monthly Brent crude oil price series for the period from January 1979 to July 2019. These datasets were considered as the global economy is currently facing significant challenges resulting from noticeable volatilities, especially in terms of the Brent crude prices, due to the outbreak of COVID-19. As a result, the following conclusions were reached: the ARFIMA-sGARCH(1,1) model and the ARFIMA-fGARCH(1,1) model under normal

distribution proved to be the best models, demonstrating the smallest values for these criteria. These models can be used to predict oil prices more accurately than others.

Camilla Ferretti's article ***Measurement of enterprise mobility among size classes, taking into account business demography*** addresses conceptual issues of enterprise mobility in terms of the capability to create or liquidate jobs. Some mobility measures offered in the literature are modified so that they also take into account newborn and exiting firms. The proposed index has all the relevant basic properties which make it a rigorous descriptive statistics. The mobility of Italian capital-owned enterprises in the years 2010–2017 is analysed in the case study. It may be considered an initial step in future research regarding its different applications (e.g. labour market flows or movements among income classes), also considering more complex theoretical backgrounds.

OlaOluwa S. Yaya, Oluwaseun A. Otekunrin, Ahamuefula E. Ogbonna present the article ***Life expectancy in West African countries: evidence of convergence and catching up with the north*** focused on the possibility of the convergence and catching up of life expectancy values observed in West African countries with those observed in North African countries. Following the theory of time series convergence, documented (e.g. Bernard and Durlauf, 1996; Greasley and Oxley, 1997) more robust unit root tests, based on the Fourier nonlinearity and instantaneous breaks (proposed in Furuoka, 2017), are used in investigating the convergence of each pair of a West African country and its North African counterpart. As no unit root in the differences of the pairs implies convergence, the results obtained by means of a new statistical approach quite outperform those produced by classical unit root tests. The results provide general evidence of the convergence of life expectancy values recorded in West Africa and North Africa.

Sunil Kumar's and **Apurba Vishal Dabgotra's** paper ***A latent class analysis on the usage of mobile phones among management studies*** starts with observation that in the past few years, wireless devices, including pocket PCs, pagers, mobile phones, etc., have gained popularity among a variety of users across the world and has increased significantly in India. Cell phones are now the most popular form of electronic communication and they have turned from a technological tool to a social tool, and as such they deserve to be thoroughly examined. Authors explore the attitude of young adults towards cell phones and identify the hidden classes of respondents according to the patterns of mobile phone use using the Latent Class Analysis (LCA) as a tool to detect any peculiarities, including those gender-based. They propose a method of selecting the most useful variables for an LCA-based detection of group structures from within the examined data applying a greedy search algorithm, where during each phase the models are compared through an approximation to their Bayes factor. The findings

demonstrate that young people display various feelings and attitudes toward cell phone usage.

In the next paper, ***Analysis of Polish mutual funds performance: a Markovian approach,*** **Dariusz Filip** and **Tomasz Rogala** discuss the fundamental issue of how mutual funds provide benefits for their clients. The performance of Polish mutual funds has been evaluated in terms of their efficiency, including their potential inertia over time. The phenomenon of economies of scale resulting from assets inflow to the fund by means of the Markovian framework has been examined. The results are consistent with the efficient market hypothesis. When assessing the market-adjusted returns, underperformance was noticed in both small and large funds. The smart money effect, recognised in the literature, is not confirmed here. However, there are some noticeable investor reactions, such as the phenomenon of chasing performance.

Shakeel A. Mir and **Immad A. Shah** in the paper ***The construction and analysis of repeated measurement designs (RDM) using trial and error method*** apply repeated measurement designs which prove broadly applicable in almost all branches of bio-sciences, including agriculture, animal husbandry, botany, zoology. Unbiased estimators for elementary contrasts among direct and residual effects are obtainable in this class of designs, which is considered their important property. Authors attempt to provide a new method of overcoming a drawback in the construction method developed by Afsarinejad (1983), where one or more treatments may occur more than once in certain sequences causing the constructed designs to no longer remain uniform in the examined periods. Nine designs were constructed and presented jointly with their corresponding mathematical analyses.

Marta Małecka's article ***Testing for a serial correlation in VaR failures through the exponential autoregressive conditional duration model*** addresses the issue of evaluation of risk models which currently remains based on the VaR measure while referring to the Basel regulations and exploring statistical properties of the exponential autoregressive conditional duration (EACD) VaR test. It was shown that the tested parameter lies at the boundary of the parameter space, which can profoundly affect the accuracy of this test. To compensate for this deficiency, a selection of chi-square distributions is applied. As a result, an improvement in the computational efficiency of the procedure is observed, as the Monte Carlo simulations used to implement the EACD VaR test in earlier studies are replaced. As s results, the EACD approach to testing VaR has shown the potential to enhance statistical inference in most problematic cases – for small samples and for those close to the null.

The *other articles* section contains two papers based on presentations given at the *Multivariate Statistical Analysis 2019* Conference in Łódź.

It starts with **Tomasz Żądło's** article ***On generalization of Quatember's bootstrap*** considering the problem of the estimation of the design-variance and the design-MSE of different estimators and predictors. Bootstrap algorithms applicable to complex sampling designs are used. A generalisation of the bootstrap procedure studied by Quatember (2014) is proposed. In most of the cases considered in simulation study it leads to more accurate estimates (or to very similar ones in remaining cases) of the design-MSE and the design-variance compared with the original algorithm and its other counterparts.

In the next paper, ***Bankruptcy prediction of small- and medium-sized enterprises in Poland based on the LDA and SVM methods,*** **Aneta Ptak-Chmielewska** analyses the impact of the last financial crisis on the small- and medium-sized enterprises (SMEs) sector across countries, affecting them on different levels and to a different extent. The economic situation in Poland during and after the financial crisis was quite stable compared to other EU member states. Since the Altman Z-Score model was devised, numerous studies on bankruptcy prediction have been written. Most of them involve the application of traditional methods, including linear discriminant analysis (LDA), logistic regression and probit analysis. However, most recent studies in the area of bankruptcy prediction focus on more advanced methods, such as case-based reasoning, genetic algorithms and neural networks. In this paper, the effectiveness of LDA and SVM predictions were compared. The hypothesis assuming that multidimensional discrimination was more effective was verified on the basis of the obtained results.

The last section, *research communicates,* contains three papers. **Danute Krapavickaitė's** article ***Estimation of the number of residents included in a population frame*** discusses statistical consequences of a high migration rate which causes the number of a country's residents to become extremely volatile and negatively affects the quality of population frames. A method for estimating the number of residents included in a study population frame is proposed, which involves the cross-classification of a population register with other databases, which contain information relating to the activities of the population elements in a given country. The estimates from an ongoing sample survey are applied to some of the cells.

In the next paper, ***A new family of robust regression estimators utilizing robust regression tools and supplementary attributes,*** by **Irsa Sajjad, Muhammad Hanif, Nursel Koyuncu, Usman Shahzad, Nadia H. Al-Noor,** a family of estimators is proposed, based on the adaptation of the estimators presented by Zaman (2019), followed by the introduction of a new family of regression-type estimators utilising robust regression tools (LAD, H-M, LMS, H-MM, Hampel-M, Tukey-M, LTS) and supplementary attributes. The mean square error expressions of the adapted

and proposed families are determined through a general formula. The study demonstrates that the adapted class of the Zaman (2019) estimators is in every case more proficient than that of Zaman and Bulut (2018a). The theoretical findings are supported by real-life examples.

Hadi Safari-Katesari and **Samira Zaroudi** in the paper entitled ***Analysing the impact of dependency on conditional survival functions using copulas*** discuss the case of insurance contract reserves for coupled lives are considered jointly, which has a significant influence on the process of determining actuarial reserves. Specifically, conditional survival distributions of life insurance reserves are computed using copulas. Subsequently, the results are compared with an independence case. These calculations are based on selected Archimedean copulas and apply when the 'death of one individual' condition exists. The estimation outcome indicates that the insurer reserves calculated by means of Archimedean copulas are far more effective than those resulting from an independence assumption. The study demonstrates that copula-based dependency modelling improves the calculations of reserves made for actuarial purposes.

Włodzimierz Okrasa

Editor

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1, pp. IX

Submission information for Authors

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

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

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

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

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

Manuscript should be submitted electronically to the Editor:
sit@stat.gov.pl,
GUS/Statistics Poland,
Al. Niepodległości 208, R. 296, 00-925 Warsaw, Poland

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

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://stat.gov.pl/en/sit-en/guidelines-for-authors/

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1, pp. XI–XII

Editorial Policy

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

***

Abstracting and Indexing Databases

Statistics in Transition new series is currently covered in:

Databases indexing the journal:

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

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 1–28, DOI 10.21307/stattrans-2021-001
Received – 02.03.2020; accepted – 17.07.2020

An application of a complex measure to model–based imputation in business statistics[1]

Andrzej Młodak[2]

ABSTRACT

When faced with missing data in a statistical survey or administrative sources, imputation is frequently used in order to fill the gaps and reduce the major part of bias that can affect aggregated estimates as a consequence of these gaps. This paper presents research on the efficiency of model–based imputation in business statistics, where the explanatory variable is a complex measure constructed by taxonomic methods. The proposed approach involves selecting explanatory variables that fit best in terms of variation and correlation from a set of possible explanatory variables for imputed information, and then replacing them with a single complex measure (meta–feature) exploiting their whole informational potential. This meta–feature is constructed as a function of a median distance of given objects from the benchmark of development. A simulation study and empirical study were used to verify the efficiency of the proposed approach. The paper also presents five types of similar techniques: ratio imputation, regression imputation, regression imputation with iteration, predictive mean matching and the propensity score method. The second study presented in the paper involved a simulation of missing data using IT business data from the California State University in Los Angeles, USA. The results show that models with a strong dependence on functional form assumptions can be improved by using a complex measure to summarize the predictor variables rather than the variables themselves (raw or normalized).

Key words: complex measure, ratio imputation, regression imputation, predictive mean matching, propensity score method.

1. Introduction

In this paper we aim to use imputation in order to obtain a data set that resembles the true data and that can be used for multiple purposes rather than a specific purpose.

[1] This paper is an extended and substantially modified version of results presented during the fourth European Establishment Statistics Workshop under the auspices of the European Network for Better Establishment Statistics (ENBES), held on 7th – 9th September 2015 in Poznań, Poland.

[2] Statistical Office in Poznań, Centre for Small Area Estimation, address: Statistical Office in Poznań, Branch in Kalisz, ul. Piwonicka 7–9, 62–800 Kalisz, Poland. E–mail: a.mlodak@stat.gov.pl and Calisia University – Kalisz, Poland. ORCID: https://orcid.org/0000-0002-6853-9163.

This is a common aim at national statistical institutes and other institutes that collect and disseminate statistical data (cf. e.g. De Waal *et al.* (2011)). For example, in order to protect the privacy of individual respondents and avoid disclosure of sensitive information, the actually collected data may not be released. Instead a national statistical institute may then opt to release imputed data that resemble the actually collected data as much as possible. Therefore, some tools of the statistical disclosure control (e.g. perturbative methods or construction of synthetic data) are, in fact, based on imputation models (cf. Hundepool *et al.* (2012)). The imputation is very important especially in the dissemination of microdata. It is because in official statistics (and not only here) a growing demand on detailed microdata has been observed in the last decades. Therefore, the production of maximally informative and secure microdata becomes crucial.

The main focus of this paper is to study the efficiency of the use of a complex measure as an auxiliary variable in some methods of model–based imputation, especially for business statistics. The complex measure reflects the diversification of objects (e.g. economic entities) in terms of a complex social or economic phenomenon, described by many variables. A measure of this kind is constructed in such a way as to ensure that information contained in the variables and mutual relationships between them are maximally exploited, which traditional models of dependency – i.e. regression function – can overlook (cf. e.g. Młodak (2014) or Malina and Zeliaś (1998)).

Using the proposed complex measure instead of using several auxiliary variables leads to a loss of less information. From a purely theoretical point of view, using the proposed complex measure instead of using several auxiliary variables cannot lead to better estimates. If it does, it means that the full potential of the imputation methods using several auxiliary variables is not used. However, from a practical standpoint, there are some compelling reasons for using the proposed complex measure, such as (1) for some imputation methods using a single complex measure instead of several auxiliary variables may be easier to implement in practice (this certainly holds for imputation methods that were designed for a single auxiliary variable, such as ratio imputation) and (2) in a single complex measure it may be easier to take outliers into account than in several auxiliary variables for which one would have to use multivariate outlier techniques. The additional motivation of the use of such an approach is that users of statistical information are looking now mainly for the provision of complex characteristics of macro–domains, such as, e.g. the labour market, infrastructure, environment, etc.

The proposed approach involves selecting from a set of possible explanatory variables for imputed information the best ones in terms of variation and correlation (called *diagnostic features*) and then replacing them with a single complex measure (called also the *meta–feature*) exploiting their whole informational potential.

This complex measure is constructed as a function of median distance of objects described by normalized diagnostic features from the benchmark of development, i.e. artificial object 'ideal' from the analysed point of view. The normalisation is performed using the Weber median (called sometimes also L_1*–median* or *geometric median*) being a point of multidimensional space minimizing the sum of Euclidean distances from the points representing given objects. In this way the resulting imputation models are simpler: they are less computationally demanding and easier to interpret, while being sufficiently efficient. Here, the utility of this solution will be verified using the following methods of model–based imputation: ratio imputation, regression imputation, regression imputation with iterative extension, predictive mean matching and propensity score method. Empirical analysis is conducted here using both simulated and real data. The assumptions of simulation account for circumstances most often observed in business statistics. That is, the values of variables were drawn from the multivariate log–normal distribution with relevant parameters such that the auxiliary variables are mutually correlated as little as possible and maximally – with the target variable. The missing data were modelled using the missing at random (MAR) approach taking into account the impact of auxiliary variables into lack of data for the target one. The efficiency of imputation is assessed using the estimates of precision (MSE of target parameter estimation using imputed data) decomposed into three components, including the term connected with the "pure" imputation effect. The MSE in the presented imputation should be not greater than when all original data are used and minimal as a measure of the precision of estimation.

This paper is structured as follows. In Section 2 we present the assumption and methods of selection of diagnostic variables in the taxonomic model as well as the construction of the complex measure on their basis using the Weber median and other ordinal statistics. Next, Section 3 provides a short description of the analysed model–based imputation methods and Section 4 discusses tools of quality assessment used in our investigation (i.e. approximate estimation of imputation precision). Section 5 contains methodology of conducted simulation study and its results. An empirical study which involved a simulation of missing data using IT business data from the California State University in Los Angeles, USA is presented in Section 6. Finally (Section 7) some conclusions are collected.

These results expand on some issues not included in the final version of relevant sections of „*Handbook on Methodology for Modern Business Statistics*", edited by L. Willenborg, S. Scholtus and R. van de Laar (Collaboration in Research and Methodology for Official Statistics), created in 2014 within the ESSnet (European Statistical System network) initiative MeMoBuSt (Methodology for Modern Business Statistics), but which were investigated during that project.

2. Construction of the complex measure

The complex measure is aimed at efficient creation of a single variable containing the information potential of many collected variables describing a composite social or economic phenomenon in given objects (e.g. economic entities). The construction of the complex measure consists of the following steps, described also (although in terms of interval data) by Młodak (2014).

Step 1. Choice of variables and data collection: one should use information which properly describes the subject of research. The collected variables containing such information should be measurable, complete and comparable. To improve data comparability, they should have the form of indices (i.e. need to be calculated per capita, per 1 km^2, per 1000 inhabitants, per enterprise, etc.). Keeping values expressed in absolute numbers (e.g. number of economic entities, total revenues, etc.) can lead to some distortion of results – some (often not numerous) objects are (by their nature or specific circumstances) characterized by values much greater than others (e.g. large cities versus rural areas). The use of indices (relatively small, from elsewhere) substantially reduces the scale–dependency of the whole procedure. Of course, if the final complex measure is used to impute or estimate values of some other target variable, the variable used to construct the complex measure should be strictly connected with the target variable.

Step 2. Verification of variables: firstly, the elimination of variables that are not efficient in discrimination of objects, i.e. dropping variables for which the absolute value of the coefficient of variation (CV) is smaller than an arbitrarily established threshold (usually 0.1 – cf. Młodak (2014)) is conducted. Such variables are regarded as not showing the diversification of the analysed objects and hence they are dropped. This procedure is justified by the assumption that taxonomic methods are applied to phenomena where the clear diversification of the analysed objects is expected and then complex measures should reflect such diversification. Next, variables are verified in terms of correlation – we eliminate variables that are too correlated with others (and, hence, carry similar information). Here the inverse correlation matrix method was used. Its diagonal entries belong to $[1, \infty)$ (cf. e.g. Neter, Wasserman and Kutner (1985)). If some of them are too large (more often greater than 10) then relevant variables are regarded as 'bad'. They can be eliminated, but not necessarily all. That is, if there are more than one 'bad' variable then one should exactly analyse correlations between such variables and on the basis of such a comparison make elimination which is as sparing as possible (i.e. as few variables as possible should be dropped) and simultaneously guarantees a sufficiently weak correlation of the remaining variables. The correlation verification requires then some subjectivity in taking decisions about elimination. In the case when the final complex measure will be used as an auxiliary

variable in the imputation or estimation model, the correlation verification is conducted usually taking into account also the correlation of possible diagnostic variables with the target one. That is, when the analysis of the correlation matrix analysis does not allow to take unambiguous decision about elimination of some 'bad' variables, the ones whose correlation with the target one is smaller than others are removed. This approach, proposed by Malina and Zeliaś (1998), exploits mutual connections between features. It is very important because the economy is a 'system of connected vessels' and therefore the variables should be perceived not separately, but rather jointly – as a whole. The set of variables which remains after verification is called the *set of diagnostic features*. Thus, each object i is described by values of diagnostic features $X_1, X_2, \dots, X_m$ and is represented by the point $\boldsymbol{\gamma}_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$, where x_{ij} denotes the value of diagnostic feature X_j for i–th object, $i = 1,2, \dots, n$, $j = 1,2, \dots, m$.

Step 3. Identification of the character of diagnostic features (variables after verification): considering the impact of variables on the situation of an entity with respect to a phenomenon of interest, we can distinguish three types of variables:

- *stimulants* – the higher the value, the better the situation of an object in this context (e.g. average monthly revenue or Gross Domestic Product per capita),
- *destimulants* – higher values indicate a deterioration of the entity's situation,
- *nominants* – variables which behave like stimulants below a certain critical point and may switch to being destimulants after crossing it. That is, below this point this feature has the characteristic of being a stimulant and above it – a destimulant. Or on the contrary – greater values (and simultaneously smaller than the optimum) are 'worse' whereas smaller (but greater than the optimum) are regarded as being 'better'.

The critical point for nominant can be identified by own experience or by consultation with famous experts. An alternative – and more formal – approach in this respect could be based on the Cramér–von Mises or Anderson–Darling test – the critical point will refer then to the extremum of theoretical distribution best adjusted to the empirical data (if it is U–shaped, of course).

Destimulants and nominants are converted into stimulants by taking their values with opposite signs (in the case of nominants this is done only to the part with destimulative properties).

Step 4. Normalization of features, aimed at obtaining a comparable form of diagnostic variables. There are many forms of normalization (see e.g. Zeliaś (2002)). To exploit all connections between them it is good to use the Weber median, i.e. the vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_m) \in \mathbb{R}^m$ minimizing the sum of Euclidean distance from points

$\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_n$ reflecting given objects (cf. Młodak (2006)). The normalisation formula is then as follows:

$$z_{ij} = \frac{x_{ij} - \xi_j}{1.4826 \cdot \underset{l=1,2,\ldots,n}{\text{med}} |x_{lj} - \xi_j|}, \quad i = 1,2,\ldots,n,\ j = 1,2,\ldots,m. \tag{1}$$

Recall that (cf. Rousseeuw and Leroy (1987)) a probabilistic premise for the use of the constant 1.4826 (approximatively equal to $1/(\varphi^{-1}(3/4))$, where φ is a distribution function of the normal distribution with an expected value of zero and a standard deviation equal to 1), is the fact that if $Y_1, Y_2, \ldots, Y_k$ are independent and identically distributed random variables having the normal distribution with a mean μ and a variance σ^2 ($\sigma > 0$), then $E\big(1.4826 \cdot \text{mad}(Y_1, Y_2, \ldots, Y_k)\big) \approx \sigma$ for sufficiently large natural k (which gives approximative standardization), where $\text{mad}(Y_1, Y_2, \ldots, Y_k) = \underset{i=1,2,\ldots,k}{\text{med}} |Y_i - \text{med}(Y_1, Y_2, \ldots, Y_k)|$ is the median absolute deviation of variables $Y_1, Y_2, \ldots, Y_k$. As we can see, the expression $\underset{i=1,2,\ldots,n}{\text{med}} \big|x_{ij} - \xi_j\big|$ used in (1) is a special modification of the median absolute deviation of X_j, i.e. $\text{mad}\big(X_j\big)$, where the classical median was replaced with the respective coordinate of the Weber median. To approximate the Weber median, Vandev (2002) proposed the iterative algorithm based on the Newton–Raphson procedure.

The normalisation (1) leads to minimization of scale–dependency of the final results of the procedure: the properties of Weber and classical median as well as relative character of the input variables ensures that outlying is usually strongly reduced without loss of information contribution.

Step 5. Definition and determination of the taxonomic benchmark of development – an artificial, ideal object is defined, with which others are compared. As it was noted by Młodak (2014), this object is usually described by the most desirable values of particular diagnostic features (in the normalized version). Because all diagnostic features are stimulants, one can assume that the benchmark is defined as being represented by the vector $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_m)$, where

$$\psi_j = \max_{i=1,2,\ldots,n} z_{ij}, \qquad j = 1,2,\ldots,m.$$

Therefore, the object described by those values is regarded as being 'ideal'. This method can be perceived as being endogenic, because the benchmark is constructed on the basis of the internal properties of the analysed empirical model[3].

Step 6. Computation of distances of entities from the benchmark. A distance being a function of the absolute differences between the respective values for a given object

[3] Alternatively, the benchmark can be defined also in an exogenous manner, i.e. arbitrarily and independently from properties of the data and the model. Such approach can be justified by some commonly adopted standard occurring in some fields. For example, if we analyse some data concerning environmental protection, the values of the benchmark can be assumed as being represented by relevant thresholds of allowable pollution generated by factories established by the European Commission and valid in the European Union.

and for the benchmark is used here. Of course, it should be nonnegative, reflexive and symmetric to be well defined. There are many ways to define it. Now we use the median distance:

$$d_i = \underset{j=1,2,\dots,m}{\text{med}} \left| z_{ij} - \psi_j \right|, \quad i = 1,2,\dots,n.$$

Step 7. Determination of a synthetic measure. The synthetic measure $\boldsymbol{\eta} = (\eta_1, \eta_2, \dots, \eta_n)$ is constructed as a statistical function of distances of analysed objects from the benchmark. The central point of this construction is a postulate that this measure should be a continuous function of a position of the distance of a given object from the benchmark taking into account the general extreme values of such a distance in the model. In our studies it is based on median and median absolute deviation of distances. The synthetic measure enables then to better identify possible outliers, where the analysed situation is especially difficult. That is, we put

$$\eta_i = 1 - \frac{d_i}{\text{med}(\boldsymbol{d}) + 2.5 \cdot \text{mad}(\boldsymbol{d})}, \tag{2}$$

$i = 1,2,\dots,n$, $\boldsymbol{d} = (d_1, d_2, \dots, d_n)$, $\text{mad}(\boldsymbol{d}) = \underset{i=1,2,\dots,n}{\text{med}} |d_i - \text{med}(\boldsymbol{d})|$, where 2.5 is called the *robust threshold value*. It ensures that $[0, \text{med}(\boldsymbol{d}) + 2.5 \cdot \text{mad}(\boldsymbol{d})]$ represents approximatively the 90% confidence interval for $\boldsymbol{d}$. It allows for achieving sufficient robustness of $\boldsymbol{\eta}$ to outliers (cf. Rousseeuw and Leroy (1987)). The values belong usually to the interval $[0,1]$. Only in special extreme cases (i.e. when an object is a strong outlier) they can be negative. The highest the value of the index (2), the better the situation in the investigated context.

The aforementioned construction can be applied in many circumstances without strong general restrictions except for strict connection with the subject of interest (of which – for imputation and estimation) with the variable to be imputed/estimated and relevant quality of the input data. They should:

- be unambiguously and precisely defined,
- describe the analysed phenomenon as exhaustively as possible,
- maintain proportionality of representing partial phenomena,
- provide measurable, available and complete statistical information for all investigated objects.

Of course, the thresholds of proper variability and sufficiently small correlation should be carefully established. The use of statistics of observations (such as 'classical' median or the Weber median) allows for an increase in the resistance to extreme, but a few, outliers. Some difficulties with application of the Weber median may occur when the complex measure will be constructed for several consecutive periods in time. Irregular perturbations at only few point may result in a large change of the position of the Weber median (cf. Durocher and Kickpatrick (2009)). However, it concerns only

the situation on the plane ($\mathbb{R}^2$) when the change is neither along nor almost along the ray starting at previous Weber median and going through previous location on a given point (i.e. if change of $\boldsymbol{\gamma}_i$ into $\boldsymbol{\gamma}_i'$ is that $\boldsymbol{\gamma}_i'$ is located on – or very close to – the ray $\overrightarrow{\boldsymbol{\theta}\boldsymbol{\gamma}_i}$, where $\boldsymbol{\theta}$ is the Weber median of $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \ldots, \boldsymbol{\gamma}_n$). In practice, however, such a situation occurs very rarely: the taxonomy based on the Weber median is most efficient when it uses at least three diagnostic features (i.e. when $m \geq 3$); on the other hand, changes in the consecutive periods in time are usually relatively small. Moreover, the Weber median is unique for $n \geq 3$ and $m \geq 2$ (cf. e.g. Milasevic and Ducharme (1987)). Otherwise, it is assumed to be equal to the classical median.

The advantages of the construction are as follows:

- the complex measure provides clearly interpretable information about the whole multivariate phenomenon,
- owing to standardization and normalization it is independent of differences between diagnostic variables in the impact on the general situation of an object and in the scale of measurement,
- it exploits maximally possible connections between diagnostic features – even those which cannot be statistically quantified and – in the case of imputation or estimation – also with the target variable,
- it contains the maximum information potential of particular diagnostic variables,
- its distribution is – in some sense – a resultant of distributions of diagnostic variables (of which in terms of variation).

The last two properties can be arguments showing greater benefits coming from the use of the complex measure then, e.g. the Principal Component Analysis (PCA) – (cf. e.g. Jolliffe (2002)). The PCA generates usually several components each of them being often of the other quality expressed by the share in total common variance. In this case, the first of them, i.e. the one whose share is the greatest, is usually assumed as the complex measure. In practice, however, the loss of original variation borne in this way is often rather considerable. Moreover, in extremal situations the shares can be even equally distributed among components obtained by PCA. In contrary to PCA, the presented complex measure reflects a variation and shape of distribution of diagnostic features as maximally as possible. Similarly, the lasso regression (cf. e.g. Tibshirani (1996)), which – by the constraint for coefficients of regression – tends to marginalize some possible auxiliary variables, seems to be in the investigated context slightly doubtful: it poses a risk of omitting some important (although maybe statistically less significant) information. In examples given by Tibshirani (1996) sometimes even over a half of primarily considered auxiliary variables were omitted as their coefficients were zeros. The complex measure enables to avoid – to a large extent – such unnecessary loss of useful information.

3. Investigated methods of model–based imputation

Now, we describe briefly the methods of model–based imputation analysed in the paper and possibilities of implementation of the complex measure to them.

3.1. Ratio imputation

Ratio imputation consists in replacing missing values with the value of a known auxiliary variable multiplied by the ratio of some descriptive summary statistics of the variable with the missing value (e.g. mean, median or sum) and the relevant statistics for the auxiliary variable. It is tacitly assumed here that the ratio of the values of these variables for a given unit is the same as the ratio of some 'total' values of these two variables. For example, if data about the value of sales for an enterprise are missing, but its total expenditure amounts to €20,000, mean sales for the whole analysed group of enterprises which the given one belongs to is €30,000 and the mean expenditure is €21,000, then the predicted value of sales is computed as 20,000×(30,000/21,000) = 20,000×(10/7) = €28,571.43. Of course, depending on current circumstances, instead of summary statistics we can also use in this context relevant values for a higher level of data aggregation (i.e. the total value for a given NUTS unit). There are also some special cases of ratio imputation such as, e.g. its weighted option, suggested by Arcaro and Yung (2001).

If there are several variables which are strictly connected with the imputed one, we can optimize the choice of the variable to be used for the imputation, e.g. by analysing the distribution of the known values of the imputed variable and appropriate values of the possible auxiliary variable (e.g. using the Wilcoxon signed rank or Cramér – von Mises – in the version for two samples – test). The auxiliary variable, for which such a 'trimmed' distribution is closest to the distribution of the known value of the variable to be imputed, will serve as the basis for the ratio imputation. Other – and faster – possibility is to compute the Pearson's correlation coefficient and chose such a variable which is most correlated with the target one. These methods do not, however, guarantee an exploitation of the whole information potential concerning the connections between original variables. Therefore, the use of the complex measure $\boldsymbol{\eta}$ seems to be desirable in this situation.

The quality of this type of imputation depends, first of all, on the degree of association between imputed and auxiliary variables. The stronger it is, the better the adjustment of imputed values is. The usefulness of this attempt depends on the availability of an appropriate auxiliary variable. The variance is much less biased than in the case of mean imputation.

3.2. Regression imputation

The main idea of the *regression imputation* is that missing values are replaced with predicted values established using a specific regression equation constructed on the basis of available data for the variable with gaps (as the value of the dependent variable resulting from the regression model) and some fully available auxiliary variables treated as explanatory variables. This approach is aimed at predicting missing values in such a way that the imputed value should be as close as possible to the unknown value. In this context it is very important to include in the model as many explanatory variables as possible (provided, however, they are not strongly correlated). This action can significantly improve the quality of prediction. Such a construction seems to be technically sophisticated and its application requires much more time than many other imputation methods (e.g. in comparison with the situation when the regression equation is constructed separately for each variable to be imputed). Of course, this method does not guarantee that the implants will be fully plausible values, but one can expect that the deviation of implants from their appropriate expectations will be relatively at their lowest.

The basic regression model is given by:

$$Y = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \boldsymbol{\varepsilon} \tag{3}$$

where $Y = (y_1, y_2, \dots, y_n)$ is the target variable with gaps, $X_1, X_2, \dots, X_m$ $(m \in \mathbb{N})$ – auxiliary variables and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$ – the disturbance. OLS estimator of coefficients has the form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_\mathrm{r}^\mathrm{T}\mathbf{X}_\mathrm{r})^{-1}\mathbf{X}_\mathrm{r}^\mathrm{T}\mathrm{Y}_\mathrm{r}$, where $\widehat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m\right)$, $\mathbf{X}_r$ and Y_r are matrix $\mathbf{X} = \left[x_{ij}\right], x_{i0} = 1,\ i = 1,2,\dots,n,\ j = 0,1,2,\dots,m$ and vector Y restricted only to those units $r_1, r_2, \dots, r_q, r_l \in \{1,2,\dots,n\}, l = 1,2,\dots,q, q < n$, for which data on Y are available, respectively[4].

One can formulate now a question: how to choose the auxiliary variable? The basic criterion in this respect should be that the strict connection with the target variable must be retained. But to effectively conduct these types of imputation, each such variable should provide a unique and large information resource. Duplicating information should be avoided. It means that we have to establish such a set of variables which are mutually weakly correlated and simultaneously retain their mutual multivariate connection. This goal may be achieved by using the reversed correlation matrix and its analysis described in Section 2 (step 2). Instead of many auxiliary variables, we can use one, the complex measure (2) containing information provided

[4] Of course, if data on the analysed variable were collected in a sample survey and some population value is to be imputed/estimated, one can include survey weights to the OLS estimator of $\boldsymbol{\beta}$. It will be then of the form $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_r^\mathrm{T}\mathbf{W}_r\mathbf{X}_r)^{-1}\mathbf{X}_r^\mathrm{T}\mathbf{W}_r Y_r$, where $\mathbf{W}_r = \mathrm{diag}(w_{r_1}, w_{r_2}, \dots, w_{r_k})$ is the matrix of sampling weights restricted to those units for which data on Y are available.

by auxiliary variables and taking their connections into account. That is, the model (3) takes the form

$$Y = \beta_0 + \beta_1 \boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{4}$$

In fact, the regression imputation is the generalized ratio imputation.

Next three methods have a specific form. They are based on iterative algorithms generating successive approximates of implants to obtain results of sufficient quality. Therefore, in fact, they are special cases of multiple imputation – an approach consisting of producing a number of complete data sets from the incomplete data by imputing the missing data finite number of times by some assumed model-based method. Then, each completed data set is analysed and the results are combined to achieve final imputed values and related inference (cf. Rubin (1987)).

3.3. Regression imputation with iterative extension

Regression imputation with iterative extension seems to be an efficient improvement of the relevant classical approach. Let $\hat{\sigma}^2\mathbf{V}$ (where $\mathbf{V} = (\mathbf{X}_r^{\mathrm{T}}\mathbf{X}_r)^{-1}$ and $\hat{\sigma}^2$ is the estimated variance of Y) be a covariance matrix for the model with Y being the explained variable in (3). Determination of imputed values for each imputation is performed such that we start from the model (3) and next new parameters $\boldsymbol{\beta}_* = (\beta_{*0}, \beta_{*1}, \dots, \beta_{*m})$ and $\hat{\sigma}_*^2$ are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m)$, $\hat{\sigma}^2$ and $\mathbf{V}$. The variance has the form $\sigma_*^2 = \hat{\sigma}^2(n - m - 1)/g$, where g is a random number from the χ^2_{n-m-1} (chi–square with $n - m - 1$ degrees of freedom) distribution and n is the number of non–missing data in Y. The regression coefficients are computed as $\boldsymbol{\beta}_* = \widehat{\boldsymbol{\beta}} + \sigma_* \mathbf{V}_{(c)}^{\mathrm{T}} Z$, where $\mathbf{V}_{(c)}^{\mathrm{T}}$ is the upper triangular matrix in the Cholesky decomposition of $\mathbf{V}$, i.e. $\mathbf{V} = \mathbf{V}_{(c)}^{\mathrm{T}}\mathbf{V}_{(c)}$ and Z is a vector of $k + 1$ independent random normal variables (cf. Yuan (2010)).

The missing values are then replaced by predictors obtained from the equation

$$Y_r = \beta_{*0} + \sum_{j=1}^{m} \beta_{*j} X_{rj} + z_r \sigma_*, \tag{5}$$

where X_{rj} are the values of covariates for such units for which data on Y are unavailable and z_r is a simulated normal deviate, $r = 1,2, \dots, n$. This operation can then be repeated starting from the formula (5) and so on. The number of iterations depends on the assumptions of the quality control (cf. Rubin (1987), Yuan (2010)). The synthetic measure (2) can be here also efficiently applied instead of the set of (sometimes numerous) covariates enabling a simplification of (5) to the form:

$$Y_r = \beta_{*0} + \beta_{*1} \boldsymbol{\eta}_r + z_r \sigma_*. \tag{6}$$

3.4. Predictive mean matching

Predictive mean matching is a method similar to the regression method with iterative extension except that instead of the main predictive equation for each missing value it imputes an observed value which is closest to the predicted value from the simulated regression model (cf. Yuan (2010), Horton and Lipsitz (2001)).

3.5. Propensity score method

Propensity score method is another way of applying regression imputation suggested by Little and Rubin (2002) and studied by Yuan (2010). The propensity score is understood as the conditional probability of assignment to a particular treatment, given a vector of observed covariates. In this method, the propensity score is generated for each variable with missing values to indicate the probability of that observation being missing. The observations are then grouped on the basis on these propensity scores and an approximate Bayesian bootstrap imputation (cf. Rubin (1987), p. 124) is applied to each group (Lavori *et al.* (1995))[5]. With a monotone missing pattern, the following steps described by Yuan (2010) are used to impute values for each variable Y with missing values:

1. Create an indicator variable Λ with the value 0 for observations with missing Y and 1 otherwise.
2. Fit a logistic regression model

$$\text{logit}(p) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j + \boldsymbol{\varepsilon}, \tag{7}$$

where $p = \Pr(\Lambda = 0 | X_1, X_2, \ldots, X_m)$ and $\text{logit}(p) = \log(p/(1-p))$.

3. Create a propensity score for each observation to estimate the probability that it is missing.
4. Divide the observations into a fixed number of groups (typically assumed to be five) based on these propensity scores. This can be done by arbitrarily establishing some structure of intervals of propensity values and indicating observations whose propensity values belong to such particular intervals.
5. Apply approximate Bayesian bootstrap imputation to each group. That is, for a given group, suppose that Y_{obs} denotes the n_1 observations with nonmissing Y values and Y_{mis} denotes the n_0 observations with missing Y (where $n_1 > n_0$). Approximate Bayesian bootstrap imputation first draws n_1 observations randomly with replacement from Y_{obs} to create a new data set Y^*_{obs}. This is a nonparametric analogy of drawing parameters from the posterior predictive

[5] Of course, in the propensity score method not only Bayesian bootstrap can be used. This procedure is, however, most popular and seems to be efficient.

distribution of the parameters. The process then draws the n_0 values for Y_{mis} randomly with replacement from Y^*_{obs}. These values are implants.

Steps 1 through 5 are repeated sequentially for each variable with missing values. In our analysis also all auxiliary variables will be replaced with one synthetic measure (2), i.e. the formula (7) will take the form

$$\text{logit}(p) = \beta_0 + \beta_1 \boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{8}$$

Yuan (2010) noted that the propensity score method was originally designed for a randomized experiment with repeated measures on the response variables. The goal was to impute the missing values to the response variables. The method uses only the covariate information that is associated with objects for which the imputed variable values are missing. It does not use correlations among variables. It is effective for inferences about the distributions of individual imputed variables, such as univariate analysis, but it is not appropriate for analyses involving relationships among variables, such as regression analysis (cf. Schafer (1999), p. 11). It can also produce badly biased estimates of regression coefficients when data for predictor variables are missing (cf. Allison (2000)).

4. MSE and its decomposition based on imputed data

The value of the imputation used should be evaluated using relevant methods of the quality control. That is, we have to assess whether the imputed values are best fitted and how the estimation precision of the population statistics using imputed data is influenced by adding or not adding disturbance terms to some models of imputations. This assessment can be done both at the stage of a preliminary simulation study or *ex post*, i.e. after performing the whole imputation process. Of course, in the latter case, it is much more difficult due to a shortage of material for efficient comparisons. Now, we will present how to estimate the Mean Squared Error (being a basic measure in this situation) for aggregated statistics based on imputed data.

Let A be the set of units in the sample, $\hat{\theta}_A$ be an estimator of θ computed using all sample data about the target variable. The variance estimation is strictly connected with θ. In general, Särndal (1992) showed that the total variance or – in terms of the theory of estimation – MSE of the estimator $\hat{\theta}$ of θ for the whole population[6], $\hat{V} = E\left(\hat{\theta} - \theta\right)^2$, can be decomposed in the sampling, imputation and mixed effect components:

$$\hat{V} = \hat{V}_{\text{SAM}} + \hat{V}_{\text{IMP}} + 2\hat{V}_{\text{MIX}}, \tag{9}$$

[6] In many cases MSE coincides with variance of estimator. However, in a missing–data context due to a bias these two quantities cannot be equivalent. So, here we will use MSE as more informative index of quality.

where $\hat{V}_{\mathrm{SAM}} = E\left(\hat{\theta}_A - \theta\right)^2$, $\hat{V}_{\mathrm{IMP}} = E\left(\hat{\theta} - \hat{\theta}_A\right)^2$, $\hat{V}_{\mathrm{MIX}} = E\left(\left(\hat{\theta}_A - \theta\right)\left(\hat{\theta} - \hat{\theta}_A\right)\right)$ are the aforementioned components, respectively. The imputation term $\hat{V}_{\mathrm{IMP}}$ shows the part of variance resulting from expected deviation of imputed values from the true ones. Of course, in a simulation study, these expected values can be approximated by arithmetic means of relevant deviations obtained by the consecutive replications of sampling. J. K. Kim *et al.* (2006) analyse the problem of estimation of MSE in the complex sample design, when imputation is repeated q times, and derive a formula expressing the difference between the expected value of multiple imputation MSE estimation and the MSE of estimator $\hat{\theta}$. They have investigated such models for subpopulations (called also domains) and linear regression models. Arcaro and Yung (2001) propose approximately unbiased statistics for $\hat{V}_{\mathrm{SAM}}$, $\hat{V}_{\mathrm{IMP}}$ and $\hat{V}_{\mathrm{MIX}}$ using weighted mean and weighted ratio imputation with weights derived from traditional generalized regression estimator (GREG) for the mean based on relevant auxiliary data. Alternatively, they have analysed the MSE estimator for weighted ratio imputation with specially adjusted jackknife GREG weights as $\hat{V} = \sum_{j=1}^{h} \frac{n_j}{n_j - 1}\left(\hat{\theta}_j - \hat{\theta}\right)^2$ where $\hat{\theta}_j$ is the imputation estimator of θ, corrected using jackknife weights in the j–th stratum, n_j is the number of units belonging to this stratum, $j = 1,2, \ldots, h$, and $h \in \mathbb{N}$ is the number of strata in the sampling design. Of course, this algorithm is also efficient if instead of strata repetitions of simple random sampling in the simulation study are used, despite the fact that the samples could not be disjoint. Assuming that $\hat{\theta}$ is unbiased, Kim (2000) proposes unbiased unweighted MSE estimators for regression and ratio imputation models. Fuller and Kim (2005) proved the formula for fully efficient fractionally imputed MSE estimator based on the squared deviation of mean estimator and response probabilities, in particular imputation cells and subpopulations. Similar research for the balanced random imputation has been conducted by Chauvet *et al.* (2011).

In the case of the *ex post* quality control, we have a much more serious problem. Because there is no exact reference platform (the 'true' distribution of the target variable is unknown), it is necessary to rely only on an approximate estimation of imputation precision using approaches (cf. e.g. Andridge and Little (2010), who divide a sample into several complete data sets and, instead of repeated sampling, investigate a division of the population into complete and disjoint data sets and then estimate the MSE). This division might be based on an auxiliary variable (most preferably categorical) strictly connected with the target one. These classes should have approximately equal number of elements. Hence, we can obtain an estimate of error.

However, using decomposition (9) seems to be a better solution. That is, assuming that the units were sampled independently and $\hat{\theta}_A = \sum_{i \in A} y_i^* / |A|$, the MSE will be approximated by

$$\tilde{V} = \frac{1}{|A|^2} \sum_{i \in A} \left(y_i^* - \hat{\theta}_A\right)^2 = \tilde{V}_{\mathrm{SAM}} + \tilde{V}_{\mathrm{IMP}} + 2\tilde{V}_{\mathrm{MIX}}, \tag{10}$$

where the relevant components are estimated using the following statistics:

- sampling effects

$$\tilde{V}_{\text{SAM}} = \frac{1}{|A|^2}\sum_{i\in A}\left(\tilde{y}_i - \hat{\theta}_A\right)^2, \tag{11}$$

- imputation effects

$$\tilde{V}_{\text{IMP}} = \frac{1}{|A|^2}\sum_{i\in A}(y_i^* - \tilde{y}_i)^2, \tag{12}$$

- mixed effects

$$\tilde{V}_{\text{MIX}} = \frac{1}{|A|^2}\sum_{i\in A}(y_i^* - \tilde{y}_i)\left(\tilde{y}_i - \hat{\theta}_A\right), \tag{13}$$

where $y_i^* = py_i + (1-p)\hat{y}_i$ and $\tilde{y}_i = py_i + (1-p)\left(\left(n\hat{\theta}_A - |U|\hat{\theta}_U\right)/(n-|U|)\right)$, with $p = 1$ if the value of Y for i–th unit is available and $p = 0$ otherwise, $\hat{\theta}_U = \sum_{i\in U}\hat{y}_i/|U|$ is the estimator of θ obtained on the basis of imputation for the set of units for which data on Y are unavailable (U), where $|\cdot|$ denotes the cardinality of a given set. The value $\tilde{y}_i$ is equal to y_i if y_i is available and to mean of known values of Y otherwise. Thus, we can perform detailed diagnostics of our imputation method.

The formulas (10) – (13) are designed for single imputation. Instead, one can also use multiple imputation in which case the variance can be imputed using the well-known pooling rules by Rubin (1987). First of them is the within–imputation variance – the average of the mean of the within variance estimate, i.e. squared standard error – which reflects the sampling variance, i.e. the precision of the parameter of interest in each completed data set:

$$\hat{V}_w = \frac{1}{l}\sum_{i=1}^{l}\widehat{SE}_i^2,$$

where l is the number of imputed data sets and $\widehat{SE}_i$ – estimate of the sum of squared standard errors observed in i–th imputed data set, $i = 1,2,\dots,l$. The smaller sample, the larger the within–imputation variance.

The second rule concerns the between–imputation variance, which reflects the extra variance occurring due to the missing data and is estimated by taking the variance of the parameter of interest estimated over imputed data sets. It is computed using the formula

$$\hat{V}_b = \sqrt{\frac{\sum_{i=1}^{l}\left(\hat{\theta}_i - \hat{\bar{\theta}}\right)^2}{l-1}},$$

where $\hat{\theta}_i$ denotes the estimate of in i–th imputed data set and $\hat{\bar{\theta}} = \sum_{l=1}^{l}\hat{\theta}_i/l$ is the pooled estimate of θ, $i = 1,2,\dots,l$. The higher the level of missing data, the larger the between–imputation variance.

The total variance is then given by

$$\hat{V}_T = \hat{V}_w + \left(1 + \frac{1}{l}\right)\hat{V}_b.$$

5. Simulation study

To verify the efficiency of our method a simulation experiment was conducted. A sample consisting of 200 units was constructed to account for circumstances observed in business statistics. We have assumed that the target variable Y and three auxiliary variables X_1, X_2 and X_3 are considered. Their values were drawn jointly from the multivariate log–normal distribution. This sampling was realized by generating random vectors from the multivariate normal distribution with the vector of expected values $\mu = (3.5, 0.1, 1.3, -4.0)$ and covariance matrix of the form

$$\mathbf{\Sigma} = \begin{bmatrix} 2.5 & 1.0 & -1.7 & 2.4 \\ 1.0 & 1.3 & 0 & 0 \\ -1.7 & 0 & 3.4 & 0 \\ 2.4 & 0 & 0 & 6.7 \end{bmatrix}$$

and next exponentiating obtained results. Thus, Y is represented by first coordinate of any such vectors and X_1, X_2, X_3 – by second, third and fourth, respectively. Such an attempt is motivated by the following premises:

- most experts experienced in business statistics argue that the log–normal distribution is the best way to approximate the distribution of variable occurring in this field,
- the auxiliary variables should be uncorrelated each with other, but they all should be clearly correlated with the target variable; in the investigated case the expected Pearson correlation coefficients of X_1, X_2 and X_3 with Y are 0.5547, -0.5831 and 0.5864 respectively,
- in commonly used professional statistical software (SAS, R, etc.) the random number/vectors from log–normal distribution can be generated only by exponentiating values/vectors drawn only from the normal distribution. Moreover, in the multivariate case it is required that covariance matrix is positive definite[7].

In practice, the possibilities of establish and that all aforementioned conditions are simultaneously satisfied is slightly restricted. We have chosen the best possible solution in this situation.

Another problem was connected with the modelling of non–response. That is, it should be decided how to choose records for which data on Y are assumed to be missing and have to be imputed. Of course, the simplest way to do it seems to be taking a random subsample of the original data and drop values of Y for them (it is the so–called *Missing Completely at Random* – MCAR condition, according to the terminology

[7] It is not difficult to prove that the $n \times n$ covariance matrix is arrowhead (i.e. it has non zero diagonal and first row and first column entries), positive definite and produces Pearson's correlation matrix, whose entries in the first row and the first column are greater (in terms of absolute values) than 0.5 only if $n \leq 4$.

introduced by Little and Rubin (2002)). Unfortunately, this attempt is rarely used in practice – mainly due to the fact that its application leads usually to unbiased estimation. Thus, also differences between various methods in this context can be just random sampling fluctuations. Therefore, the MAR (*Missing at Random*) scenario, where it is assumed that the missing data mechanism depends on the auxiliary (X) variables, could be a more appropriate solution here. That is, the missingness can be explained by variables, for which full information is available. It makes MAR different than MCAR, where it is assumed that gaps in data result from random results. According to the relevant proposals, which can be found in literature (cf. e.g. Pampaka *et al.* (2016)), we use the logit model here, in which the importance of particular auxiliary variable is taken into account, i.e.

$$\text{logit}(p_i) = -0.5 + 0.3x_{i1} - 1.2x_{i2} + 0.2x_{i3},$$

where p_i is the probability that the value y_i is missing and x_{ij} denotes the value of X_j for i–th object, $j = 1,2,3$. The value y_i was regarded as missing if $\hat{p}_i \geq 0.5$ where $\hat{p}_i$ is the estimate of p_i obtained from this model, $i = 1,2,\dots,n$. The structural parameters (-0.5, 0.3, -1.2 and 0.2) in this model were established a priori so that to ensure various connections of auxiliary variables with possibility of lack of data on Y and simultaneously usually reasonable expected number of such missing items.

The following imputation methods were used:

- ratio imputation (denoted as R),
- ratio imputation with the complex measure (RM),
- regression imputation (RG),
- regression imputation with complex measure based on (4) (RGM),
- regression imputation with iteration (RGI),
- regression imputation with iteration based on complex measure using (6) (RGIM),
- predictive mean matching (PMM),
- predictive mean matching with the complex measure (PMMM),
- propensity score method (PSM),
- propensity score method with the complex measure using (8) (PSMM).

The classical ratio imputation was conducted using such an auxiliary variable which was best correlated with Y (in terms of available data). The complex measure used instead of the set of three independent auxiliary variables was determined using the formula (2) based on the normalization using the Weber median (1), taxonomic benchmark and distance of units from the benchmark indicated in steps 5 and 6 of the procedure described in Section 3. The experiment was conducted using especially constructed algorithm prepared in the SAS Enterprise Guide 4.3. software (and especially its IML environment). In the case of the regression, predictive mean

matching and propensity score method the mi procedure was used. In the case of RGI, RGIM, PMM and PMMM methods 10 iterations and for PSM and PSMM – 2 iterations were done. It was sufficient to ensure relevant quality of imputed data. The whole experiment was replicated 1000 times. It is worth noting that each replication covered both drawing new samples and generating the missing value.

Using the results of imputation we have computed the analysed measures of quality. That is, we have used the maximum distance from imputed values and the following measures of quality of estimation of expected value of Y using the empirical arithmetic mean of the available and imputed values: bias, MSE, estimated MSE (using only available data for Y supplemented with the imputed ones, formula (10)), components of the MSE (according to (10)) and the estimated MSE (formulas (11), (12), (13)).

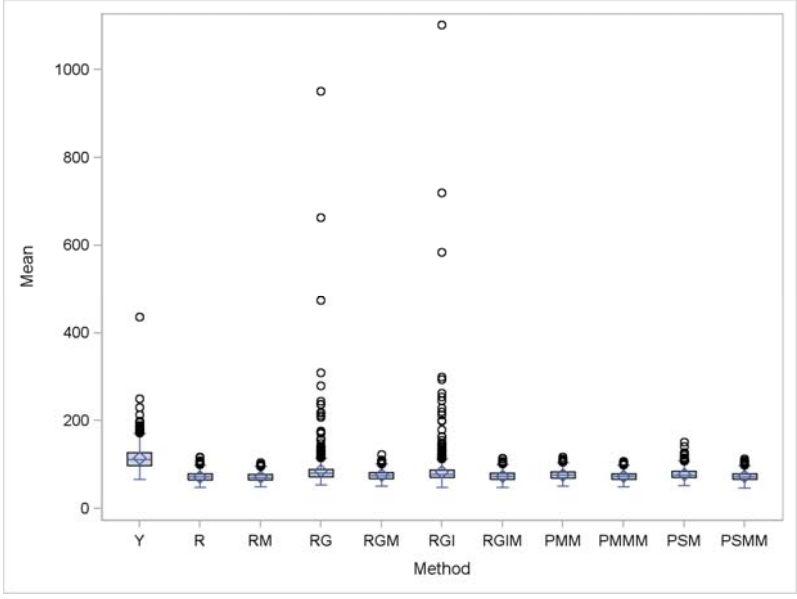


Figure 1. Box–and–whisker visualisation of distributions of the target variable Y means – original and with imputed values – by methods of imputation.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

Figure 1 presents box–and–whisker plots reflecting the distribution of means of original, complete, values of Y and means of modified Y where modelled missing values are replaced with implants generated using particular methods presented above. These means were computed for relevant units over 1000 replications. One can observe that using the complex measure instead of the simple collection of auxiliary variables a substantial reduction of outlying (expressed especially by extreme values of the means) results of regression imputation (also with iteration) is achieved. Such an improvement in this context is considerable also for the propensity score method. In any case, the population mean of Y seems to be slightly underestimated, but this bias is not great.

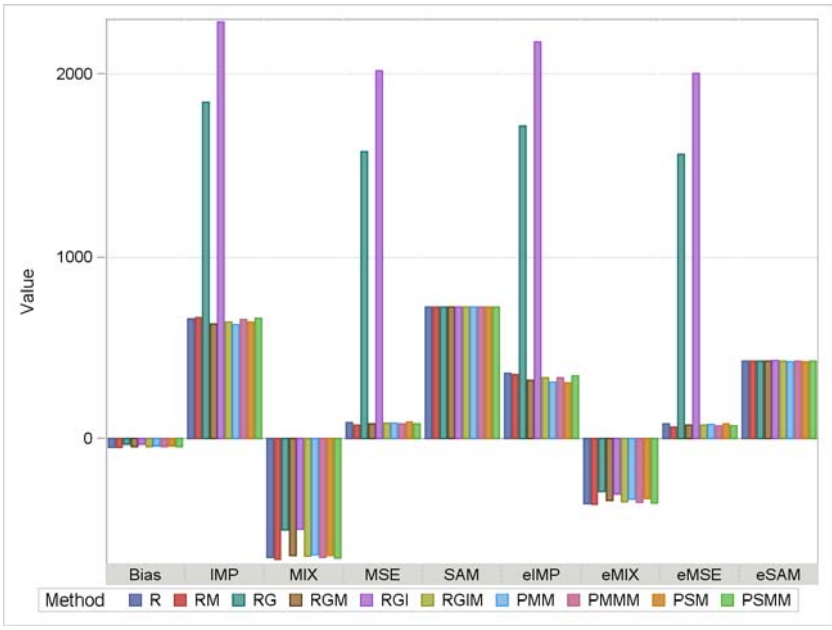


Figure 2. Comparison of quality indicators for estimation of mean of Y using original and imputed data by methods of imputation.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

The results of the experiment in terms of quality indicators of imputation are presented in Figure 2. It contains the average values of these indices computed over 1000 trials. MSE, SAM, IMP and MIX denote here the mean square error and its sampling, imputation and mixed effect components and eMSE, eSAM, eIMP and eMIX their estimators, computed using formulas (10), (11), (12) and (13), respectively. One can observe that using the complex measure substantially improves the quality of imputation for R, RG and RGI methods. Such an improvement is especially considerable in terms of MSE and its imputational component. For the remaining methods this advantage is lower, but also observable. The sampling component (SAM) does not depend on the imputation methods, by the assumption. Some very small increase in IMP and MIX (in absolute values) components after using the complex measure in R, PMM and PSM methods may be a result of slightly higher bias of imputed values in this case, which was observed already in Figure 1.

It would be also interesting to employ the imputation methods when the complex measure (2) is used and when all variables are used, but in application of imputation methods based on individual variables, these variables are normalized according using the formula (1). Figures 3 and 4 show distribution of target variable means and quality indicators obtained after 1000 replications made according to the above described principles.

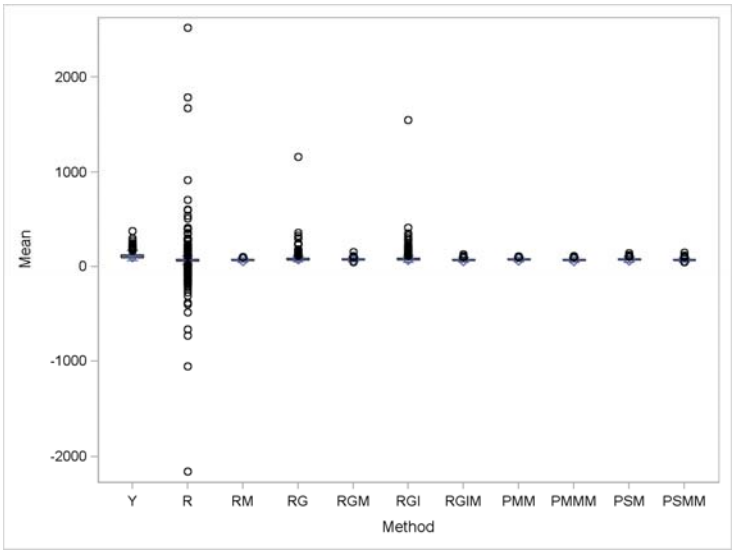


Figure 3. Box–and–whisker visualisation of distributions of the target variable Y means – original and with imputed values – by methods of imputation, variables normalized using the Weber median.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

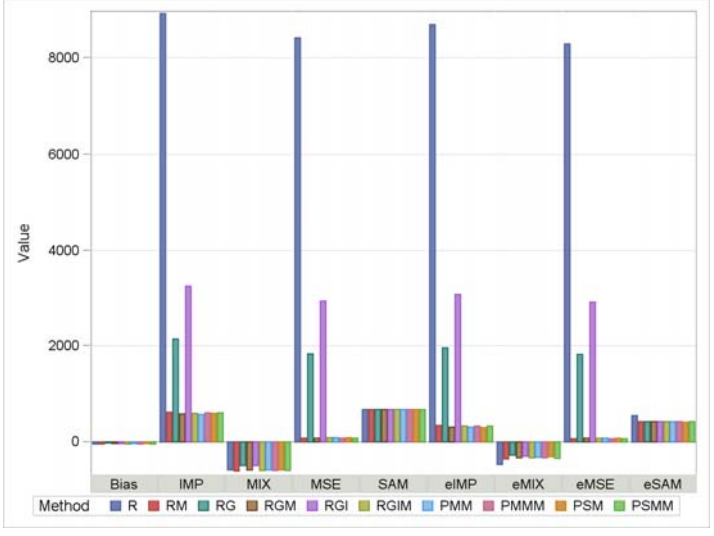


Figure 4. Comparison of quality indicators for estimation of mean of Y using original and imputed data by methods of imputation, variables normalized using the Weber median.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

One can observe that majority of our previous observations was confirmed. That is, the reduction of imputation error when using the complex measure instead of

variables normalized using the Weber median is especially considerable for R, RG and RGI methods. In other cases, the advantage of such replacement is much smaller. The imputation based on the complex measure allows also for reduction of distorting outliers in the imputed values. These results lead to the conclusion that for some methods the use of the complex measure is more efficient than when the individual variables in model-based imputation are used – independently from that whether these variables are normalized or not.

6. Empirical study

Our alternative study was based on data on 36 firms representing the IT sector in Bermuda, Canada, China, Denmark, Finland, Germany, Greece, Indonesia, Japan, Mexico, Russia, South Korea, Sweden, the UK and the USA, stored on Instructional Web Server of the California State University in Los Angeles, USA (http://instructional1.calstatela.edu/mfinney/Courses/491/hand/sas_exercise/tech3.xls). Here, the following five variables are recorded: Return on Equity (ROE, %), Revenues (in million $), Revenue Growth (RGR, %), Total Shareholder Return (TSR, %) and Profits (PR, in million $).

The set originally contained 39 firms, but due to missing data for ROE three of them had to be dropped. For the purposes of the study, the revenues were chosen as the variable to be imputed. To implement the non-response we use – similarly as in the case of the simulation study (Section 5) – the MAR condition for missing data. Two independent options of the logit model of missingness of data on revenues were applied. The first of them takes the correlation of available variables with revenues into account, the second one underlies the variability of ROE, RGR, TSR and PR. On the other hand, we had also to remember about upper limits of computational capability of the software (connected with exponentiating used in determination of estimates of the missingness probability). Finally, the analysed logit models had the following forms:

- option 1: $\text{logit}(p_i) = 0.001ROE_i + 0.5RGR_i - 1.1TSR_i - 0.004PR_i$,
- option 2: $\text{logit}(p_i) = 0.27ROE_i + 0.005RGR_i - 0.005TSR_i - 0.04PR_i$.

Again, y_i is regarded as missing when $\hat{p}_i \geq 0.5$, $i = 1,2, \ldots, n$. The use of the option 1 generated 7 gaps in data whereas the option 2 resulted in 8 non–response items.

Similarly as in Section 5, we have now constructed the complex measure which will serve as a support for imputation. It was based on three indicators which are strongly diversified and weakly correlated with the revenues, selected according to Step 2 of the procedure described in Section 1, i.e. ROE, RGR and TSR. The variation of all possible variables was very high (i.e. the coefficient of variation amounted to 77.2% for ROE, 96.2% for RGR, 137.8% for TSR and 299.2% for PR). The diagonal entries of the inverse of correlation matrix are respectively: 1.5149, 1.9651, 1.6123 and 1.2122. Hence, the

possible auxiliary variables are mutually weakly correlated. Because PR is weaker correlated with the target variable than others (the relevant Pearson's correlation coefficient amounted to 0.1905, whereas for the remaining variables it exceeded 0.21 or – for TSR – even 0.3), we decided to remove it. Thus, we have obtained the set of diagnostic features. All of them are stimulants for revenues. We have computed the Weber median for them. Its coordinates amounted to 21.5131, 21.9906 and 26.9717 respectively. Next, – according to Step 4, we have normalized the diagnostic features using the formula (1). These normalized values were the basis of computation of the final complex measure by determination of the taxonomic benchmark of development (step 5), which was described by the vector (3.5726, 2.4161, 5. 3996), computation of distance of entities from the benchmark (Step 6) and the values of the complex measure (Step 7, formula (2)).

To impute simulated missing data, the same methods as in Section 5 were applied, i.e. ratio imputation (R), regression imputation (RG), regression imputation with iteration (RGI), predictive mean matching (PMM) and propensity score method (PSM), each of them also with the option based on the complex measure (RM, RGM – with formula (4), RGIM – formula (6), PMMM and PSMM – with formula (8), respectively). In classical ratio imputation the Total Shareholder Return was used as a reference variable, because it is the one most correlated with the target variable.

In Figure 5, the means of revenues and their relevant 95% confidence intervals for original (Y) and imputed data – when the option 1 is applied – are visualised. The vertical dashed line shows the true mean of Y. We can observe that the use of the complex measure in the case of ratio imputation substantially improves the precision of mean estimation. Better adjustment in this context is considerable also for RG and PMM when explanatory variables are replaced with the complex measure.

Table 1 shows the comparison of basic statistics describing the shape of original distribution of Y with complete data and distribution of Y where data gaps were filled up by imputed values obtained using a relevant imputation method. One can notice here that differences between respective extremes (minimums and maximums), median and quartiles of the original variable and its imputed version confirm our earlier conclusions to a very large extent. It is possible that many of them are caused by existing incidentally very small or very high values of imputed values, which is the integral part of risk connected with any imputation. The relatively high coefficient of variation seems to be a good justification of this view. Moreover, the skewness and kurtosis of distribution of 'complete' Y is better approximated by option based on the complex measure for the predictive mean matching (PMMM) and diversification (expressed by CV=136.2%, whereas for Y CV=130.4%) – by the propensity score method (PSMM).

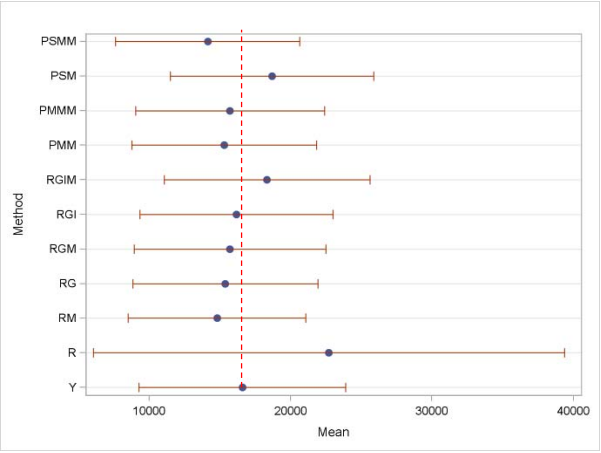


Figure 5. Means and their 95% confidence intervals for original and imputed data on revenues (in million $) – option 1 of missing data model.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

Respective results of the use of option 2 in modelling data missingness are presented in Figure 6 and Table 2. They show again that most of the imputation methods underestimate the mean of Y. The use of the complex measure improves the quality of estimation of the mean in the case of ratio, regression and propensity score methods. Also, in this case the PMMM approach better approximates skewness and kurtosis whereas PSMM – CV and skewness of the original values (Y) than PMM and PSM, respectively.

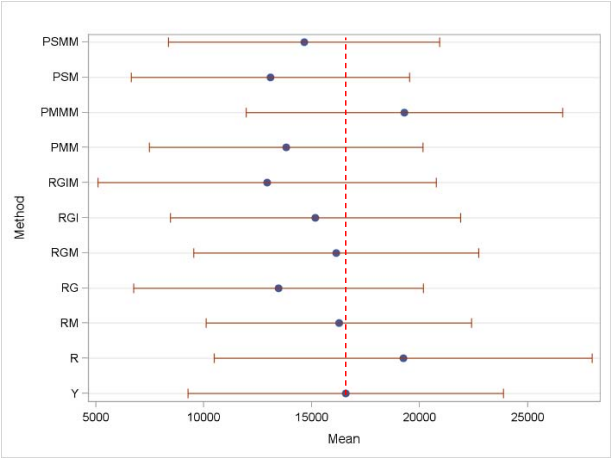


Figure 6. Means and their 95% confidence intervals for original and imputed data on revenues (in million $) – option 2 of missing data model.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

Of course, in contrast to the simulation study, data used here are arbitrary and fixed. Hence, the coefficients in the MAR condition formulas are established once and, in consequence, the imputation was done also once. Since most of the analysed imputation methods are stochastic and the result of one execution is random, it is difficult to formulate general conclusions on this basis. However, we have taken one more trial which allowed for an increase of random contribution to MAR condition and can to some extent verify the aforementioned observations.

In the second attempt for imputation based on individual variables we have taken their values normalized using the Weber median according to the formula (1). Moreover, we ensure that the number of removed (and imputed) values of the target variable will be relatively large, but not too large. More precisely, we assumed that the number of removed values should be not smaller than 11 and nor greater than 17. The MAR condition was still logit: $\text{logit}(p_i) = \beta_1 \cdot ROE_i + \beta_2 \cdot RGR_i + \beta_3 \cdot TSR_i - \beta_4 \cdot PR_i$, where – to take the correlation of auxiliary variables with the target one – β_1 and β_2 were sampled from the uniform $U\left(-\left(0.7 \cdot \frac{k}{1000}\right), 1 - \left(0.7 \cdot \frac{k}{1000}\right)\right)$ distribution whereas β_1 and β_1 – from $U\left(0.7 \cdot \frac{k}{1000}, 1 + \left(0.7 \cdot \frac{k}{1000}\right)\right)$ where k= 2349 and takes the following values: $\beta_1 = 0.0000399$, $\beta_2 = 3.7883793$, $\beta_3 = 2.44456$ and $\beta_4 = 0.5172775$. This way, we have obtained data with 15 gaps. Figure 7 shows the comparison of means and their confidence intervals in this case.

One can observe that the application of the complex measure in the model-based imputation allows often for tightening the confidence imputation for means. We have also repeated this simulation for some other k, which allows for satisfaction of all above described assumptions and the results were quite similar.

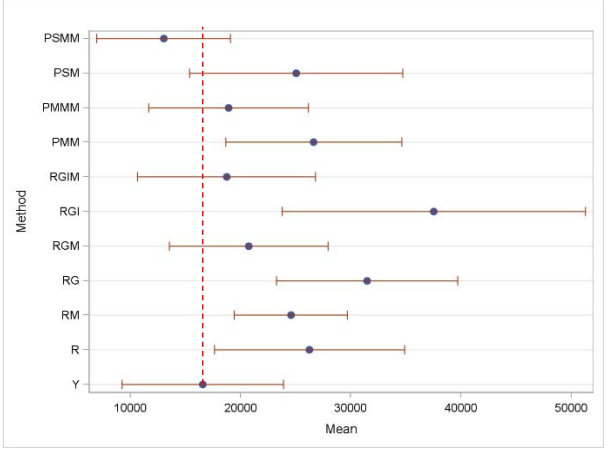


Figure 7. Means and their 95% confidence intervals for original and imputed data on revenues (in million $) – data normalized using the Weber median and random coefficients in MAR.

Source: Author's work using the SAS Enterprise Guide 4.3 software (with IML environment).

7. Conclusions

The main conclusions which can be formulated on the basis of our studies are as follows. Construction of a complex measure using ordinal statistics (of which the Weber median) ensures a more efficient exploitation of mutual connections between possible auxiliary variables and therefore more informative imputation (cf. Młodak (2006, 2014)). Using the complex measure instead of one or more auxiliary variables is especially favourable for ratio and regression imputation, where the improvement in quality expressed by bias or Mean Squared Error is usually large. In many cases similar observation can be done for the regression imputation with iteration. For the predictive mean matching and propensity score method replacing the summarized predictor variables themselves by the complex measure gives not such considerable advantages – probably due to additional correction mechanism built into these procedures (i.e. indicating best 'donor' of imputed value in the former and grouping observations and bootstrap in the latter case). However, also in these cases using the complex measure can lead to better reflection of some important characteristics of distribution of the original variable, i.e. variation, skewness or kurtosis, which describe its shape.

It is worth noting that the efficiency of particular imputation algorithms in the simulation study depends, among other things, on the MAR condition being regarded as inherently strong. The logit model used in the presented simulation study assumes that the occurrence of data gaps of the target variable depends on all fully available variables. Such an attempt enables to exploit connections between analysed variables (both with incomplete and complete data) and, in consequence, the whole information potential of the database. It then proves to be more systematic than random reasons of the gaps.

The complex measure provides more stable results, i.e. gives substantially lower risk of excessive outliers. However, one should remember that the conditions for the efficient use of the complex measure are: proper choice of auxiliary variables on the basis of which it is constructed and the method of its construction. The choice of model–based imputation method which has to be applied with the complex measure depends on the main aim of the imputation: for estimation of mean or similar population statistics ratio or regression imputation is recommended; if a scientist is more interested in approximation of the shape of unknown distribution, the predictive mean matching or the propensity score method can provide better effects for him/her.

Acknowledgement

The author is very grateful to two anonymous reviewers for valuable comments and suggestions, which substantially contributed to improvement of quality of this study.

References

ALLISON, P. D., (2000). Multiple Imputation for Missing Data: A Cautionary Tale, Sociological Methods and Research, Vol. 28, pp. 301–309.

ANDRIDGE, R. R. and LITTLE, R. J. A., (2010). A Review of Hot Deck Imputation of Survey Non–response, International Statistical Review, Vol. 70, pp. 40–64.

ARCARO, C. and YUNG, W., (2001). Variance estimation in the presence of imputation, SSC Annual Meeting, Proceedings of the Survey Method Section, pp. 75–80.

CHAUVET, G., DEVILLE, J.–C. and HAZIZA, D., (2011). On Balanced Random Imputation in Surveys, Biometrika, Vol. 98, pp. 459–471.

DE WAAL, T., PANNEKOEK, J. and SCHOLTUS, S. (2011). Handbook of Statistical Data Editing and Imputation, Wiley Handbooks in Survey Methodology, John Wiley & Sons, Inc., Hoboken, New Jersey.

DUROCHER, S. and KICKPATRICK, D., (2009). The projection median of a set of points, Computational Geometry, Vol. 42, pp. 364–375.

HORTON, N. J. and LIPSITZ, S. R., (2001). Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables, Journal of the American Statistical Association, Vol. 55, pp. 244–254.

HUNDEPOOL, A., DOMINGO–FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K., DE WOLF, P.–P., (2012). Statistical Disclosure Control, Series: Wiley Series in Survey Methodology, John Wiley & Sons, Ltd.

JOLLIFFE, I. T. (2002). Principle Component Analysis. Second Edition. Springer – Verlag, New York, Berlin, Heidelberg.

KIM, K., (2000). Variance estimation under regression imputation model, Proceedings of the Survey Research Methods Section, American Statistical Association.

KIM, J. K., BRICK, M., FULLER, W. A. and KALTON, G., (2006). On the bias of the multiple-imputation variance estimator in survey sampling, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 68, pp. 509–521.

LAVORI, P. W., DAWSON, R. and SHERA, D., (1995). A Multiple Imputation Strategy for Clinical Trials with Truncation of Patient Data, Statistics in Medicine, Vol. 14, pp. 1913–1925.

LITTLE, R. J. A. and RUBIN, D. B., (2002). Statistical Analysis with Missing Data. Second Edition, John Wiley & Sons, Inc., New York.

MALINA, A. and ZELIAŚ, A., (1998). On Building Taxonometric Measures on Living Conditions, Statistics in Transition, Vol. 3, No. 3, pp. 523–544.

MILASEVIC, P. and DUCHARME, G. R., (1987). Uniqueness of the Spatial Median, The Annals of Statistics, Vol. 15, No. 3, pp. 1332–1333.

MŁODAK, A., (2014). On the construction of an aggregated measure of the development of interval data, Computational Statistics, Vol. 29, pp. 895–929.

MŁODAK, A., (2006). Multilateral normalisations of diagnostic features, Statistics in Transition, vol. 7, pp. 1125–1139.

NETER, J., WASSERMAN, W. and KUTNER, M. H., (1985). Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs, 2nd edition, Homewood, IL: Richard D. Irwin, Inc., U.S.A.

PAMPAKA, M., HUTCHESON, G. and WILLIAMS, J., (2016). Handling missing data: analysis of a challenging data set using multiple imputation, International Journal of Research & Method in Education, vol. 39, No. 1, pp. 19–37.

ROUSSEEUW, P. J. and LEROY, A. M., (1987). Robust Regression and Outlier Detection, ed. by John Wiley & Sons, New York.

RUBIN, D. B., (1987). Multiple Imputation for Nonresponse in Surveys, John Wiley & Sons, New York.

SÄRNDAL, C. E. (1992). Methods for estimating the precision of survey estimates when imputation has been used, Survey Methodology, vol. 18, pp. 241–252.

SCHAFER, J. L., (1997). Analysis of Incomplete Multivariate Data, New York: Chapman and Hall.

TIBSHIRANI, R., (1996). Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B (Methodological), Vol. 58, No. 1, pp. 267–288.

VANDEV, D. L., (2002). Computing of Trimmed L1 – Median, Laboratory of Computer Stochastics, Institute of Mathematics, Bulgarian Academy of Sciences, (preprint), available at

http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/papers/aspap.pdf .

YUAN, Y. C., (2010). Multiple Imputation for Missing Data: Concepts and New Development (Version 9.0), SAS Institute Inc, Rockville, MD, U.S.A.

ZELIAŚ, A., (20042). Some Notes on the Selection of Normalization of Diagnostic Variables, Statistics in Transition, vol. 5, No. 5, pp. 787–802.

APPENDIX

Table 1. Descriptive statistics of distribution of original and imputed revenues – option 1 of missing data model

Specification	Minimum (in million $)	Lower quartile (in million $)	Median (in million $)	Upper quartile (in million $)	Maximum (in million $)	Coefficient of Variation (%)	Skewness	Kurtosis
Y	623.10	2479.85	7928.40	22353.50	83221.00	130.4	1.7949	2.5318
R	-113226.99	1961.50	8270.11	28538.50	189397.87	216.9	1.2942	5.1495
RM	623.10	2528.20	8559.45	18095.47	83221.00	125.7	2.2671	5.5879
RG	-1356.38	1287.05	7928.40	25849.59	83221.00	125.9	1.9174	4.0467
RGM	-16694.32	1961.50	8559.45	26042.13	83221.00	127.4	1.6282	3.2586
RGI	-12566.21	1961.50	8559.45	28412.30	83221.00	125.0	1.6147	2.8658
RGIM	623.10	2528.20	8955.52	29907.79	83221.00	117.4	1.4907	1.5600
PMM	623.10	2528.20	7863.75	25702.00	83221.00	126.2	1.9797	4.1440
PMMM	623.10	1961.50	7863.75	28538.50	83221.00	125.7	1.8569	3.4739
PSM	623.10	2528.20	8559.45	31375.00	83221.00	113.9	1.4717	1.7915
PSMM	623.10	1128.60	6834.45	19005.00	83221.00	136.2	2.1422	4.8090

Source: Own work using the algorithm written in SAS Enterprise Guide 4.3 (with IML environment).

Table 2. Descriptive statistics of distribution of original and imputed revenues – option 2 of missing data model

Specification	Minimum (in million $)	Lower quartile (in million $)	Median (in million $)	Upper quartile (in million $)	Maximum (in million $)	Coefficient of Variation (%)	Skewness	Kurtosis
Y	623.10	2479.85	7928.40	22353.50	83221.00	130.4	1.7949	2.5318
R	623.10	4161.80	9485.78	20753.68	126578.72	134.4	2.7496	8.6040
RM	623.10	4161.80	9939.75	19445.90	83221.00	111.7	2.2160	5.5749
RG	-23310.83	2479.85	7863.75	18391.27	83221.00	147.4	1.8536	4.5729
RGM	-8927.63	2530.55	9276.10	26041.16	83221.00	120.8	1.7166	3.6362
RGI	-4920.05	1961.50	7928.40	26399.92	83221.00	131.1	1.8004	3.4871
RGIM	-28352.42	1106.80	7863.75	22353.50	83221.00	179.1	1.1556	1.8188
PMM	623.10	2479.85	7983.90	14172.45	83221.00	135.4	2.3786	5.8865
PMMM	623.10	4161.80	9939.75	28538.50	83221.00	112.3	1.7784	2.9581
PSM	623.10	1287.05	5168.95	14172.45	83221.00	145.6	2.3500	5.6783
PSMM	623.10	3244.15	8020.10	19005.00	83221.00	126.7	2.3127	5.7103

Source: Own work using the algorithm written in SAS Enterprise Guide 4.3 (with IML environment).

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 29–54, DOI 10.21307/stattrans-2021-002
Received – 31.12.2019; accepted – 07.09.2020

Modelling and forecasting monthly Brent crude oil prices: a long memory and volatility approach

Remal Shaher Al-Gounmeein[1], **Mohd Tahir Ismail**[2]

ABSTRACT

The Standard Generalised Autoregressive Conditionally Heteroskedastic (sGARCH) model and the Functional Generalised Autoregressive Conditionally Heteroskedastic (fGARCH) model were applied to study the volatility of the Autoregressive Fractionally Integrated Moving Average (ARFIMA) model, which is the primary objective of this study. The other goal of this paper is to expand on the researchers' previous work by examining long memory and volatilities simultaneously, by using the ARFIMA-sGARCH hybrid model and comparing it against the ARFIMA-fGARCH hybrid model. Consequently, the hybrid models were configured with the monthly Brent crude oil price series for the period from January 1979 to July 2019. These datasets were considered as the global economy is currently facing significant challenges resulting from noticeable volatilities, especially in terms of the Brent crude prices, due to the outbreak of COVID-19. To achieve these goals, an R/S analysis was performed and the aggregated variance and the Higuchi methods were applied to test for the presence of long memory in the dataset. Furthermore, four breaks have been detected: in 1986, 1999, 2005, and 2013 using the Bayes information criterion. In the further section of the paper, the Hurst Exponent and Geweke-Porter-Hudak (GPH) methods were used to estimate the values of fractional differences. Thus, some ARFIMA models were identified using AIC (Akaike Information Criterion), BIC (Schwartz Bayesian Information Criterion), AICc (corrected AIC), and the RMSE (Root Mean Squared Error). In result, the following conclusions were reached: the ARFIMA(2,0.3589648,2)-sGARCH(1,1) model and the ARFIMA(2,0.3589648,2)-fGARCH(1,1) model under normal distribution proved to be the best models, demonstrating the smallest values for these criteria. The calculations conducted herein show that the two models are of the same accuracy level in terms of the RMSE value, which equals 0.08808882, and it is this result that distinguishes our study. In conclusion, these models can be used to predict oil prices more accurately than others.

Key words: ARFIMA, volatility, fGARCH, sGARCH, modelling and forecasting, hybrid model.

[1] Corresponding author. School of Mathematical Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia and Department of Mathematics, Faculty of Science, Al-Hussein Bin Talal University, Ma'an, Jordan. E-mail: r.gounmeein@ahu.edu.jo. ORCID: https://orcid.org/0000-0002-2415-6937.

[2] School of Mathematical Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia. E-mail: m.tahir@usm.my. ORCID: https://orcid.org/0000-0003-2747-054X.

1. Introduction

Over the years, the study of oil price and volatility has remained one of the most important economic trends in terms of increasing investment and minimizing risk. Therefore, it is necessary to use an accurate statistical method to know the changes in price in terms of increase and decrease, through what is known as long memory (Mostafaei and Sakhabakhsh, 2012).

Long memory is a phenomenon we may sometimes face when analysing the time series data where long-term dependence between two points increases the amount of distance between them (Bahar et al., 2017). Usually when modelling long memory behaviour for any time series, such as mathematics, economics, among others; the operation can be more accurate by relying on the Autoregressive Fractionally Integrated Moving Average (ARFIMA) models compared with Autoregressive Integrated Moving Average (ARIMA) models. It can also have an important impact in the financial field (Bhardwaj and Swanson, 2006), where long memory models are one of the most important models used in the analysis of time series (Karia et al., 2013). ARFIMA model was fitted for the time series data either to better understand the data or to predict the future points in the series (forecasting). The use of forecasting in economic and financial fields is very important at the national, regional and international levels. It helps investors to reduce financial risks and increase the profits in the volatility of the global economy. The ARFIMA model was created by Granjer and Joyeux (1980) as mentioned by Mostafaei and Sakhabakhsh (2012) to capture the long memory behaviour of this time series data. The long memory feature exists if the autocorrelation function (ACF) decays more slowly than the exponential decay described by Bahar et al. (2017) or detected by using the statistical methods, namely Hurst Exponent as explained in Beran (1994). Besides, it is a known fact that long memory characteristics observed in data can be generated by a nonstationary structural break, as mentioned by Ohanissian et al. (2008). Therefore, the importance of testing for structural breaks in the conditional mean of a time series is necessary, as it determines that long memory is real or fake, as pointed out by Diebold and Inoue (2001), Granger and Hyung (2004) and Ohanissian et al. (2008). Therefore, the break detection procedure exhibits desirable properties both in the presence of breaks (stable potency across multiple breaks), as pointed out in Pretis et al. (2016). Besides, performing structural break testing when estimating the ARFIMA model is of great importance as it increases accuracy and prediction confidence.

On the other hand, volatility is an important consideration for any time series, especially in oil prices. Volatility is noticeable in studies related to financial, economic, tourism and other areas, where the data is widely scattered (Tendai and Chikobvu, 2017; Akter and Nobi, 2018). As it is known, there are obvious volatilities shown in some types of time series especially in crude oil prices (Lee and Huh, 2017). Therefore,

it was necessary to study these volatilities to avoid inaccuracies in the development of plans and strategies for making important decisions or for future predictions necessary. Moreover, to know their impact when forecasting to avoid any financial risks that may cause losses to the investor as the forecasting of financial time series data is yet as one of the most difficult tasks due to the non-stationary and non-linearity, as studied by Ismail and Awajan (2017). Also, Ramzan et al. (2012) showed that one category of models which has confirmed successful in forecasting volatility in many cases is the GARCH family of models. This is studied here by Standard Generalized Autoregressive Conditionally Heteroskedastic (sGARCH) model and Functional Generalized Autoregressive Conditionally Heteroskedastic (fGARCH) model. Based on the reason above we are choosing to study the long memory and volatility in this study, due to the modality of Brent crude oil prices grow exponentially, nonstationary and are volatile. These phenomena are popular features found in many large-scale data.

2. Literature review

For many years, many studies have been published that relate to the modelling and forecasting crude oil price (Yu et al., 2008; Aamir and Shabri, 2015; Sehgal and Pandey, 2015; Bahar et al., 2017; Lee and Huh, 2017; Yu et al., 2017; He, X. J., 2018; Yin et al., 2018). One of the old studies was by Yu et al. (2008), where they proposed using an empirical mode decomposition that depends on the learning model of the neural network group. The experimental results showed that the proposed model is a very capable approach to predicting international crude oil prices. Later, Aamir and Shabri (2015) used Auto-regressive Integrated Moving Average (ARIMA), Generalized Auto-regressive Conditional Heteroscedasticity (GARCH) and hybrid ARIMA-GARCH for modelling and forecasting monthly crude oil price of Pakistan. They found that the ARIMA-GARCH model is suitable and perform best based on the value of Akaike's Information Criterion (AIC) and Root Mean Squared Error (RMSE). Meanwhile, Sehgal and Pandey (2015) showed that Artificial Intelligent methods widely use forecasting oil prices as an alternative to traditional methods. Then, Lee and Huh (2017) suggested an alternative model which predicts the oil price using a Bayesian approach with informative priors. While Yu et al. (2017) found that Support Vector Machine (SVM) model outperformed Feed-Forward Neural Networks (FNN), Auto-Regressive Integrated Moving Average (ARIMA) model, Fractional Integrated ARIMA (ARFIMA) model, Markov-Switching ARFIMA (MS-ARFIMA) model, and Random Walk (RW) model for forecasting one-step or multi-step of crude oil price. In contrast, Bahar et al. (2017) used West Texas Intermediate daily data from 2/January/1986 to 31/August/2016, where the result showed that the price of crude oil has structural breaks feature. Moreover, the forecasting result showed high accuracy with geometric Brownian motion when compared with the mean-reverting Ornstein-Uhlenbeck

process for the short term. In 2018, He, X. J. identified the appropriate model for crude oil price prediction among several models used for weekly price data during the period 2009-2017. Machine learning Support Vector Regression (SVR) was found the best model. On the other hand, Yin et al. (2018) used numerous predictor variables with a new time-varying weight combination method where the results showed strong performance in forecasting the oil price.

Recently, many authors such as Fazelabdolabadi (2019) and Nyangarika et al. (2019) have still been interested in oil price prediction in terms of choosing the best predictive model. Nyangarika et al. (2019) used exponential smoothing to modify an Auto-Regressive Integrated Moving Average (ARIMA) model for the Brent crude oil price and Gas price data during the period from Jan/1991 to Dec/2016. In contrast, Fazelabdolabadi (2019) proposed the forecasting of the crude oil prices by applying a hybrid Bayesian Network (BN) method. The results showed that the proposed method is a good choice for short-term forecasting.

As mentioned previously, several models were used for modelling and forecasting the price of crude oil. The ARFIMA model is one of the famous models used in the analysis of time series (Karia et al., 2013) and understands the behaviour of the data, specifically crude oil prices. Jibrin et al. (2015) also used the ARFIMA model to study and forecast crude oil prices using weekly West Texas Intermediate and Brent series for the period 15/5/1987 to 20/12/2013, and explained that the WTI series and the Brent series have three breaks in the years 1999, 2004, 2008 and 1999, 2005, 2009, respectively. Bahar et al. (2017) used West Texas Intermediate daily data from 2/January/1986 to 31/August/2016, and the result showed that the price of crude oil had structural breaks feature. Also, there are previous studies that used volatility and hybrid models to describe the movement of crude oil prices. Hybrid models are an important method for studying the relationship between long memory and volatility. Among these studies, Manera et al. (2004) estimated the dynamic conditional correlations in the returns on Tapis oil spot and one-month forward prices for the period from 2 June 1992 to 16 January 2004, using CCCMGARCH (Constant Conditional Correlation Multivariate GARCH) model, VARMAGARCH (Vector Autoregressive Moving Average GARCH) model, VARMA-AGARCH (VARMA–Asymmetric GARCH) model, and DCC (Dynamic Conditional Correlation) model. The result shows that the ARCH and GARCH effects for spot (forward) returns are significant in the conditional volatility model for spot (forward) returns. Moreover, the multivariate asymmetric effects are significant for both spot and forward returns. Also, the calculated constant conditional correlations between the conditional volatilities of spot and forward return are virtually identical using CCC-GARCH(1,1), VAR(1)-GARCH(1,1) and VAR(1)-AGARCH(1,1). After that, in (2013) Kang and Yoon examined the volatility models and their forecasting abilities for three types of petroleum futures contracts traded on the New York Mercantile Exchange (West Texas Intermediate crude oil, heating oil #2,

and unleaded gasoline) particularly regarding volatility persistence (or long-memory properties). These models are ARIMA–GARCH, ARFIMA–GARCH, ARFIMA–IGARCH, and ARFIMA–FIGARCH. Although the ARFIMA–FIGARCH model better captures long-memory characteristics of returns and volatility, the out-of-sample analysis indicates that there is no single model for all three types of petroleum futures contracts, and this calls on investors to exercise caution when measuring and forecasting volatility (risk) in petroleum futures markets. As for Akron and Ismail (2017), they proposed a hybrid GA-FEEMD (Genetic Algorithm and Fast Ensemble Empirical Mode Decomposition) model for forecasting crude oil price time-series data. The results showed that the proposed hybrid model improved the forecasting accuracy of the data, compared with ARIMA and artificial neural network methods. On the other hand, Daniel Ambach and Oleksandra Ambach (2018) conducted a study on the application of a periodic regression model with the ARFIMA-GARCH residual process to model and predict the oil price, whereas the hybrid model provided some advantages, including that it captures long memory and conditional heteroscedasticity, but it failed to capture the periodicity in a good way. Besides, for the first lag of the squared standardized residuals, the proposed model showed a remaining presence of correlation, which is not satisfying at all. Therefore, it should be extended.

As a summary, previous studies have shown that there are mixed results in terms of selecting the appropriate model for the modelling and forecasting of crude oil prices. Thus, the current study focuses on constructing a time series model to forecast the monthly Brent crude oil price using ARFIMA with the GARCH family approach. Furthermore, due to the lack of studies in which crude oil prices have been predicted by comparing the ARFIMA-sGARCH hybrid model versus the ARFIMA-fGARCH hybrid model. Also, it will extend the works in the previous literature by examining long memory and volatilities in Brent crude oil prices simultaneously, by using the comparison of these models: ARFIMA-sGARCH model versus the ARFIMA-fGARCH model. Finally, this study also focuses on the interest in taking the two smallest values for accuracy criteria such as AIC and not just one value when choosing the best model. Thus, these points will be highlighted in this study. So, the purpose of this work is to identify structural breaks, verify that long memory is present for monthly Brent crude oil price data. Then, to determine the best model among ARFIMA models using some criteria and accuracy measures, such as AIC (Akaike information criterion) and RMSE (root mean squared error) in the sample. Besides, to check the residuals of the model for the existence of volatility or not, to determine the optimal model that can be used to study conditional variation (volatility) in series data through sGARCH and fGARCH models, to obtain the best hybrid model to predict the price of Brent crude oil in the short-term with the smallest error value.

3. Materials and Methods

3.1. The Dataset

Monthly data of the Brent crude oil price (all prices are per barrel in USA $) were used in this study from January 1979 to July 2019, obtained from the website: www.indexmundi.com/commodities/?commodity=crude-oil-brent. The data were divided into two parts: the first included data from January 1979 to July 2018 consisting of 475 observations, which were used to fit the forecasting model; while the second part has data from August 2018 to July 2019 consisting of 12 observations, which were used to test the accuracy of the in-sample forecast. Thereafter, a 13-month prediction was carried out outside the sample. This study uses the R-software version (3.5.3) to implement all statistical analyses.

3.2. Long Memory Test and Estimation

To check the presence of the long memory feature, there are several statistical methods that can be used, as described in Boutahar et al. (2007). These methods are R/S analysis, the aggregated variance method, and the Higuchi method. In particular, the range over standard deviation (R/S) analysis, which is a diagnostic of long memory, was the role played by Mandelbrot (1972), then Lo (1991) modified it. Mandelbrot (1972) found that the R/S analysis shows good properties over autocorrelation function (ACF) analysis and variance time function (VTF) analysis. After that, Lo (1991) modified R/S analysis, it is robust to short-range dependence, non-normal distributions, and conditional heteroscedasticity under the null hypothesis of no long-term dependence. This analysis achieves the following formula, as shown by Mandelbrot (1972) and Lo (1991):

$$Q_{(\mathrm{n})} = \frac{R_{(\mathrm{n})}}{S_{(n)}} = \frac{\max\limits_{1\le k\le n} \sum_{i=1}^{k}(X_i - \overline{X_n}) - \min\limits_{1\le k\le n} \sum_{i=1}^{k}(X_i - \overline{X_n})}{(n^{-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2)^{\frac{1}{2}}} \quad (1)$$

where $$\overline{X_n} = n^{-1}\sum_{i=1}^{n} X_i \quad (2)$$

and (n) is the sample size.

Besides, Jibrin et al. (2015) mentioned that a single structural break test is a test to determine the presence of break. It was introduced by Chow (1960) and modified to the Quandt Likelihood Ratio (QLR) test for the break between (t_0 and t_1) or called the Supremum F-statistic, given by:

$$Sup\ F = \max\{F(t_0), F(t_0+1), \dots, F(t_1)\} \quad (3)$$

where if Supremum F-statistic > 0.05, then the test rejects the null hypothesis which is H_0: no structural breaks.

On the other hand, the value of the fractional difference (d) was estimated by several methods, which were illustrated by (Hosking, 1981; Reisen, 1994; Boutahar et al., 2007; Palma, 2007; Telbany and Sous, 2016). These methods are:

The Hurst Exponent method: Reisen (1994) mentioned that this method, proposed by Hurst (1951, 1956) and then reviewed by McLeod and Hipel (1978), is based on the range $(R^*_{(n)})$ of the subtotals to deviate values from their mean in the time series divided by the standard deviation $(D^*_{(n)})$, which is denoted by $(R_{(n)})$ and written as follows:

$$R_{(\mathrm{n})} = \frac{R^*_{(\mathrm{n})}}{D^*_{(n)}} = \frac{\max\limits_{1\le k\le n} \sum_{i=1}^{k}(X_i-\bar{\mathrm{X}}) - \min\limits_{1\le k\le n} \sum_{i=1}^{k}(X_i-\bar{\mathrm{X}})}{(n^{-1}\sum_{i=1}^{n}(X_i-\bar{\mathrm{X}})^2)^{\frac{1}{2}}} \tag{4}$$

where
$$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i \tag{5}$$

The Geweke and Porter-Hudak (GPH) method: Based on the regression equation (Y_i), Geweke and Porter-Hudak (1983) suggested the estimation for the parameter $(\hat{d}_n)$, according to the following equations:

$$\hat{d}_n = -(\sum_{i=1}^{n}(X_i - \bar{X})^2)^{-1}(\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})) \tag{6}$$

where
$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{7}$$

$$\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i \tag{8}$$

In contrast, the smoothed periodogram (Sperio) and fractionally-differenced (Fracdiff) are just functions in R-software, used to estimate the value of the fractional difference (d), according to the following formulas respectively:

Reisen (1994) clarified the Sperio function, which estimates the fractional difference (d) in the ARFIMA(p,d,q) model. This function, represented by $f_s(w)$ and that is through the Parzen lag window, is as follows:

$$f_{\mathrm{s}}(w_j) = \frac{1}{2\pi}\sum_{-c}^{c} L\left(\frac{s}{c}\right) R(s)\cos(sw_j) \tag{9}$$

where
$$\mathrm{L}(u) = \begin{cases} 1 - 6u^2 + 6|u|^3 & , \quad |u| \le 1/2 \\ 2(1-|u|)^3 & , \quad -1/2 < u \le 1 \\ 0 & , \quad |u| > 1 \end{cases} \tag{10}$$

$\mathrm{L}(u)$ is called the Parzen lag window generator (we select the Parzen lag window as it has the feature that always yields positive estimates of the spectral density), (c) is the parameter (commonly indicated to the 'truncation point') and

$$R(\mathrm{s}) = \frac{1}{n}\left(\sum_{i=1}^{n-s}(X_i - \bar{X})(X_{i+s} - \bar{X})\right) \quad , \; s = 0, \pm 1, \ldots, \pm(n-1) \tag{11}$$

indicate the sample autocovariance function.

Hosking (1981) defined the fractionally-differenced operator, which uses the regression estimation method to estimate the fractional difference (d) for the ARFIMA

model (Olatayo and Adedotun, 2014). The (d) value is calculated by a binomial series, as follows:

$$\nabla^d = (1-B)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-B)^k = 1 - dB - \frac{1}{2}d(1-d)B^2 - \frac{1}{6}d(1-d)(2-d)B^3 - \cdots \quad (12)$$

3.3. Models Specification

The definitions of the ARIMA model were proposed by Box and Jenkins (Box et al., 2008) as follows. A stationary time series $\{x_t\}$ is called an Autoregressive Integrated Moving Average model of order (p,d,q) denoted by ($ARIMA(p,d,q)$), if

$$\phi_p(B)\nabla^d x_t = \theta_q(B)\epsilon_t \quad (13)$$

whereas,

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - .... - \phi_p B^p) \quad (14)$$

$$\theta_q(B) = (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q) \quad (15)$$

$$\nabla^d = (1-B)^d \quad (16)$$

where $\phi_p(B)$ is a polynomial of autoregressive for order (p) denoted by AR(p); $\theta_q(B)$ is a polynomial of moving average for order (q) denoted by MA(q). The integer number (d) is the non-seasonal difference order. (B) are the backward shift operators defined by $B^k X_t = X_{t-k}$. (∇) are the non-seasonal difference operators. Furthermore, (ϵ_t) is a white noise process.

ARFIMA model is the same as the ARIMA model above, but the essence of the difference between them is in the value of (d). If $d \in (0, 0.5)$, then the data has a long memory. While intermediate memory if $d \in (-0.5,0)$. However, when $d = 0$, then the data has a short memory (for mathematical details, see Beran (1994) page 60).

3.4. GARCH Models

In 1986, Bollerslev expanded the Autoregressive Conditional Heteroscedasticity (ARCH) model with order (q), which Engle developed in 1982, to become the Generalised Autoregressive Conditional Heteroscedasticity (GARCH) model with order (p, q) (Francq and Zakoian, 2019). The first model depends on uncorrelated random error values (ε_t). In contrast, the GARCH model relies on conditional variation. The general form of the $GARCH(p, q)$ model is given by Francq and Zakoian (2019) as follows:

$$\varepsilon_t = \eta_t \sigma_t, \qquad with \quad \eta_t \overset{iid}{\sim} N(0,1) \quad (17)$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \quad (18)$$

where $\omega > 0$, $\alpha_i \geq 0$ and $\beta_j \geq 0$ are $constants,$ $i = 1,2, \dots q,$ $j = 1,2, \dots p,$ and $t \in \mathbb{Z}$. Whereas when $\beta_j = 0,$ then equation (18) is called $ARCH(q)$. In contrast, if $p = q = 0$, then equation (18) becomes white noise. When the conditional variance of the process is unknown, the Asymptotic Quasi-likelihood (AQL) methodology is merging the kernel technique to estimate the parameter of the GARCH model, such as in Alzghool (2017).

3.5. Standard GARCH (sGARCH) Model

The conditional variance (σ_t^2) at a time (t) is expressed by the standard GARCH (p,q) model, the same as the equation (18), where (ε_t) is considered the residual returns, as the equation (17), which have been mentioned above (Miah and Rahman, 2016).

3.6. Functional GARCH (fGARCH) Model

Given the urgent need to describe the high-frequency volatilities that abound in the financial statements, a proper rational description of this problem, known as the function, had to be found (Francq and Zakoian, 2019). In 2013, Hörmann et al. suggested the functional approach of the ARCH model, then expanded this approach in 2017 by Aue et al., as mentioned by Francq and Zakoian (2019), through focusing on fGARCH(1,1) process such as in Aue et al. (2017), as shown below:

$$\sigma_i^2 = \delta + \alpha\varepsilon_{i-1}^2 + \beta\sigma_{i-1}^2 \tag{19}$$

where (ε_t) is a sequence of random functions satisfying the equation (17), $\delta \geq 0,$ $\alpha \geq 0$, $\beta \geq 0$ and $i \in \mathbb{Z}$.

Note that for $t \in [0,1]$ and (x) is an arbitrary element of the Hilbert space $\mathcal{H} = L^2[0,1]$, the integral operators (α) and (β) are defined by $(\alpha x)_{(t)} = \int_0^1 \alpha(t,s)x(s)ds$ and $(\beta x)_{(t)} = \int_0^1 \beta(t,s)x(s)ds$. The integral kernel functions $\alpha(t,s)$ and $\beta(t,s)$ are elements on $L^2[0,1]^2$.

As mentioned above, their approach depends on a daily division of the data (Francq and Zakoian, 2019), with the possibility for other time units (Aue et al., 2017), for example a monthly time unit, (see Aue et al., 2017; Francq and Zakoian, 2019 for more details).

3.7. Hybrid ARFIMA-GARCH Models

If the variance of the $ARFIMA(p,d,q)$ model can be modelled by a $GARCH(p,q)$ process, then this model is to be termed a hybrid $ARFIMA(p,d,q) - \mathrm{GARCH(p,q)}$. This model was defined by Palma (2007) as follows:

$$\phi_p(B)\nabla^d x_t = \theta_q(B)\varepsilon_t \tag{20}$$

where $\varepsilon_t = \eta_t \sigma_t \ , \qquad with \quad \eta_t \overset{iid}{\sim} N(0,1)$ (21)

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2 \quad (22)$$

According to the above, each model of the hybrid models will be estimated $ARFIMA(p,d,q) - sGARCH(p,q)$ and $ARFIMA(p,d,q) - fGARCH(p,q)$ under different distributions. Namely, normal (norm) distribution, student's t (std) distribution and generalized error (ged) distribution, see Tendai and Chikobvu (2017).

3.8. Criteria and Accuracy Measures for Choosing the Best Model

The fit model selection among several models is based on many criteria such as Akaike Information Criterion (AIC), Schwartz Bayesian Information Criterion (BIC) and corrected AIC (AICc) as indicated by Cryer and Chan (2008). They are given by the following formulas:

$$AIC = -2\ln(l) + 2k \quad (23)$$

$$BIC = -2\ln(l) + kln(n) \quad (24)$$

$$AICc = -2\ln(l) + 2k + \frac{2(k+1)(k+2)}{n-k-2} \quad (25)$$

where (l) is a maximum likelihood for the model, (k) is the total number of parameters $(meaning\ k = p + q)$ through the equations (14) and (15) respectively, while (n) is the number of observations. Therefore, the best model which gives the lowest value for these AIC, BIC and AICc criteria. Furthermore, the Root Mean Square Error (RMSE) is one of the accuracy measures which is used for evaluation of the performance of the model, as explained by Montgomery et al. (2015), as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2} \quad (26)$$

where (Y_t) is the actual value and $(\hat{Y}_t)$ is the forecasted value.

4. Results and Discussion

The monthly Brent crude oil price series is denoted by $\{X_t\}$ and (t) represents the time in months. Figure 1 displays the time series plot $\{X_t\}$ for the dataset from January 1979 to July 2019. Through the $\{X_t\}$ series, large fluctuations are observed over time, especially in 2008. The descriptive statistics of the monthly Brent crude oil price that consists of 487 observations have a mean of 42.95, a median of 30.20 and a positive skewness of 1.177466. For this reason, the tail of the series is on the left side.

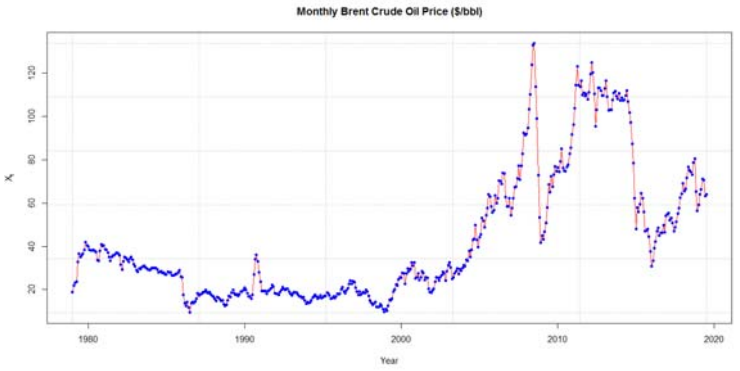


Figure 1. Time series plot for monthly Brent crude oil price ($/bbl)

Therefore, this series was studied in terms of having a long memory feature, through graphing and necessary statistical methods. The graph of the Autocorrelation Function (ACF) for time series data in Figure 2 shows a slow decrease.

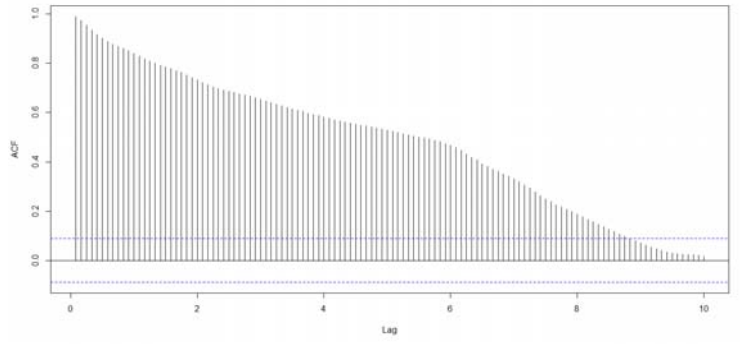


Figure 2. ACF plot for X_t

This gives a preliminary conclusion that there is a long memory, this is confirmed by Table 1 through several statistical methods.

Table 1. Long Memory Tests

R/S Analysis	Aggregated Variance Method	Higuchi Method
H = 0.8531864	H = 0.7910981	H = 0.9578515

The above table shows the results of three tests to check for the presence of long memory. It is noted that all values of (H) are greater than 0.5, which gives a firm conclusion to the existence of the long memory of the price for Brent crude oil data. Furthermore, the structural breaks are visible in the dataset series. Where, there exists four breaks for Brent are displayed in Figure 3 with the first, second, third, and last break captured in 1986, 1999, 2005, and 2013 respectively.

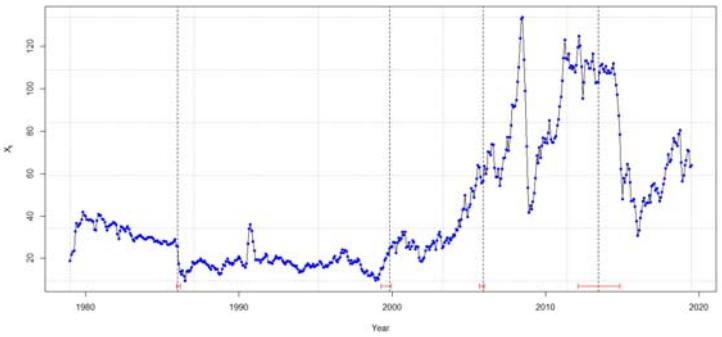


Figure 3. Monthly Brent crude oil price ($/bbl) with breaks and their confidence intervals

Besides, Table 2 displays the test result of the structural breaks using the QLR test. Note that the null hypothesis for the structural breaks is rejected as the Sup-F statistic is too large (1190) and the P-values < 0.05.

Table 2. Structural Break Test

QLR	P-value
1190	< 2.2e-16

Based on the P-value for the Jarque Bera test (< 2.2e-16) and the coefficient of skewness mentioned earlier, this series is considered not normal. So, the $\{x_t\}$ transformation must be done. Assume that $\{Y_t\}$ represents the growth rate for $\{X_t\}$ as in the following formula:

$$Y_t = \log(X_t) \tag{27}$$

Figure 4 displays the growth rate time series plot $\{Y_t\}$ for the series $\{X_t\}$.

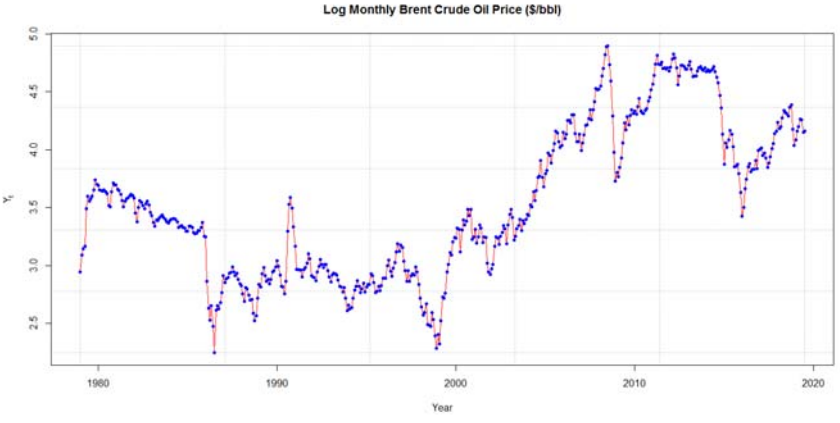


Figure 4. Time series plot for Y_t

In contrast, the Autocorrelation Function (ACF) and Partial Autocorrelation Functions (PACF) for $\{Y_t\}$ are given in Figure 5. It shows that the series is not white noise.

Based on the above results, the fractional difference (d) values for the $\{ Y_t \}$ series will be estimated in several different methods and functions, as shown in Table 3, where the value of the fractional difference using the Hurst Exponent method is 0.3589648, the value by the Sperio function estimate is 0.4984955, and the value of 0.4994726 was the result of Fractionally-Differenced function estimate. In contrast, the value of 0.7676326, which is due to Geweke and Porter-Hudak method estimate, is excluded because it is greater than 0.5.

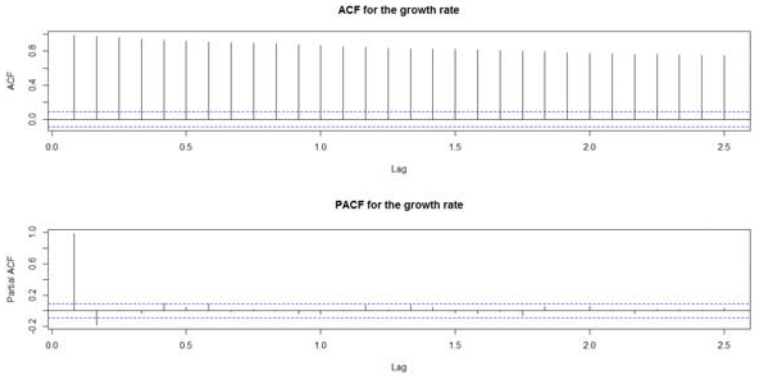


Figure 5. ACF and PACF plot for Y_t

Table 3. Fractional Difference Values for Y_t Series

Method / Function	d	State
Hurst Exponent (d = H-0.5)	d_1 = 0.3589648	$0 < d_1 < 0.5$
Sperio (bandw.exp = 0.3, beta = 0.74)	d_2 = 0.4984955	$0 < d_2 < 0.5$
Fractionally-Differenced (Fracdiff)	d_3 = 0.4994726	$0 < d_3 < 0.5$
Geweke and Porter-Hudak (GPH)	d_4 = 0.7676326	$0.5 < d_4$

Tables (4-6) and Figure 6 respectively, illustrate the stationary test using the Augmented Dickey-Fuller (ADF) test and Phillips-Perron (PP) test. Note that the $\{ Y_t \}$ series is stationary after taking the fractional difference (d) based on the different methods and functions shown in tables. Whereas, the fractional difference for the series $\{Y_t\}$ will be treated according to equation (12) as follows:

$$Z_{t(di)} = diff(Y_t) = Y_t \nabla^{di} \tag{28}$$

where, $d_i = d_1, d_2 \ and \ d_3$, respectively.

Table 4. The Stationary Test for $Z_{t(d1)}$ Series Using the Hurst Exponent Method

Method	Test	Value	p-value	State
Hurst (d_1 = 0.3589648)	ADF Test	- 4.1727	0.01	Stationary
	PP Test	- 82.923	0.01	Stationary

Table 5. The Stationary Test for $Z_{t(d2)}$ Series Using Sperio Function

Function	Test	Value	p-value	State
Sperio (d_2 = 0.4984955)	ADF Test	- 5.1927	0.01	Stationary
	PP Test	- 151.34	0.01	Stationary

Table 6. The Stationary Test for $Z_{t(d3)}$ Series Using Fractionally-Differenced Function

Function	Test	Value	p-value	State
Fractionally-Differenced (Fracdiff) (d_3 = 0.4994726)	ADF Test	- 5.2001	0.01	Stationary
	PP Test	- 151.89	0.01	Stationary

According to equations 23-26 above, a qualifying model is one that has the lowest value for AIC, AICc, BIC and RMSE. As a result of Table 7, ARFIMA(1,0.3589648,0) model, ARFIMA(2,0.3589648,1) model and ARFIMA(2,0.3589648,2) model have the lowest values for these criteria. Also, it is noted that these models are within the Hurst Exponent estimate, which has the lowest value for the fractional difference estimate (d). As a result, the three models will be taken and compared to choose the best among them by moving to the next step of testing the residuals (see Al-Gounmeein and Ismail, 2020). While in this step, residuals testing is a necessary step to examine any model through several methods, including the graph for the ACF and the P-value for the Ljung-Box residuals test, because these methods are important measures to consider correlations of residuals (Montgomery et al., 2015).

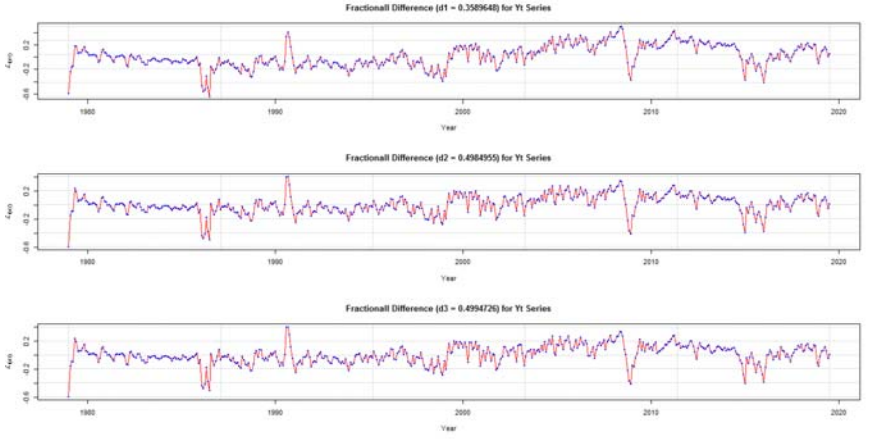


Figure 6. Time series plot for $Z_{t(di)}$ using the fractional difference values, respectively

By looking at Table 8, the three models do not have the property of the unit root for the residuals, using the P-value for the Ljung-Box test statistics at Lag(12), Lag(24) and Lag(36). We note that the P-value for the residuals of the third model is larger than the first and the second. In contrast, that model has the smallest Chi-Square statistic (χ^2) at the same different Lags. This is one of the indicators that gives the conclusion that the ARFIMA$(2, 0.3589648\ ,2)$ model is the best. Furthermore, it is observed that

through Figures 7-9, where all residuals of these models are very small by looking at the ACF plot and all the Ljung-Box P-values lie above the dashed line.

Table 7. AIC, AICc, BIC, and RMSE of ARFIMA Models

d	Model	AIC	AICc	BIC	RMSE
Hurst $d_1 = 0.3589648$	**(1, d_1 ,0)**	- 962.91	- 962.89	**- 954.59**	0.08730141
	(0, d_1 ,1)	- 639.85	- 639.83	- 631.52	0.1227511
	(1, d_1 ,1)	- 961.07	- 961.02	- 948.58	0.08728667
	(2, d_1 ,0)	- 961.04	- 960.99	- 948.55	0.08728949
	(0, d_1 ,2)	- 769.74	- 769.69	- 757.25	0.1068218
	(2, d_1 ,1)	**- 966.25**	**- 966.16**	- 949.6	0.08661429
	(1, d_1 ,2)	- 962.7	- 962.62	- 946.05	0.08695112
	(2, d_1 ,2)	**- 966.07**	- 965.95	- 945.26	**0.08644647**
Sperio $d_2 = 0.4984955$	(1, d_2 ,0)	- 956.67	- 956.65	- 948.35	0.08793743
	(0, d_2 ,1)	- 818.6	- 818.57	- 810.27	0.1017299
	(1, d_2 ,1)	- 954.8	- 954.75	- 942.31	0.08792554
	(2, d_2 ,0)	- 954.77	- 954.72	- 942.28	0.08792809
	(0, d_2 ,2)	- 876.59	- 876.54	- 864.1	0.09549913
	(2, d_2 ,1)	- 959.16	- 959.08	- 942.51	0.08732158
	(1, d_2 ,2)	- 955.83	- 955.75	- 939.18	0.08764342
	(2, d_2 ,2)	- 957.37	- 957.24	- 936.55	0.0873032
Fractionally – Differenced (Fracdiff) $d_3 = 0.4994726$	(1, d_3 ,0)	- 956.64	- 956.61	- 948.31	0.08794099
	(0, d_3 ,1)	- 819.47	- 819.44	- 811.14	0.1016366
	(1, d_3 ,1)	- 954.77	- 954.72	- 942.28	0.08792893
	(2, d_3 ,0)	- 954.74	- 954.69	- 942.25	0.08793152
	(0, d_3 ,2)	- 877.09	- 877.04	- 864.6	0.09544912
	(2, d_3 ,1)	- 959.11	- 959.02	- 942.45	0.0873277
	(1, d_3 ,2)	- 955.79	- 955.71	- 939.14	0.08764758
	(2, d_3 ,2)	- 957.31	- 957.18	- 936.49	0.08730851

Table 8. Ljung-Box Test Statistic for the Residuals

Model	Lag (12)		Lag (24)		Lag (36)	
	χ^2	P-value	χ^2	P-value	χ^2	P-value
ARFIMA(1,0.3589648,0)	23.247	0.0257	39.895	0.02195	51.031	0.04968
ARFIMA(2,0.3589648,1)	19.037	0.08763	35.148	0.06623	44.401	0.1588
ARFIMA(2,0.3589648,2)	17.594	0.1286	32.708	0.1104	41.946	0.2287

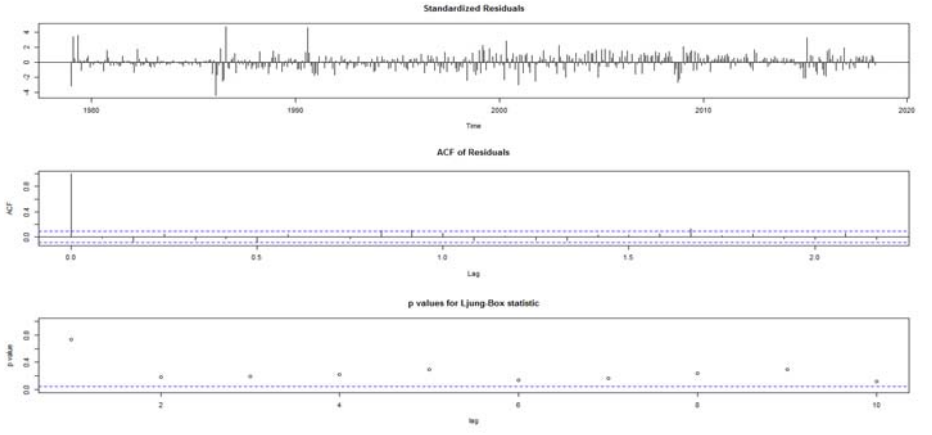


Figure 7. Plots of the residuals for the ARFIMA$(1, d_1, 0)$ model

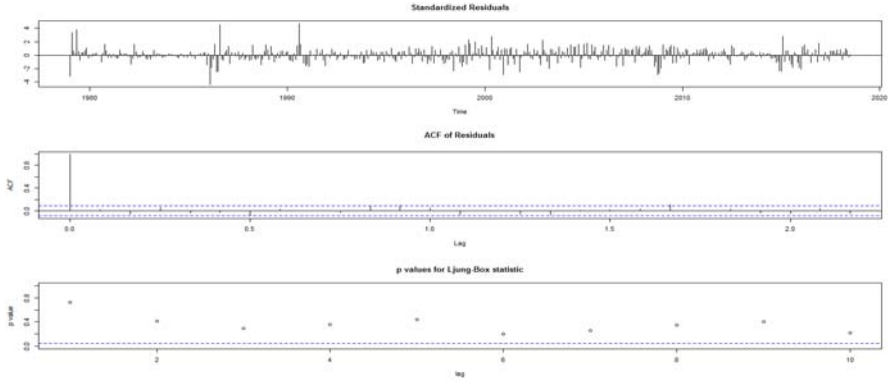


Figure 8. Plots of the residuals for the ARFIMA$(2, d_1, 1)$ model

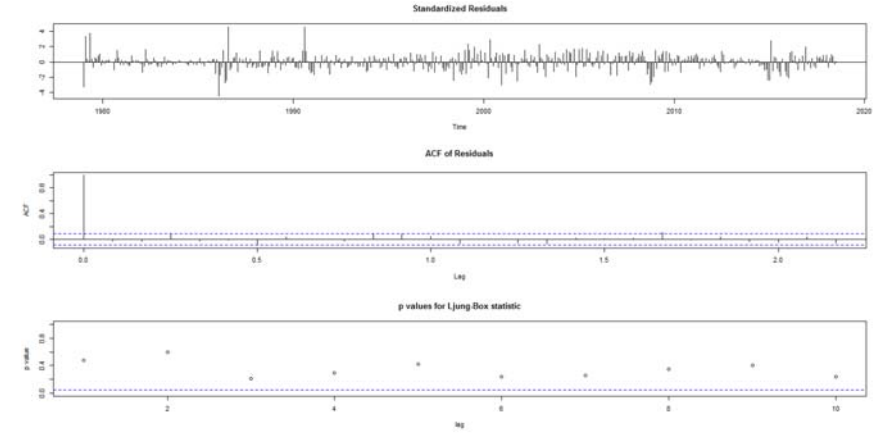


Figure 9. Plots of the residuals for the ARFIMA$(2, d_1, 2)$ model

While that the P-value for the Jarque Bera test for the residual's models ARFIMA$(1, d_1, 0)$, ARFIMA$(2, d_1, 1)$ and ARFIMA$(2, d_1, 2)$ is $< 2.2e^{-16}$. As for the P-value for the Shapiro-Wilk normality test for these models' residuals are $1.585e^{-8}$, $1.919e^{-9}$, and $1.686e^{-9}$, respectively. As a result, these models' residuals are not normally distributed.

On the other hand, when examining the residuals, it was observed that there exists Heteroscedasticity and ARCH effect using the ARCH Lagrange Multiplies (ARCH-LM) test for the residuals and residuals squared as shown in the P-value for the three models of ARFIMA in Table 9. Where all P-values for residuals and residuals squared in the ARCH-LM test less than 0.05. Therefore, we reject H_0. This means there exists an ARCH effect. As well as when returning to Table 8 using the Ljung-Box test statistic. Certainly, illustrated by showing the Heteroskedasticity for the squared residuals in Figures 10-12 for these models, respectively.

Table 9. ARCH-LM Test for the Residuals and Residuals Squared

Model	P-value	
	$residuals$	$(residuals)^2$
ARFIMA (1, 0.3589648 ,0)	1.275e-06	2.338e-06
ARFIMA (2, 0.3589648 ,1)	3.92e-07	0.000185
ARFIMA (2, 0.3589648 ,2)	8.217e-07	9.66e-06

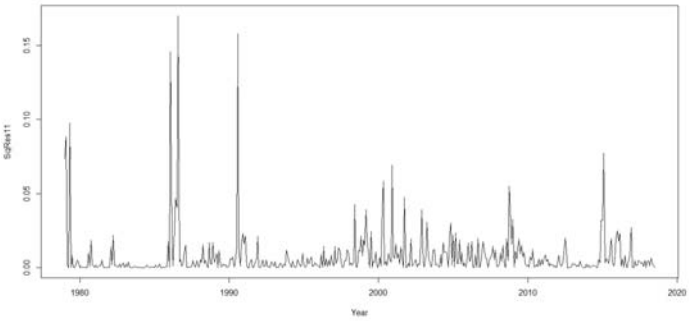


Figure 10. Plots of squared residuals for ARFIMA$(1, d_1, 0)$ model

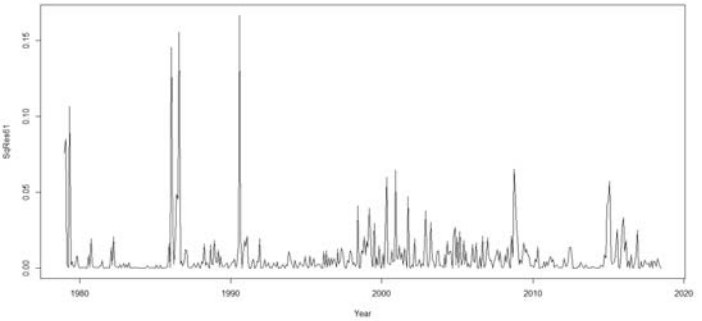


Figure 11. Plots of squared residuals for ARFIMA$(2, d_1, 1)$ model

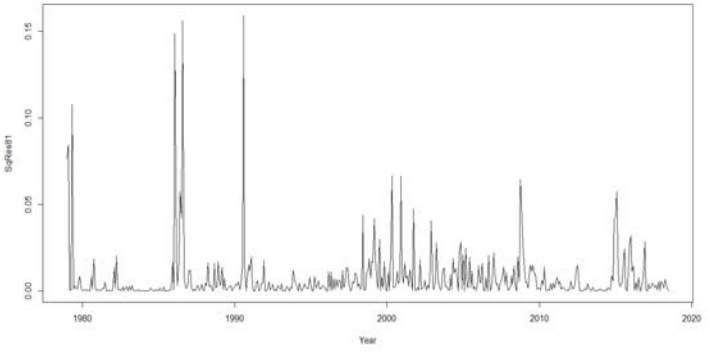


Figure 12. Plots of squared residuals for ARFIMA$(2, d_1, 2)$ model

We note from these figures that the volatility of the squared residuals changes significantly throughout the year in the three models. Therefore, the presence of these high volatility leads to the handling of GARCH family models. In other words, in the next steps, we will choose the appropriate GARCH model using the AIC criterion to get the ARFIMA-GARCH hybridization model. Thus, the AIC criterion will be used to choose the best model for the study of the volatility. This is illustrated by the following Figures 13-15 of the residuals of these models, which clarifies dealing with the type of Standard GARCH (sGARCH) model.

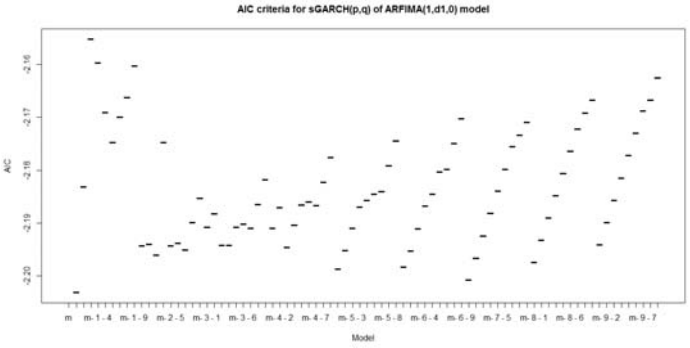


Figure 13. AIC criteria for sGARCH(p , q) of ARFIMA$(1, d_1, 0)$ model

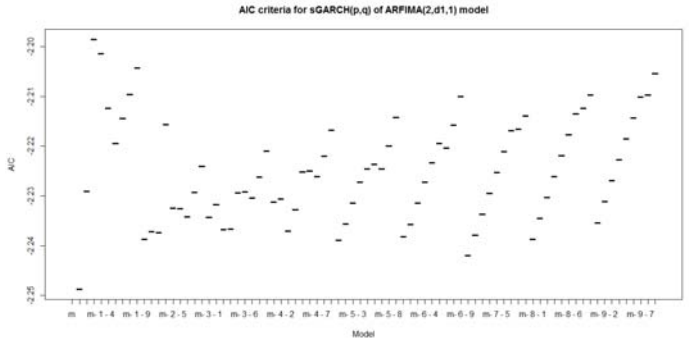


Figure 14. AIC criteria for sGARCH(p , q) of ARFIMA$(2, d_1, 1)$ model

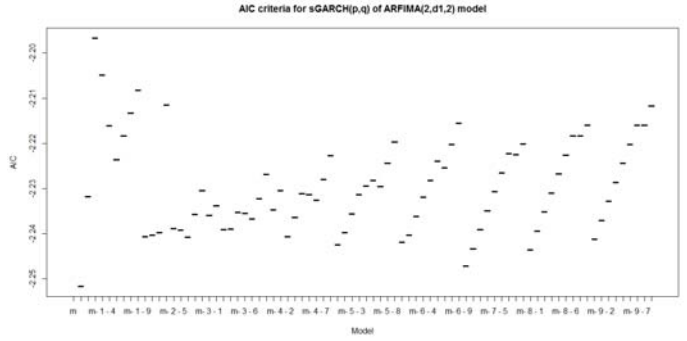


Figure 15. AIC criteria for sGARCH(p , q) of ARFIMA$(2, d_1, 2)$ model

The graphical test of AIC criteria above (Figures 13-15) indicates that the best volatility model for these three models is sGARCH(1,1). Also, based on the smallest value of this criterion for sGARCH(1,1) of ARFIMA$(1, d_1, 0)$ model, ARFIMA$(2, d_1, 1)$ model and ARFIMA$(2, d_1, 2)$ model is (-2.203081, -2.248740 and -2.251659), respectively.

Table 10 presents the RMSE result of nine hybrids (ARFIMA-sGARCH) models, implemented in three distributions (normal distribution, student's t distribution and generalized error distribution). Accordingly, the ARFIMA(2,0.3589648,2)-sGARCH(1,1) model under normal distribution is the best in modelling and forecasting Brent crude oil price volatility, as this model has the smallest value for RMSE.

Table 10. RMSE of ARFIMA-sGARCH Models

d	Model	RMSE
Hurst $d_1 = 0.3589648$	ARFIMA(1, d_1 ,0)-sGARCH (1,1) norm	0.09680273
	ARFIMA(1, d_1 ,0)-sGARCH (1,1) std	0.09122890
	ARFIMA(1, d_1 ,0)-sGARCH (1,1) ged	0.09063099
	**********	
	ARFIMA(2, d_1 ,1)-sGARCH (1,1) norm	0.08934022
	ARFIMA(2, d_1 ,1)-sGARCH (1,1) std	0.08964552
	ARFIMA(2, d_1 ,1)-sGARCH (1,1) ged	0.09002362
	**********	
	ARFIMA(2, d_1 ,2)-sGARCH (1,1) norm	**0.08808882**
	ARFIMA(2, d_1 ,2)-sGARCH (1,1) std	0.08974343
	ARFIMA(2, d_1 ,2)-sGARCH (1,1) ged	0.08871313

Hence, from the result in Table 10, the optimal parameters of this model summarised in Table 11, where all the estimated coefficients of the ARFIMA(2,0.3589648,2)-sGARCH(1,1) model in Table 11 have statistical significance at the 5% level according to the normal distribution of the model, where the values for (α_1 and β_1) indicate that the conditional variance is positive.

Table 11. ARFIMA(2, 0.3589648 ,2) − $sGARCH(1,1)$ Parameters

Parameters	Estimate	Standard Error	Prob.
mu	- 0.451487	0.025846	0.000000
ar(1)	1.720112	0.000499	0.000000
ar(2)	- 0.720268	0.000340	0.000000
ma(1)	- 0.853978	0.019381	0.000000
ma(2)	- 0.097814	0.004752	0.000000
omega	0.000531	0.000190	0.005154
α_1	0.419994	0.080176	0.000000
β_1	0.579006	0.057405	0.000000

On the other hand, volatility in the price of Brent crude oil has been studied through the fGARCH model. It is clear from Table 12 that the normal distribution of the ARFIMA(2,0.3589648,2)-fGARCH(1,1) model with -2.2515 AIC criterion has the

smallest RMSE. Also, the same hybridization model was obtained in Table 10 with a different type of GARCH family. This is shown in Table 12.

By looking at Table 13, which shows the optimal parameters for this model, it has the same statistical significance as the model of ARFIMA(2,0.3589648,2)-sGARCH(1,1) in Table 11. In other words, the significance of alpha and beta values in the two models indicates that the price's volatilities in the past period affect the current price's volatilities.

As a result of this study, it was found that ARFIMA(2,0.3589648,2)-sGARCH(1,1) model and ARFIMA(2,0.3589648,2)-fGARCH(1,1) model under normal distribution are equal in the value of RMSE. Thus, these two models will be taken and moved to the next step, the model validation phase preceding the prediction phase.

Table 12. RMSE of ARFIMA-fGARCH Models

d	Model	RMSE
Hurst $d_1 = 0.3589648$	ARFIMA(1, d_1 ,0)-fGARCH (1,1) norm	0.09681124
	ARFIMA(1, d_1 ,0)-fGARCH (1,1) std	0.09122962
	ARFIMA(1, d_1 ,0)-fGARCH (1,1) ged	0.09063172
	**********	
	ARFIMA(2, d_1 ,1)- fGARCH (1,1) norm	0.08934022
	ARFIMA(2, d_1 ,1)- fGARCH (1,1) std	0.08964552
	ARFIMA(2, d_1 ,1)- fGARCH (1,1) ged	0.09002547
	**********	
	ARFIMA(2, d_1 ,2)-fGARCH (1,1) norm	**0.08808882**
	ARFIMA(2, d_1 ,2)- fGARCH (1,1) std	0.08974902
	ARFIMA(2, d_1 ,2)- fGARCH (1,1) ged	0.08871313

Table 13. ARFIMA$(2,\ 0.3589648\ ,2) - fGARCH(1,1)$ Parameters

Parameters	Estimate	Standard Error	Prob.
mu	- 0.451487	0.025847	0.000000
ar(1)	1.720112	0.000499	0.000000
ar(2)	- 0.720268	0.000340	0.000000
ma(1)	- 0.853977	0.019381	0.000000
ma(2)	- 0.097815	0.004752	0.000000
omega	0.000531	0.000190	0.005154
α_1	0.419992	0.080176	0.000000
β_1	0.579007	0.057405	0.000000

Based on the above outcomes of the identification, estimation and diagnosis stages, the final validation of the two hybridization models is necessary by testing the residuals. By using the P-value for the Ljung-Box statistical test, we note that the P-value for the residuals of two hybridization models equals 0.993 > 0.05. It means that these two models have the property of the unit root or independent residuals. On the other hand, this can be confirmed by the figures for ACF of standardized residuals and ACF of squared standardized residuals shown in Figures 16-17 respectively (see Iqelan, 2015).

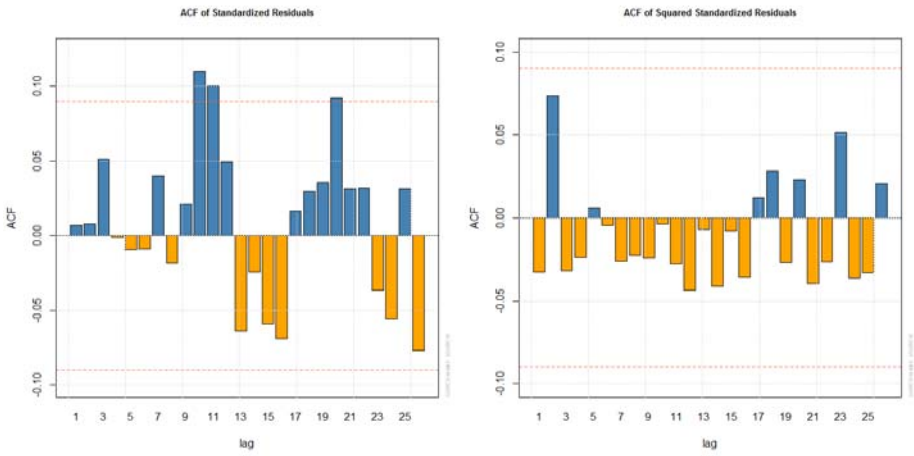


Figure 16. ACF of standardized residuals and squared standardized residuals for ARFIMA(2, 0.3589648 ,2)-sGARCH(1,1) model

Thus, the result of this study is that one or both models can be used to modelling and forecasting Brent crude oil price volatility in the short-term. Due to its accuracy in the performance with the least predictive error, the forecast out-of-sample for the best two models presented (Table 14) is from August 2019 to August 2020, where the table shows that the forecast of conditional variance for these models is increasing slowly over the future period. In other words, the volatility values in the same table are increasing at a slow rate. This indicates uncertainty in knowing the future monthly price of Brent crude oil. This is confirmed by the apparent decline in monthly price - the series column - which will affect the future growth of the global economy and the price of the dollar. Consequently, it will affect global oil production.

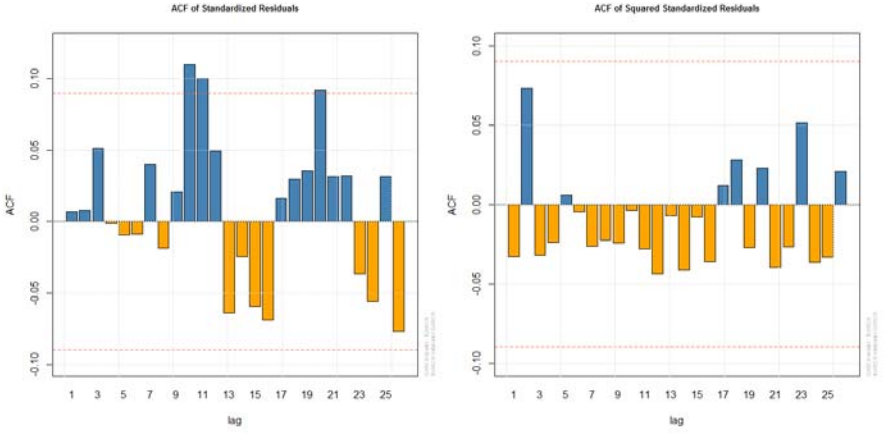


Figure 17. ACF of standardized residuals and squared standardized residuals for ARFIMA(2, 0.3589648 ,2)-fGARCH(1,1) model

Table 14. Forecast Out-of-Sample for ARFIMA(2, 0.3589648 ,2) − $sGARCH(1,1)$ Model and ARFIMA(2, 0.3589648 ,2) − $fGARCH(1,1)$ Model

		ARFIMA$(2, d_1, 2) - sGARCH(1,1)$		ARFIMA$(2, d_1, 2) - fGARCH(1,1)$	
Year	Month	Series	Sigma	Series	Sigma
2019	Aug	64.436	4.4553	64.436	4.4553
	Sep	63.810	4.4756	63.810	4.4756
	Oct	63.311	4.4959	63.311	4.4959
	Nov	62.709	4.5160	62.709	4.5160
	Dec	62.214	4.5360	62.214	4.5360
2020	Jan	61.634	4.5559	61.634	4.5559
	Feb	61.144	4.5758	61.144	4.5757
	Mar	60.584	4.5955	60.584	4.5954
	Apr	60.100	4.6150	60.100	4.6150
	May	59.558	4.6345	59.559	4.6345
	Jun	59.081	4.6539	59.081	4.6539
	Jul	58.557	4.6732	58.557	4.6732
	Aug	58.088	4.6924	58.088	4.6924

The result of this study calls for the development of the best strategic plans and vision for the future by economists, investors, and analysts to take advantage of the uncertainty in Brent crude oil prices in the future. Also, it is possible to conduct similar studies on Brent crude oil price when the case study for the fractional difference value is greater than 0.5 for ARFIMA models or the study of ARFIMA models in terms of seasonal presence.

5. Conclusion

This paper is designed to determine the modelling and forecasting of monthly Brent crude oil price and its volatility. Also, it extended the works from the previous literature by examining long memory and volatilities in the dataset simultaneously, by using the comparison of the ARFIMA-sGARCH models versus the ARFIMA-fGARCH models. It was noted that the ARFIMA(2,0.3589648,2)-sGARCH(1,1) model and ARFIMA(2,0.3589648,2)-fGARCH(1,1) model under normal distribution with RMSE, which equals (0.08808882) are the best for these data, where these models outperform other several models in modelling and forecasting the volatility. The forecasts for these models indicated a decline in the price in the short-term. On the other hand, the Hurst Exponent method outperformed constructing an appropriate hybridization model to predict. Finally, we obtained distinct results for our study that distinguish it from other previous studies, namely: two-hybrid models of long memory phenomenon (ARFIMA) were obtained with two members of the GARCH family (sGARCH and fGARCH) having the same accuracy in RMSE value. Also, the best model does not have the smallest AIC value, which gives the conclusion that taking a single value for the AIC

criterion is not sufficient to choose the best model among the models. Therefore, it is proposed to consider taking the two smallest values in accuracy criteria such as AIC and not just the smallest value, and this is what this study showed.

References

AAMIR, M., SHABRI, A. B., (2015). Modelling and Forecasting Monthly Crude Oil Prices of Pakistan: A Comparative Study of ARIMA, GARCH and ARIMA-GARCH Models. Sci.Int. (Lahore), 27(3), pp. 2365–2371.

AKRON, N., ISMAIL, Z., (2017). A hybrid GA-FEEMD for forecasting crude oil prices. Indian Journal of Science and Technology, 10(31), pp. 1–6.

AKTER, N., NOBI, A., (2018). Investigation of the Financial Stability of S&P 500 Using Realized Volatility and Stock Returns Distribution. Journal of Risk Financial Management, 11(22), pp. 1–10.

AL-GOUNMEEIN, R. S., ISMAIL, M. T., (2020). Forecasting the Exchange Rate of the Jordanian Dinar versus the US Dollar Using a Box-Jenkins Seasonal ARIMA Model. International Journal of Mathematics and Computer Science, 15(1), pp. 27–40.

AMBACH, D., AMBACH, O., (2018). Forecasting the oil price with a periodic regression ARFIMA-GARCH process. IEEE Second International Conference on Data Stream Mining & Processing, Lviv, Ukraine, pp. 212–217.

ALZGHOOL, R., (2017). Parameters estimation for GARCH (p,q) model: QL and AQL approaches. Electronic Journal of Applied Statistical Analysis, 10(1), pp.180–193.

AUE, A., HORVATH, L. and PELLATT, D. F., (2017). Functional generalized autoregressive conditional heteroskedasticity. Journal of Time Series Analysis, 38(1), pp. 3–21.

BAHAR, A., NOH, N. M. and ZAINUDDIN, Z. M., (2017). Forecasting model for crude oil price with structural break. Malaysian Journal of Fundamental and Applied Sciences, pp. 421–424.

BERAN, J., (1994). Statistics for Long Memory Processes, Chapman and Hall, p. 315.

BHARDWAJ, G., SWANSON, N. R., (2006). An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series. Journal of Econometrics 131, pp. 539–578.

BOUTAHAR, M., MARIMOUTOU, V. and NOUIRA, L., (2007). Estimation Methods of the Long Memory Parameter: Monte Carlo Analysis and Application. Journal of Applied Statistics, 34(3), pp. 261–301.

BOX, G. E. P., JENKINS, G. M. and REINSEL, G. C., (2008). Time series analysis forecasting and control, Fourth Edition, Wiley & Sons, Inc, p. 746.

CRYER, J. D., CHAN, K., (2008). Time Series Analysis With Application in R, Second Edition, Springer, p. 491.

DIEBOLD, F. X., INOUE, A., (2001). Long Memory and Regime Switching. Journal of Econometrics, 105, pp. 131–159.

FAZELABDOLABADI, B., (2019). A hybrid Bayesian-network proposition for forecasting the crude oil price. Financial Innovation, 5(30), pp. 1–21.

FRANCQ, C., ZAKOIAN, J. M., (2019). GARCH Models: Structure, Statistical Inference and Financial Applications, Second Edition, John Wiley & Sons Ltd, p. 487.

GRANGER, C. W. J., HYUNG, N., (2004). Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. Journal of Empirical Finance, 11, pp. 399–421.

HE, X. J., (2018). Crude Oil Prices Forecasting: Time Series vs. SVR Models. Journal of International Technology and Information Management, 27(2), pp. 25–42.

HOSKING, J. R. M., (1981). Fractional differencing. Biometrika, 86(1), pp. 165–176.

IQELAN, B. M., (2015). Time Series Modeling of Monthly Temperature Data of Jerusalem / Palestine. MATEMATIKA, 31(2), pp. 159–176.

ISMAIL, M. T., AWAJAN, A. M., (2017). A new hybrid approach EMD-EXP for short-term forecasting of daily stock market time series data. Electronic Journal of Applied Statistical Analysis, 10(2), pp. 307–327.

JIBRIN, S. A., MUSA, Y., ZUBAIR, U. A. and SAIDU, A. S., (2015). ARFIMA Modelling and Investigation of Structural Break(s) in West Texas Intermediate and Brent Series, CBN Journal of Applied Statistics, 6(2), pp. 59–79.

KANG, S. H., YOON, S., (2013). Modeling and forecasting the volatility of petroleum futures prices. Energy Economics, 36, pp. 354–362.

KARIA, A. A., BUJANG, I. and AHMAD, I., (2013). Fractionally integrated ARMA for crude palm oil prices prediction: case of potentially over difference. Journal of Applied Statistics, 40(12), pp. 2735–2748.

LEE, C. Y., HUH, S. Y., (2017). Forecasting Long-Term Crude Oil Prices Using a Bayesian Model with Informative Priors. Sustainability. 9, 190, DOI: 10.3390/su9020190.

LO, A. W., (1991). Long-term memory in stock market prices. Econometrica, 59(5), pp. 1279–1313.

MANDELBROT, B., (1972). Statistical Methodology for Nonperiodic Cycles: From the Covariance to R/S Analysis. Annals of Economic and Social Measurement, 1(3), pp. 259–290.

MANERA, M., MCALEER, M. and GRASSO, M., (2004). Modelling dynamic conditional correlations in the volatility of spot and forward oil price returns, 2nd International Congress on Environmental Modelling and Software - Osnabrück, Germany, 183, pp. 1–6.

MIAH, M., RAHMAN, A., (2016). Modelling Volatility of Daily Stock Returns: Is GARCH(1,1) Enough?. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 18(1), pp. 29–39.

MONTGOMERY, D. C., JENNINGS, C. L. and KULAHCI, M., (2015). Introduction To Time Series Analysis And Forecasting, Second Edition, Wiley & Sons, Inc, p. 643.

MOSTAFAEI, H., SAKHABAKHSH, L., (2012). Using SARFIMA Model to Study and Predict the Iran's Oil Supply. International Journal of Energy Economics and Policy, 2(1), pp. 41–49.

NYANGARIKA, A., MIKHAYLOV, A. and RICHTER, U. H., (2019). Oil Price Factors: Forecasting on the Base of Modified Auto-regressive Integrated Moving Average Model. International Journal of Energy Economics and Policy, 9(1), pp. 149–159.

OHANISSIAN, A., RUSSELL, J. R. and TSAY, R. S., (2008). True or Spurious Long Memory? A New Test. Journal of Business & Economic Statistics, 26(2), pp. 161–175.

OLATAYO, T. O., ADEDOTUN, A. F., (2014). On the Test and Estimation of Fractional Parameter in ARFIMA Model: Bootstrap Approach. Applied Mathematical Sciences, 8(96), pp.4783–4792.

PALMA, W., (2007). Long-Memory Time Series: Theory and Methods, John Wiley & Sons, Inc, p. 285.

PRETIS, F., SCHNEIDER, L., SMERDON, J. E. and HENDRY, D. F., (2016). Detecting volcanic eruptions in temperature reconstructions by designed break-indicator saturation. Journal of Economic Surveys, 30(3), pp. 403–429.

RAMZAN, S., RAMZAN, S. and ZAHID, F. M., (2012). Modeling and Forecasting Exchange Rate Dynamics In Pakistan Using ARCH Family of Models. Electronic Journal of Applied Statistical Analysis, 5(1), pp. 15–29.

REISEN, V. A., (1994). Estimation of the Fractional Difference Parameter in the ARIMA(p,d,q) Model Using the Smoothed Periodogram. Journal of Time Series Analysis, 15(3), pp. 335–350.

SEHGAL, N., PANDEY, K. K., (2015). Artificial intelligence methods for oil price forecasting: a review and evaluation, Springer-Verlag Berlin Heidelberg, DOI: 10.1007/s12667-015-0151-y.

TELBANY, S., SOUS, M., (2016). Using ARFIMA Models in Forecasting Indicator of the Food and Agriculture Organization. IUGJEBS, 24(1), pp. 168–187.

TENDAI, M., CHIKOBVU, D., (2017). Modelling international tourist arrivals and volatility to the Victoria Falls Rainforest, Zimbabwe: Application of the GARCH family of models. African Journal of Hospitality, Tourism and Leisure, 6(4), pp. 1–16.

YIN, X., PENG, J. and TANG, T., (2018). Improving the Forecasting Accuracy of Crude Oil Prices. Sustainability. 10, 454, DOI: 10.3390/su10020454.

YU, L., WANG, S. and LAI, K. K., (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Economics, 30, pp. 2623–2635.

YU, L., ZHANG, X. and WANG, S., (2017). Assessing Potentiality of Support Vector Machine Method in Crude Oil Price Forecasting. EURASIA Journal of Mathematics, Science and Technology Education, 13(12), pp. 7893-7904.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1, pp. 55–73, DOI 10.21307/stattrans-2021-003
Received – 28.03.2020; accepted – 22.12.2020

Measurement of enterprise mobility among size classes, taking into account business demography

Camilla Ferretti[1]

ABSTRACT

A extensive body of literature is devoted to the production of mobility measures based on transition matrices. The applications often involve panel data, and yet the impact of demographic events on enterprise mobility is not considered. The article aims provide a definition of enterprise mobility, in terms of the capability to create or liquidate jobs. Moreover, some existing mobility measures are modified so that they also take into account newborn and exiting firms. The proposed index has all the relevant basic properties which make it a rigorous descriptive statistics. The mobility of Italian capital-owned enterprises in the years 2010 – 2017 is analysed in the case study.
What we propose may be an alternative tool for practitioners to measure the degree of mobility in the presence of demographic events. It may be considered an initial step in future research regarding its different applications (e.g. labour market flows or movements among income classes), also considering more complex theoretical backgrounds.

Key words: mobility measure; transition matrix; firm size; business demography.

1. Introduction

The importance of having available some descriptive statistics for measuring the mobility in an evolving sample has been widely recognized both in the past and today. Typically, a set of k non-overlapping and discrete states (also said *classes* or *categories*) is given, based on an economically relevant variable (e.g. employment or unemployment state, firm size, income classes). Movements among such classes happened between time t and time $t+1$ are usually recorded into a $k \times k$ *transition matrix*. Measuring the degree of mobility corresponds to choosing a suitable indicator able to summarize in a unique real number the global amount of movements. There are mainly two distinct ways to face this issue: $\imath$) by comparing two distributions among states at time t_0 and t_1 and measuring their "distance", as in Shorrocks (1982) and Fields and Ok (1996) (among others); $\imath\imath$) by applying a suitable function $I : \mathbb{R}^{k\times k} \to \mathbb{R}$ on the $k \times k$ transition matrix $P = \{p_{ij}\}$ as, for example, in Prais (1955); Shorrocks (1978); Bourguignon and Morrisson (2002) (see Ferretti, 2012, for an exhaustive survey). Proposals in the more recent literature include also Ferretti and Ganugi (2013); Chen and Cowell (2017); Cowell and Flachaire (2018); Paul (2019). Furthermore, the relationship of the mobility measures with some theoretical stochastic processes, possibly underlying the dynamics under study, and the sampling properties of a given mobility index have been treated in Geweke et al. (1986); Schluter (1998); Formby et al. (2004) and Ferretti (2014).

[1]DISES, Univ. Cattolica del Sacro Cuore, Piacenza, Italy. E-mail: camilla.ferretti@unicatt.it; cami.ferretti@gmail.com. ORCID: https://orcid.org/0000-0003-4508-5127.

In this work, we focus specifically on the Firms Size and Business Demography, and on the issue of measuring firms mobility using transition matrices (TMs herein). It is an attempt to construct an indicator able to unify two relevant features in any set of firms evolving with respect of time: on the one hand we aim to measure the global mobility, intended as capacity to move among size classes (a more refined definition will be provided in the following sections); on the other hand we would like to measure the effect of demographic events (births and deaths) on the mobility as a whole. As a matter of fact, these two features are usually treated separately: both for descriptive and estimating purposes, movements among different states are often stored in a transition matrix; whereas national statistical bureaus usually provide the birth and death rates without considering the mobility among size classes, as in the technical report ISTAT (2019).

In the following we propose a step-by-step procedure to define mobility in the Firm Size framework, accounting at the same time for the effect of Business Demography. Starting from the fact that the existing mobility indices are additively decomposable in k contributions, representing the mobility in different parts of the size range, we introduce some ad-hoc modifications to obtain a new index which measures both the prevalent tendency towards upsizing/downsizing, and the number of created/destroyed job places. The new index represents for now an alternative and quite easy-to-handle descriptive measure for practitioners, because it can be evaluated having at disposal the sole aggregated data (transition matrix, birth rate and death rate).

This work is organized as follows: Section 2 provides a short survey about mobility indices measured with transition matrices; Section 3 introduces our concept of mobility in the Firm Size framework; Section 4 recalls the usual way to set TMs when births and deaths are considered and proposes a formalization of the effect of newborns and exiting enterprises on the mobility; in Section 5 the main properties of the new index are analysed (in particular the decomposition in terms of birth and death events); Section 6 contains the empirical example regarding the mobility of Italian capital-owned Employer Enterprises belonging to B – E sectors in NACE Rev. 2 (Industry except construction) for the years 2010 – 2017. The last Section concludes.

2. Measuring mobility with TMs

Considering a set of k non-overlapping size classes (for example, firms can be divided according the official definition of "micro", "small", "medium" and "large"), we observe the size of firms in two consecutive instants t and $t+1$. Note that a set of *Incumbents* is required[2], intended as a group of firms being already active at time t and still active at time $t+1$, to count the number of transitions among the size classes. We choose here to work in a non-parametric framework, in the sense that we aim to build a mobility indicator which does not require underlying theoretical assumptions. In consequence of that, p_{ij} will be considered as a relative frequency (or empirical probability) instead of a theoretical probability. The empirical transition matrix (TM) is thus defined as usual by $P = \{p_{ij}\}_{i,j=1,\dots,k}$

[2]The term *Incumbents* is generally referred to a set of firms already in position in a market. For extension we use the same term to indicate firms observed for the whole considered interval of time.

such that

$$p_{ij} = \frac{n_{ij}}{\sum_{l=1}^{k} n_{il}}, \tag{1}$$

where n_{ij} is the number of firms moved from the i-th to the j-th class, for every couple $i, j = 1, \ldots, k$.[3] Consequently, p_{ij} is the relative frequency of movements between i and j, conditioned to the starting size class i.

In this framework, a *mobility index* is a function $I : [0,1]^{k \times k} \to \mathbb{R}$. Given two generic groups A and B of enterprises, and the corresponding TMs P_A and P_B, by definition A is said to have a higher degree of mobility than B if and only if $I(P_A) > I(P_B)$. Main proposals in the literature are, among others:

- the trace index $I_{tr}(P) = \frac{1}{k} \sum_i (1 - p_{ii})$ (Prais, 1955; Shorrocks, 1978);
- the index $I_b(P) = \frac{1}{k} \sum_{i,j} p_{ij} |j - i|$ (Bartholomew, 1982);
- the up/downward index $I_{up}(P) = \sum_i f_i \sum_{j>i} p_{ij}$ and $I_{down}(P) = \sum_i f_i \sum_{j<i} p_{ij}$, where f_i is the percentage of firms moving from i (Bourguignon and Morrisson, 2002);
- the directional index $I^v_{dir}(P) = \sum_i f_i \sum_j p_{ij} \cdot sign(j - i) \cdot v(|j - i|)$, where $sign(j - i) \cdot v(|j - i|)$ is included to grasp both the direction and the magnitude of jumps from i to j for every possible couple i, j (Ferretti and Ganugi, 2013).[4]

As explained in Ferretti (2012), mobility has many facets, each one measured by one or more specific indices. To better illustrate this fact, we consider the trivial and fictitious case of 300 firms with 1, 2 or 3 employees, and we suppose to observe the following numbers of transitions:

		Size at time $t+1$			
	$N = \{n_{ij}\}$	1	2	3	TOT at time t
Size at time t	1	50	40	10	100
	2	20	60	20	100
	3	10	20	70	100
	TOT at time $t+1$	80	120	100	TOT = 300

N provides many different pieces of information about the firms mobility: for example, we may be interested in measuring 1) the tendency to move one class up from the lowest class; or 2) the tendency to remain in the same size class; or 3) the global tendency to downsize. The goal of mobility indices is twofold: to provide a summarizing measure for the mobility facet we are interested in (choosing the correct index) and to furnish a rigorous indicator which is a pure number and does not depend on the total amount of considered

[3] Note that n_{ij} depends on t. The time label is discarded for the sake of simplicity, when possible.

[4] The function $sign(x)$ is defined to be equal to +1 if $x \geq 0$ and to -1 if $x < 0$. Note also that v is required to satisfy $v(0) = 0$, see Ferretti and Ganugi (2013) for more details.

firms (using P instead of N). For example, p_{12} is a possible index for the case 1) in the previous example, whereas the trace index and the downward index can be suitable for the cases 2) and 3). To conclude the example, P is easily obtained dividing the N's rows by 100, and here we display some indices evaluated following the definitions listed above:

$$I_{tr} = 0.4, \quad I_{up} = 0.233, \quad I_{down} = 0.167, \quad I_b = 0.467.$$

2.1. Additively decomposable mobility measures

To better explain the concept of mobility, we first recall the fact that mobility measures are additively decomposable, with few exceptions[5]. All the aforementioned indices, together with the most part of the existing ones, can be indeed written in the following general form:

$$I(P) = \sum_{i=1}^{k} \omega_i \cdot I_i(P_i), \tag{2}$$

where ω_i are weights such that $0 \leq \omega_i \leq 1$ for every $i = 1, \ldots, k$ and $\sum_i \omega_i = 1$. Given the i-th size class, the function $I_i : [0,1]^k \to \mathbb{R}$ measures the mobility of individuals leaving from such class, whose transitions are ruled by the i-th row $P_i = (p_{i1}, \ldots, p_{ik})$ of P. We also point out that I_i may be not well defined for vectors not belonging to the set

$$\Delta^{k-1} := \{(x_1, \ldots, x_k) \in \mathbb{R}^k : \sum_{i=1}^{k} x_i = 1, x_i \geq 0, \forall i = 1, \ldots, k\},$$

which contains all the possible probability mass functions on k discrete states.

Consequently, Equation 2 mirrors the fact that every individual starting from i has a certain degree of mobility depending on the empirical probability mass function given by P_i. All the individuals starting from the same category are supposed to have the same degree of mobility: this is a reasonable approximation because the use of TMs implies that categories are homogeneous enough to include like-minded individuals. Thus, mobility indices on TMs basically measure the mobility in different parts of the firms size' distribution, defined by the selected categories, and furnish a weighted mean of such contributions.

3. Definition of firms' mobility

Decomposability helps to give a suitable interpretation to the aforementioned mobility measures, when the firm size framework is considered. Let P be the TM obtained in the previous example, and consider the trace index decomposed as follows:

$$I_{tr}(P) = \omega_1 \underbrace{(1 - p_{11})}_{I_1(P_1)} + \omega_2 \underbrace{(1 - p_{22})}_{I_2(P_2)} + \omega_3 \underbrace{(1 - p_{33})}_{I_3(P_3)}.$$

[5]For example, the index $I(P) = det(P)$ examined in Shorrocks (1978), which is not directly decomposable in k terms related to the k size classes. However, it is implicitly based on the Markov assumption and consequently goes beyond the scope of this work.

Note that $\omega_i = 1/3$ for every i, by definition. Given i, the function I_i results to be the empirical probability to move away from i. Consequently, the trace index is a (weighted) mean probability that a generic firm will leave its starting class. If $\omega_i = f_i$ (the empirical probability to move from i), the trace index is exactly equal to the probability that a generic firm will move from its starting class. Analogously, I_{up} and I_{down} are mean probabilities that a generic firm will respectively upsize or downsize.

More relevant is the interpretation of the Bartholomew's index I_b. Considering again the aforementioned 3×3 matrix P, for every i the function I_i is defined by

$$I_i(P_i) = p_{i1}|1-i| + p_{i2}|2-i| + p_{i3}|3-i|.$$

In the previous example the quantity $|j-i|$ measures exactly the amount of job places involved by an enterprise moving from i to j, counted without considering whether they are created or destroyed. Given i, the term $|j-i|$ is equal to 0, 1, or 2 with probability p_{ij}, thus $I_i(P_i)$ is the expected number of job places which "move" together with a generic firm starting from i, under the simplistic assumption that firms hire/fire only one or two employees at once. Lastly, $I_b(P)$ is the analogous mean number considering all the starting classes. In the previous example we obtain that every firm moves on average 0.467 job places. It is worth noting that I_b results to be always not negative but it has no upper bound, because every moving firm may move potentially more than one job place. On the other hand, I_{tr}, I_{up} and I_{down} are bounded by definition in $[0,1]$.

As a last step, we revise the directional index I_{dir} from the firms' mobility point of view. The quantity $sign(j-i) \cdot v(|j-i|)$ is here included, instead of $|j-i|$, where $v(|j-i|)$ is a generalized measure for the number of involved job places by firms moving from i to j, and $sign(j-i)$ indicates job creation if $j > i$ and job destruction if $j < i$. If we set $v(|j-i|) = |j-i|$ in the previous example, we obtain:

$$I_1(P_1) = p_{12} + 2p_{13} = 0.6; \quad I_2(P_2) = -p_{21} + p_{23} = 0; \quad I_3(P_3) = -2p_{31} - p_{32} = -0.4.$$

Thus, on average, a firm in the first class creates 0.6 job places and a firm in the last class destroys 0.4 job places. Note that firms in the intermediate class move and cause both job creation and destruction, which however cancel one another. The global index is $I_{dir} = 0.067$, which indicates a modest tendency to upsize, as signalled by the positive sign of the index, and every firm creates 0.067 job places on average, under the same simplistic assumption as before.

To conclude this Section, we propose the following concept of mobility regarding Firm Size:

A suitable mobility measure should quantify the tendency to leave the starting size class, together with the direction towards up/downsize and the number of job places created/destroyed.

With this aim we will focus on the indices able to measure the tendency to upsize or downsize: I_{up} and I_{down} as defined in Section 2. We also propose a modified version of such indices, which mixes together the original definition proposed by Bourguignon and Mor-

risson (2002) and the concept of "directional mobility" introduced in Ferretti and Ganugi (2013):

$$\begin{cases} I^v_{up} = \sum_i f_i \sum_{j>i} p_{ij} v(|j-i|); \\ I^v_{down} = \sum_i f_i \sum_{j<i} p_{ij} v(|j-i|). \end{cases} \tag{3}$$

As explained before, v is a suitable function used to assign different weight to firms making movements of different magnitude (note that $|j-i|$ is the number of size classes crossed in moving from i to j). Formulas in Equation 3 mirror the concept of mobility we aim to use in the firm size analysis. In addition, being I_{up}, I_{down} and I_{tr} as defined in the previous section, it is easy to prove that:

1. $I_{up} + I_{down} = I_{tr}$, and $I^v_{up} + I^v_{down} = I^v_{tr}$, which is the trace index modified to take into account the number of job places;

2. $I_{up} - I_{down} = I_{dir}$, the directional index with $v \equiv 1$, and $I^v_{up} - I^v_{down} = I^v_{dir}$.

Hence, given v, it is worth noting that summing I^v_{up} and I^v_{down} we obtain a measure of the absolute mobility, without regarding at the direction, that recalls the *business churn* as defined in the Eurostat database (birth rate + death rate). On the other hand, subtracting I^v_{down} from I^v_{up} we obtain a net measure of the tendency to upsize or downsize, which can be considered as a mobility turnover (the *net turnover rate* is usually defined as birth rate - death rate).

4. Business Demography and Firms Mobility

Demographic events considered in a general way (appearing and disappearing individuals that can be persons, workers or firms) clearly represent main features to be analyzed in many research fields. Among others, we recall 1) the credit ratings dynamics and the Default probability estimation (Jafry and Schuermann, 2004; Violi, 2008; Ferretti et al., 2019); 2) the analysis of labour forces and the estimation of flows in and out the unemployment state (Gomes, 2012).

In the analysis of enterprises development Business Demography is relevant so that it is monitored yearly by many national statistical bureaus (see Business Demography Database, Eurostat or the Italian last official report ISTAT, 2019). Quoting the Eurostat main page about this topic: *"Business demography statistics present data on the active population of enterprises; their birth; survival ... and death. Special attention is paid to the impact of these demographic events on employment levels."* In addition, they claim that *"Business demography data can be used to analyse ... the entrepreneurship in terms of the propensity to start a new business, or the contribution of newly-born enterprises to the creation of jobs."*

Defining the mobility as the tendency to move and create job places we perfectly mirror the scope of the official statistics about Business Demography. We now revise the definition of TM in the presence of demographic events and we propose a way to merge together concepts coming from the official Business Demography and the analysis of mobility. Thus, together with the aforementioned Incumbents, we consider for every year t the enterprises which are active for the first time (*Newborns*) and the enterprises which are active for the

last time (*Exiting* firms). In evaluating movements between t and $t+1$, demographic events are treated according to the seminal papers by Adelman (1958) and Schoen (1988): the additional category O is considered and, together with the aforementioned n_{ij}, we define the number n_{io} of deaths from $i = 1,\dots,k$ (i.e. firms active at t and no longer active at $t+1$) and the number n_{oj} of births in $j = 1,\dots,k$ (firms not active at t and active at $t+1$).

Crudely, the mobility could be measured on the augmented $(k+1)\times(k+1)$ TM $P^* = \{p^*_{ij}\}_{i,j=1,\dots,k,o}$ such that:

$$\begin{cases} p^*_{ij} = \frac{n_{ij}}{\sum_{l=1}^{k} n_{il}+n_{io}} = \text{ empirical probability to move towards } j, \text{ starting from } i; \\ p^*_{io} = \frac{n_{io}}{\sum_{l=1}^{k} n_{il}+n_{io}} = \text{ empirical probability to definitely exit from } i; \\ p^*_{oj} = \frac{n_{oj}}{\sum_{l=1}^{k} n_{ol}+n_{oo}} = \text{ empirical probability to newly enter in } j; \end{cases} \tag{4}$$

for every $i, j = 1,\dots,k$.[6] The $(k+1)\times(k+1)$ matrices $N^* = \{n_{ij}\}_{i,j=1,\dots,k,o}$ and P^* indeed represent the usual and easy-to-handle way to consider births and deaths and to measure mobility in empirical applications (see Violi, 2008; Macchiarelli and Ward-Warmedinger, 2014, among others). Note that n_{ij} is not affected by entries or exits, for every couple of regular classes i and j, whereas the number of firms in every class at time t or $t+1$ is changed (for example, the number of firms leaving i is now equal to $\sum_{j=1}^{k} n_{ij} + n_{io}$).

In evaluating the mobility using P^* two main drawbacks arise. First, n_{oo} is often missing, being the non-observable number of potential newborns "waiting outside" (the unbiased estimation of the probability to be outside and to be born in the future is not trivial and goes beyond the scope of this work). Consequently, the values $p_{o1},\dots,p_{ok},p_{oo}$ may be biased and mislead the mobility measurement. Second, the outer state O is often not ordered with respect to the other categories $1,\dots,k$, and loses its exceptional nature if treated as a regular category. Hence, mobility measures evaluated on P^* may be ambiguous in the case of indices considering the ordering or the distance among categories, as in the firm size analysis. Hence, in the following paragraphs we will propose some modifications to avoid such drawbacks, with the aim to find the better choice for the weight ω^*_i and the function $I^*_i(\cdot)$, $i = 1,\dots,k$, to obtain a mobility measure in line with the context of Business Demography.

Mobility and Newborn Enterprises Official statistics usually contain information about the starting number of employees in the cohort of enterprises newly born at time t and their gain/loss of employees at $t+s$ (often with s =1,2,3,4,5). We now remark that, according with Equations 1 and 4, the number of births between time t and time $t+1$ does not affect the values p^*_{ij}, if $i \neq O$. Instead, it will affect the future mobility because it increases the number of individuals moving at time $t+1$. Therefore, the O-th row $(p_{o1},\dots,p_{ok},p_{oo})$ of P^* is not useful for measuring the mobility: it indeed represents the percentage of births in every class with respect to the total amount of newborns, quantifying the "attractiveness" of a given class, rather than its impact on the mobility as a whole.

For this reason, we propose to mirror the official statistics and to divide active enterprises

[6] From now on, the superscript * will indicate objects such as TMs or mobility indices measured considering also the demographic events.

into *long-term Incumbents* (firms active at $t-1$, t, and $t+1$) and *Newborns* (firms not active at $t-1$ and active at t and $t+1$). Deaths are temporarily ignored. The following decomposition holds: let $n_i^*(t)$ be the total number of firms being active in i at time t, which includes both the long-term Incumbents arrived from the regular classes (indicated with $n_{li}(t-1)$) and the newly born enterprises which were not active at time $t-1$ ($n_{oi}(t-1)$). Consequently, the percentage of firms in i at time t is given by:

$$f_i^* = \frac{n_i^*(t)}{n(t)} = \frac{\sum_{l=1}^{k} n_{li}(t-1) + n_{oi}(t-1)}{n(t)} = f_{i,inc}^* + f_{i,new}^*,$$

where $n(t)$ is the total number of firms active at time t, $f_{i,inc}^*$ is the fraction composed by Incumbents and the remaining part $f_{i,new}^*$ regards Newborns. Following the Eurostat birth rate's definition (*"number of enterprise births in the reference period t divided by the number of enterprises active in t - percentage"*), we can write:

$$f_i^* = f_{i,inc}^* + \frac{n_{oi}(t-1)}{n(t)} = f_{i,inc}^* + \frac{n_i^*(t)}{n(t)} \frac{n_{oi}(t-1)}{n_i^*(t)} = f_{i,inc}^* + f_i^* b_i^*,$$

where $\frac{n_{oi}(t-1)}{n_i^*(t)} = b_i^*$ is by definition the birth rate in i at time t. Thus, assuming to have a suitable function $I^*(P_i^*)$ for every i, we propose to modify Equation 2 in the following way:

$$I^*(P^*) = \sum_{i=1}^{k} (f_{i,inc}^* + f_i^* b_i^*) \cdot I_i^*(P_i^*). \tag{5}$$

Mobility and Declining Enterprises We now temporarily discard Newborns and we consider Exiting firms, intended as firms which are active at time t and no longer active at time $t+1$. As noted before, calculating $I_i^*(P_i^*)$ we lump deaths from i with regular transitions, potentially losing the peculiar information provided by p_{io}^*.

To rightly consider mortality, we first note that the effect of *Exits* is quite different from the effect of Newborns, mainly because $p_{io}^* = \frac{n_{io}(t)}{n_i^*(t)}$ corresponds by definition to the Eurostat's death rate d_i^* in $i = 1, \ldots, k$ (*"number of enterprise deaths in the reference period t divided by the number of enterprises active in t - percentage"*). Thus, the death rate is still comprised in the TM P^*. In addition, we observe that:

$$p_{ij}^* = \frac{n_{ij}(t)}{n_i^*(t)} = \frac{n_{ij}(t)}{\sum_{l=1}^{k} n_{il}(t) + n_{io}(t)} = \frac{n_{ij}(t)}{\sum_{l=1}^{k} n_{il}(t)} \frac{\sum_{l=1}^{k} n_{il}(t)}{\sum_{l=1}^{k} n_{il}(t) + n_{io}(t)} = p_{ij}(1 - p_{io}^*).$$

The same result is obtained through probabilistic facts, noting that p_{ij}^* is a conjoint probability that two events happen: 1) to not exit between t and $t+1$ and 2) to move towards j, given the starting state i. Analogously, p_{ij} is the (empirical) probability to move towards j, conditioned on both the starting state i and the survival until $t+1$. Consequently, under the same p_{ij}, the value p_{ij}^* increases when the death rate d_i^* decreases and vice-versa. Lastly, we observe that the function I^* used in the existing indices is in most cases linear with respect to its variables. Thus, to rightly consider the effect of mortality on the whole mobility, we propose the following substitution: instead of using P^*, we measure the mobility consider-

ing the sole-Incumbents $k \times k$ matrix P and we use the death rate as a rescaling factor. The resulting formula is:

$$I^*(P) = \sum_{i=1}^{k} (1 - d_i^*) \cdot I_i(P_i). \tag{6}$$

5. A measure of mobility including birth and death events

Merging Equations 5 and 6, we propose to measure mobility of enterprises using the following formula:

$$I^{bd}(P) = \sum_{i=1}^{k} (f_{i,inc}^* + f_i^* b_i^*)(1 - d_i^*) \cdot I_i(P_i), \tag{7}$$

where $I_i(P_i)$ can be retrieved from existing indices.

Equation 7 can be motivated in terms of probabilities and expected number of job places, in analogy with Section 3. Considering the trace index, and given the starting state i, only surviving firms contribute to the mobility among regular classes and their contribution is equal to the probability to move away from i, given the survival: $I_i = 1 - p_{ii}$. Such term is multiplied by the empirical probability to be a firm moving from i at t and surviving until $t+1$: $f_i^*(1 - p_{io}^*)(1 - p_{ii})$. The sum over i finally provides the global probability to leave the current size class. The same explanation holds if we are interested in the mean number of job places created/destroyed, adding a suitable measure v.

5.1. Main properties

To be a suitable descriptive measure, a generic index I is required to satisfy some properties, such as those described in Shorrocks (1978), Ferretti and Ganugi (2013) and Paul (2019). We first highlight the fact that the proposed index does not require the knowledge of the individual path of every enterprise under analysis, which are often not publicly available. It can indeed be calculated only using the transition matrix and the birth and death rates, possibly retrieved from different sources.

More rigorous properties are analyzed in the following list, reminding that the term I_i derives from an existing decomposable measure such as the trace index. Hereon we will specifically refer to the up/down mobility measures I_{up}^{bd} and I_{down}^{bd} obtained by setting I_i as displayed in Equation 3.

P1) *Immobility:* $I(Id) = 0$. Immobility requires to choose I so that the identity matrix Id, which obviously describes the absence of any type of movement, corresponds to the value 0 of mobility. Given the i-th row e_i of Id, we have $I_i(e_i) = 0$ for both I_{up}^{v} and I_{down}^{v}. Immobility is consequently proved.

P2) *Boundedness:* $I(P) \leq M$ for every TM P. Boundedness is related to the existence of a TM P corresponding to the case of maximum mobility M. For example when we measure the tendency to upsize, mobility is maximal if all the firms start from the lowest size class and jump immediately in the highest class. The maximal TM thus exists for I_{up}^{bd}, as well as for I_{down}^{bd}, and can be retrieved as suggested in Ferretti and

Ganugi (2013). Nevertheless, given two groups A and B of enterprises, it is more relevant to compare $I(P_A)$ with $I(P_B)$ instead of evaluating their distance from the maximum mobility scenario, which is quite unrealistic.

P3) *Normalization:* $0 \leq I(P) \leq 1$, for every TM P. Normalization is useful to express any mobility index as a percentage, improving its readiness. If we consider mobility as the pure probability to up/downsize, the mobility index is normalized by definition. Otherwise, as said before, I^{bd}_{up} and I^{bd}_{down} are bounded and consequently they could be re-scaled to satisfy normalization. Nevertheless, such indices can be used to measure the mean number of job places created/destroyed, which may reasonably be greater than one: therefore, normalization can be discarded without consequences.

5.2. Additional specific properties

By construction, I^{bd}_{up} and I^{bd}_{down} are equipped with some additional properties as listed in the following:

A1) *Newborns make the mobility to increase*. If the birth rate is equal to zero then $f^*_i = f^*_{i,inc}$, otherwise $f^*_i = f^*_{i,inc} + f^*_i b^*_i > f^*_{inc}$. Thus, it proved that Newborns bring a gain in the mobility, because they increase the number of firms potentially moving from every size class.

A2) *Exits make the mobility to decrease*. In analogy with Newborns, we observe that if the death rate is greater than zero then $f^*_i(1-d^*_i) < f^*_i$. Thus, deaths cause a loss in the mobility because they reduce the number of transitions among the regular size classes.

A3) *Weights do not sum up to one*. In Equation 7 we use the peculiar weights $\omega^*_i = (f^*_{i,inc} + f^*_i b^*_i)(1-d^*_i)$ such that $\sum_i \omega^*_i = 1 - \sum_i f^*_i d^*_i \leq 1$. Obviously, it is possible to substitute ω^*_i with $\frac{\omega^*_i}{1-\sum_i f^*_i d^*_i}$. Nevertheless, reminding that the considered indices are defined to be the empirical probability to upsize/downsize (if $v \equiv 1$) or the expected number of job places created/destroyed, in our opinion the weights normalization is not useful to improve the index readiness. In addition, the missing fraction $\sum_i f^*_i d^*_i$ is related to the loss caused by exiting enterprises, as explained by property A2.

A4) *The index is decomposable in terms of births and deaths effects*. Indeed, we can write:

$$I^{bd}(P) = \sum_{i=1}^{k} (f^*_{i,inc} + f^*_i b^*_i)(1-d^*_i) \cdot I_i(P_i) = M_{inc} + M_{new} + M_{ex} + M_{newex}.$$

In details:

- $M_{inc} = \sum_{i=1}^{k} f^*_{i,inc} \cdot I_i(P_i)$ is the mobility due to the long-term Incumbents (enterprises active at $t-1$, t and $t+1$);
- $M_{new} = \sum_{i=1}^{k} f^*_i b^*_i \cdot I_i(P_i)$ is the gain in the mobility due to Newborns (enterprises active at t and $t+1$);

- $M_{ex} = -\sum_{i=1}^{k} f^*_{i,inc} d^*_i \cdot I_i(P_i)$ is the loss in the mobility due to exiting Incumbents (enterprises active at $t-1$ and t);
- $M_{newex} = -\sum_{i=1}^{k} f^*_i b^*_i d^*_i \cdot I_i(P_i)$ is the loss due to immediately exiting Newborns (enterprises active only at t).

Note that the above decomposition is meaningful is we use the not-normalized weights ω^*_i, as explained by property A3. Lastly, the same decomposition can be based also on a wider time window, considering for example transitions between t and $t+5$, instead of $t+1$, with the aim to mirror the official statistics, which often provide information about the cohort of Newborns after five years.

6. The mobility of Italian capital-owned manufacturing firms

6.1. Data description

Our source is the AIDA database by Bureau van Dijk (https://aida.bvdinfo.com), which collects annual accounts of the Italian enterprises, with a strong prevalence of capital-owned firms. To our knowledge, no other databases are publicly available to observe firms transitions, thus we will focus the analysis on this specific legal form. The last release covers the years 2009 – 2018, and we discard the first and the last year, which are potentially biased. In addition, to reduce the computational effort and for the sake of homogeneity, we select only firms belonging to sectors B - E in NACE Rev. 2 (Industry except Construction) which is considered separately also by the Italian Statistical Office in the annual report (see ISTAT, 2019). Data are cleaned considering only firms having a defined number of employees for every year of activity. As further check, we also exclude enterprises with missing values in the Annual Turnover or in the Balance Sheet Total. Lastly, we discard reactivated firms and firms that ceased by merging or division, with the aim to reproduce the official technique for identifying true births and deaths.

The resulting dataset covers eight years 2010 – 2017 and 52937 enterprises. To make possible the comparison with the official indicators, size categories are defined according to the Eurostat definition in terms of employees: no employees, from 1 to 4, from 5 to 9, more than 10 employees (see Business Demography Database, Eurostat). The observed numbers of transitions are displayed in Table 8 in the Appendix. Unfortunately, we note that enterprises with no employees are severely under-represented in the AIDA dataset with respect to the official data, given that they cover around 14% of the set against 40% on average resulting from the Eurostat database (see section (b) of Table 8). To avoid drawbacks, we thus choose to work with the subset of *Employer Enterprises*, defined as firms with at least one employee, as explained in Eurostat-OECD manual (2007). According to the same manual:

- *"an Employer Enterprise Birth occurs either as an enterprise birth with at least one employee in the year of birth, or as an entry by growth reaching the threshold of one employee"*;
- *"an Employee Enterprise Death occurs either as an enterprise death with at least one employee in the year of death or as an exit by decline, moving below the threshold of one employee"*.

Table 1 contains the main indicators about the employer enterprises. Panel A displays the annual fraction of firms moving from each size category. We can see that, except for the first two years, the percentage of "1 – 4" firms (51% on average) is similar to official data (56%), whereas percentages regarding the other two classes are reversed, given that "5 – 9" firms represent on average 31% of the AIDA set and less than 20% in the Eurostat data.

Panels C and D contain the annual birth and death rates. Years 2010 and 2011 show exceptional and not explainable values which are consequently excluded in evaluating the average with respect of time. In the remaining years, capital-owned firms result to have a different demography with respect to Eurostat data, being the birth rate very high for small firms and generally higher than the death rate. In opposition, the official Italian turnover rate is always negative, as revealed by the Eurostat's averages (see also ISTAT, 2016, which provides rates disaggregated by size class). Consequently, if only the birth and death rates are considered, Italian capital-owned enterprises seem to enjoy good health, and the economic crisis seems to solely cause the negative turnover of the largest firms since 2012. Reminding that we are working with a specific set of firms (capital-owned employer enterprises belonging to Industry except Constructions), the analysis of the determinants of this quite surprising behaviour goes beyond the scope of this work.

Table 1: Main indicators for the Italian capital-owner Employer Enterprises.

	2010	2011	2012	2013	2014	2015	2016	2017	Average	EUROSTAT^{+}
	PANEL A: Percentage of firms by size class									
From 1 to 4	44.60%	47.99%	49.61%	51.67%	52.48%	52.39%	53.62%	53.40%	50.72%	55.81%
From 5 to 9	28.13%	33.00%	32.21%	30.65%	30.08%	30.89%	29.92%	30.49%	30.67%	19.31%
10 or more	27.28%	19.02%	18.17%	17.69%	17.43%	16.73%	16.46%	16.11%	18.61%	24.88%
Tot Firms	17777	29994	31360	31731	32476	34225	35165	36454	31148	257407
	PANEL B: Mean number of employees by size class									
From 1 to 4	2.44	2.47	2.43	2.41	2.38	2.39	2.37	2.37	2.41	1.83
From 5 to 9	6.48	6.48	6.44	6.38	6.36	6.37	6.37	6.41	6.41	6.63
10 or more	46.67	42.01	42.48	42.72	42.38	42.49	43.05	43.20	42.62	46.11
Tot Employees	278034	339347	344984	341366	342697	353389	360946	371102	350547	3520079
	PANEL C: Birth rate by size class									
From 1 to 4	20.17%	53.79%	12.45%	12.24%	14.73%	17.60%	13.13%	12.93%	13.85%*	9.35%
From 5 to 9	18.20%	44.97%	4.33%	5.18%	5.52%	7.56%	4.55%	4.96%	5.35%*	1.90%
10 or more	5.26%	14.78%	1.74%	1.51%	1.78%	1.76%	0.97%	0.95%	1.45%*	0.75%
TOT	15.55%	43.46%	7.89%	8.18%	9.70%	11.85%	8.56%	8.57%	9.12%*	5.78%
	PANEL D: Death rate by size class									
From 1 to 4	7.06%	5.32%	10.39%	11.28%	10.55%	9.16%	7.89%	9.07%	9.72%*	10.90%
From 5 to 9	4.40%	2.80%	4.85%	4.67%	4.10%	3.24%	2.60%	2.43%	3.65%*	2.53%
10 or more	0.80%	1.12%	2.05%	1.84%	1.89%	1.48%	1.28%	1.14%	1.61%*	1.13%
TOT	4.61%	3.69%	7.09%	7.58%	7.10%	6.05%	5.22%	5.77%	6.47%*	6.86%
	PANEL E: Net Turnover (birth rate - death rate)									
From 1 to 4	13.11%	48.47%	2.06%	0.96%	4.18%	8.44%	5.25%	3.86%	4.12%*	-1.54%
From 5 to 9	13.80%	42.17%	-0.52%	0.51%	1.41%	4.31%	1.95%	2.53%	1.70%*	-0.63%
10 or more	4.45%	13.66%	-0.32%	-0.32%	-0.11%	0.28%	-0.31%	-0.19%	-1.61%*	-0.38%
TOT	10.94%	39.77%	0.79%	0.60%	2.60%	5.80%	3.34%	2.80%	2.66%*	-1.08%

Source: Own calculations on the AIDA data.* Averages are calculated excluding 2010 and 2011.

$^{+}$Source: Eurostat's database "Employer Business Demography by size class (from 2004 onwards, NACE Rev. 2)"

7. Results

We finally measure the mobility in the set under analysis. Reminding the definition of mobility in the firm size framework, and Equations 3 and 7, we propose to measure the tendency to upsize and create job places in the following way:

$$I_{up}^{bd} = \sum_{i=1}^{k} (f_{i,inc}^{*} + f_i^{*} b_i^{*})(1 - d_i^{*}) \sum_{j>i} p_{ij} v(|j-i|).$$

As a first step we set $v(|j-i|) \equiv 1$ to evaluate the probability to move upward. As a second step, in line with the example proposed in Section 3, we choose a suitable v to measure the effort in terms of job places required to upsize. In particular, we propose to define:

$$v(|j-i|) = |\text{(mean number of employees in } j\text{) - (mean number of employees in } i\text{)}|.$$

To better explain this choice, consider for example enterprises in the first and in the second size class. On average, they employ respectively 2.41 and 6.41 workers (see Panel B in Table 1). Thus, the mean distance, intended as the size difference, between firms in the first and in the second class is equal to 4. In other words, a firm in the first class is required to hire four employees on average, to move to the second class. Analogous facts hold for I_{down}^{bd}.

Figure 1 shows the results of the annual mobility indices (the anomalous values in 2010 – 2011 are discarded). Values are equipped with standard errors obtained simulating 1000 bootstrapped copies of the same set (the index sampling distribution results to be Gaussian as proved in Ferretti, 2014). In the left panel we note that the probability to move towards every direction among size classes is very low, being both the upward and the downward indices not higher that 5%. In particular, the probability to move downward is predominant for active enterprises in 2012, which was the year of economic crisis in Italy (Il Sole 24 Ore, 2013). The same negative peak is reached in the right panel, followed by a recovery during which the mean number of job places created per firm (around 0.4) is neatly higher than the amount of job places destroyed (lower than 0.3). Note also that in 2015 firms have equal probability to upsize or downsize, but upsizing firms involve more job places than the downsizing ones.

From now on we will focus only the concept of mobility intended as the tendency to create or destroy job places. Table 2 contains the mobility indices I_{up}^{bd} and I_{down}^{bd} with v defined as the mean difference between size classes, decomposed by size category. For every category we also evaluate the corresponding contribution as a percentage of the whole mobility. Note that being aggregated in a unique class, largest firms cannot become even larger and contribute to the upward mobility, as well as smallest enterprises cannot contribute to the downward mobility. Considering the mobility turnover, defined as (job places created - job places destroyed), it is worth noting that it mirrors the turnover between the birth rate and the death rate, in the sense that it is positive for the first two size classes, and negative for the largest enterprises. Nevertheless, the total turnover is negative in 2012 – 2013, as revealed by Figure 1.

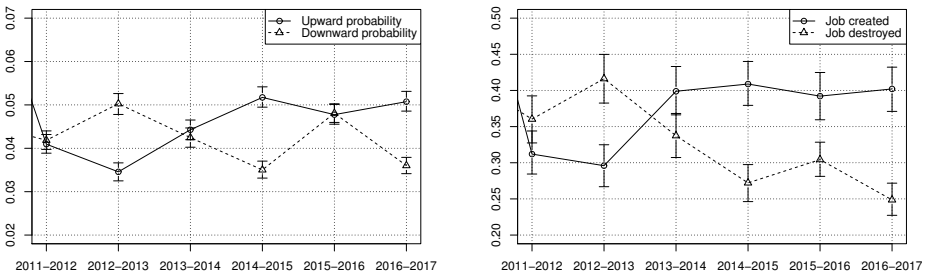


Figure 1: Mobility of Italian capital-owner Employer Enterprises recorded in AIDA. Left panel: yearly probability to move upward (solid line) or downward (dotted line). Right panel: yearly mean number of job places created (solid line) or destroyed (dotted line) per firm.

Table 2: Mean number of job places created and destroyed per firm, disaggregated by size category. In italic, the corresponding contribution as a percentage of the whole mobility.

No. of employees	2010 – 11	2011– 12	2012 – 13	2013 – 14	2014 – 15	2015 – 16	2016 – 17	Average
	PANEL A: Job places created on average per firm							
From 1 to 4	0.3850	0.1650	0.1586	0.1968	0.2050	0.1953	0.1985	0.1865*
	56.08%	*52.91%*	*53.66%*	*49.29%*	*50.18%*	*49.88%*	*49.37%*	*50.67%*
From 5 to 9	0.3015	0.1469	0.1370	0.2025	0.2035	0.1963	0.2036	0.1816*
	43.92%	*47.09%*	*46.34%*	*50.71%*	*49.82%*	*50.12%*	*50.63%*	*49.33%*
10 or more	-	-	-	-	-	-	-	-
TOT	0.6865	0.3119	0.2955	0.3993	0.4084	0.3916	0.4021	0.3681*
	100.00%	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*
	PANEL B: Job places destroyed on average per firm							
From 1 to 4	-	-	-	-	-	-	-	-
From 5 to 9	0.1568	0.1442	0.1751	0.1498	0.1242	0.1789	0.1315	0.1506*
	37.35%	*40.02%*	*42.02%*	*44.34%*	*45.61%*	*58.92%*	*52.97%*	*46.61%*
10 or more	0.2630	0.2160	0.2417	0.1880	0.1480	0.1248	0.1168	0.1725*
	62.65%	*59.98%*	*57.98%*	*55.66%*	*54.39%*	*41.08%*	*47.03%*	*53.39%*
TOT	0.4199	0.3602	0.4168	0.3377	0.2722	0.3037	0.2483	0.3231*
	100.00%	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*	*100.00%*
	PANEL C: Net turnover (job places created - job places destroyed)							
From 1 to 4	0.3850	0.1650	0.1586	0.1968	0.2050	0.1953	0.1985	0.1865*
From 5 to 9	0.1447	0.0027	-0.0382	0.0527	0.0793	0.0174	0.0721	0.0310*
10 or more	-0.2630	-0.2160	-0.2417	-0.1880	-0.1480	-0.1248	-0.1168	-0.1725*
TOT	0.2666	-0.0483	-0.1213	0.0615	0.1363	0.0879	0.1538	0.0450*

* Averages are calculated excluding 2010 – 2011.

Lastly, Table 3 shows the main property of the proposed mobility index: the decomposition in terms of births and deaths. It is relevant that the 2012's negative turnover between job places created and destroyed is mainly caused by Incumbents, which hire (resp. fire) 0.29 (resp. 0.41) employees per firm. On the other hand, newborn firms show a certain liveliness for the whole considered period, creating 0.06 job places per firm on average, against the 0.02 job places destroyed. Nevertheless, their contribution is very small, reflecting the fact that Newborns have to wait some years to become capable to create job places. In addition, until 2014 they cannot withstand the loss due to exiting enterprises. It is also worth noting that exits burden more on the tendency to upsize than the opposite tendency, causing a loss of 0.02 on average in the upsize mobility and of 0.008 in the downsize mobility. These last values can be interpreted as the mobility that exiting enterprises would have had if they had survived. Last, the loss due to immediately exiting firms is negligible.

Table 3: Upsize and downsize mobility decomposition in terms of Incumbents, Newborns and Exiting enterprises.

	2010 – 11	2011– 12	2012 – 13	2013 – 14	2014 – 15	2015 – 16	2016 – 17	Average
	Decomposition of the Upward Mobility							
M_{inc}	0.5887	0.1637	0.2927	0.3961	0.3959	0.3647	0.3867	0.3333*
M_{new}	0.1410	0.1617	0.0283	0.0382	0.0455	0.0532	0.0378	0.0608*
M_{ex}	-0.0347	-0.0066	-0.0228	-0.0314	-0.0289	-0.0223	-0.0200	-0.0220*
M_{newex}	-0.0084	-0.0069	-0.0026	-0.0036	-0.0040	-0.0040	-0.0025	-0.0039*
	Decomposition of the Downward Mobility							
M_{inc}	0.3854	0.2678	0.4186	0.3375	0.2705	0.2954	0.2460	0.3173*
M_{new}	0.0438	0.0990	0.0122	0.0110	0.0098	0.0162	0.0073	0.0285*
M_{ex}	-0.0079	-0.0044	-0.0135	-0.0104	-0.0078	-0.0074	-0.0049	-0.0080*
M_{newex}	-0.0014	-0.0022	-0.0005	-0.0004	-0.0003	-0.0005	-0.0002	-0.0008*

* Averages are calculated excluding 2010 – 2011.

8. Conclusion and discussion

Here, we propose a descriptive index for measuring the mobility in an evolving set of enterprises taking into account the impact of demographic events. Firms are subdivided into k size classes, and we define "mobility" as the tendency to upsize (or downsize) together with the capability to create (or destroy) job places: in this sense the proposed index is flexible enough and can be used for measuring both the probability to upsize/downsize and the expected number of job places created/destroyed. The index is characterized by two main parameters: the weights used to measure the contribution of every size class to the mobility as a whole, and the function which quantifies the distance among size classes. Here, we mathematically prove that weights can be chosen as functions of the birth and death rates, which consequently make the mobility respectively increase and decrease. We also propose to measure the distance among size classes as the mean difference in the number of employees. As a case study, we measure the upward and downward mobility of a set of Italian capital-owned Employer Enterprises in the years 2010 – 2017. Results show that if only birth and death rates are considered, the effect of the economic crisis is not perceived, being

the turnover always positive, in contrast with the official results, which are used as benchmark. On the other hand, the new index furnishes a description which correctly mirrors the mainstream idea about the effect of the economic crisis in the years 2010 - 2017.

This work is an initial attempt to formalize the effect of birth/death events on the degree of mobility of an evolving set, using transition matrices. We do not aim to analyze the determinants of the observed mobility values, rather we try to propose an alternative tool for practitioners to better measure the mobility in the presence of demographic events. Generally speaking, the construction of a mobility indicator depends on two main issues: 1) the field of application and 2) the specific feature we are interested in (turbulence, prevailing direction, speed of convergence toward an equilibrium, if it exists, see Ferretti, 2012 for more details). For example, in Geweke et al. (1986) time-continuous economic variables are considered, together with an underlying Markov Chain model; in Ferretti and Ganugi (2013) the proposed mobility measure is thought to grasp both the direction and the distance covered moving from the i-th to j-th category; lastly, Paul (2019) refers to the case of income mobility and attaches the highest weight to mobility of poorest workers. Here, we aim to build a suitable tool for empirical applications: with this in mind we avoid specific assumptions about any possible underlying model ruling the transitions among categories. Furthermore, as explained before, our proposal is based on the fundamental idea that mobility strictly depends on the number of moving individuals: consequently, births and deaths represent respectively a gain and a loss in the whole degree of mobility. In addition, we suppose that individuals are independent one from each other in terms of birth's, death's and transition's probabilities, and that the newborns effect on the mobility is one-year lagged. Such hypotheses are considered valid having in mind the firm size dynamics. For future research they may be questioned and generalized to other relevant issues, for example considering the interaction and the competition among firms. In addition, possible further developments will consider a theoretical description of the birth and death probabilities through suitable probabilistic models, as in Duncan and Lin (1972), and a robustness/inferential analysis of the proposed index.

References

ADELMAN, I., (1958). A stochastic analysis of the size distribution of firms. J Am Stat Ass, 53(284), pp. 893–904.

BARTHOLOMEW, D. J., (1982). Stochastic Models for Social Processes, 3rd ed. London, Wiley.

BOURGUIGNON, F., MORRISSON, C., (2002). Inequality among World Citizens. Am Econ Rev, 92(4), pp. 727–744.

EUROSTAT BUSINESS DEMOGRAPHY, retrieved from https://ec.europa.eu/eurostat/web/structural-business-statistics/business-demography, accessed May, 18, 2020.

CHEN, Y., COWELL F., (2017). Mobility in China. Rev Income Wealth, 63, pp. 203–218.

COWELL, F., FLACHAIRE E., (2018) Measuring Mobility. Quant Econ, 9, pp. 865–901.

DUNCAN G., LIN L., (1972). Inference for Markov chains having stochastic entry and exit. J Am Stat Ass, 67(340), pp. 761–767.

EUROSTAT – OECD, (2007). Manual on Business Demography Statistics, Luxembourg: Office for Official Publications of the European Communities, ISBN: 978-92-79-04726-8.

FERRETTI, C., (2012). Mobility as prevailing direction for transition matrices. Stat Appl, 10(1), pp. 67–79 .

FERRETTI, C., (2014). Sampling properties of the directional mobility index and the income of Italian families. Statistica, 4, pp. 455–466.

FERRETTI, C., GANUGI, P., (2013). A new mobility index for transition matrices. Stat Method Appl, 22(3), pp. 403–425.

FERRETTI, C., GABBI, G., GANUGI, P., SIST, F., VOZZELLA, P., (2019). Credit Risk Migration and Economic Cycles. Risks, 7(109), pp. 1–18.

FIELDS, G., OK, E., (1996). The Meaning and Measurement of Income Mobility. J Econ Th, 71(2), pp. 349–377.

FORMBY, J.P., SMITH, W.J., ZHENG, B., (2004). Mobility measurement, transition matrices and statistical inference. J Econometrics, 120, pp. 181–205.

GEWEKE, J., MARSHALL, R., ZARKIN, G., (1986). Mobility Indices in Continuous Time Markov Chains. Econometrica, 54(6), pp. 1407–1423.

GOMES, P., (2012). Labour Market Flows: Facts from the United Kingdom. Labour Economics, 19(2), pp. 165–175.

ISTAT, (2016). Demografia d'impresa 2009–2014. Tech. rep., Centro diffusione dati.

ISTAT, (2019). Demografia d'impresa 2012–2017. Tech. rep., Centro diffusione dati.

IL SOLE 24 ORE, 26th March 2013, https://st.ilsole24ore.com/art/notizie/2013-03-26/italia-paese-colpito-crisi-123024.shtml#comments.

JAFRY, Y., SCHUERMANN, T., (2004). Measurement, estimation and comparison of credit migration matrices. J Bank Financ, 28(11), pp. 2603–2639.

MACCHIARELLI, C., WARD-WARMENDINGER, M., (2014). Transitions in labour market status in EU labour markets. IZA J Eur Labour Stud, 3(1), pp. 1–17.

PAUL, S., (2019). A measure of Income Mobility based on Transition Matrices and application to China and the United States. J Asia Pacific Econ., 25(3), pp. 389–401.

PRAIS, S. J., (1955). Measuring social mobility. J Roy Stat Soc, Series A, part I, 188, pp. 56–66.

SHOEN, R. (1988). Practical uses of Multistate Population Models. Ann Rev Sociol, 14, pp. 341–361.

SCHLUTER, C., (1998). Statistical inference with mobility indices. Econ Lett, 59, pp. 157–162.

SHORROCKS, A., (1978). The measurement of Mobility. Econometrica, 46(5), pp. 1013 – 1024.

SHORROCKS, A., (1982). On the distance between income distributions [inequality measures between income distributions with applications]. Econometrica 50(5), pp. 1337–1339.

VIOLI, R., (2008). Credit Ratings Transition in Structured Finance. J Financ Transf, 22, pp. 1–36.

APPENDIX

A. AIDA's transition matrices

Table A: (a) Observed transitions among size classes and births/deaths per size class in the AIDA dataset. (b) AIDA's and Eurostat's initial frequencies per size class. (c) Transition matrix for the subset of incumbent Employer Enterprises.

	(a)						(b)		(c)		
	No emp.	From 1 to 4	From 5 to 9	10 or more	Deaths	TOT	Obs. %	Eurostat %$^{+}$	TM for employer enterprises		
2010 – 2011											
No employees	2946	6968	4140	794	409	15257	46.19%	40.63%			
From 1 to 4	282	5923	1416	29	278	7928	24.00%	32.74%	80.39%	19.22%	0.39%
From 5 to 9	88	697	3937	146	132	5000	15.14%	11.65%	14.58%	82.36%	3.05%
10 or more	12	31	93	4686	27	4849	14.68%	14.97%	0.64%	1.93%	97.42%
Births	905	774	311	49	17864	19903					
TOT	4233	14393	9897	5704	18710	52937					
2011 – 2012											
No employees	2632	1198	175	33	195	4233	12.81%	39.57%			
From 1 to 4	370	12518	1095	14	396	14393	43.57%	33.97%	91.86%	8.04%	0.10%
From 5 to 9	62	1081	8419	120	215	9897	29.96%	11.61%	11.24%	87.52%	1.25%
10 or more	16	23	151	5466	48	5704	17.27%	14.86%	0.41%	2.68%	96.91%
Births	877	739	262	66	16766	18710					
TOT	3957	15559	10102	5699	17620	52937					
2012 – 2013											
No employees	2570	999	155	36	197	3957	11.98%	39.68%			
From 1 to 4	846	12974	938	30	771	15559	47.10%	34.16%	93.06%	6.73%	0.22%
From 5 to 9	143	1373	8122	117	347	10102	30.58%	11.53%	14.28%	84.50%	1.22%
10 or more	29	41	161	5380	88	5699	17.25%	14.64%	0.73%	2.88%	96.38%
Births	1078	1007	349	49	15137	17620					
TOT	4666	16394	9725	5612	16540	52937					
2013 – 2014											
No employees	2831	1301	168	38	328	4666	14.12%	40.71%			
From 1 to 4	999	13314	1195	36	850	16394	49.63%	33.67%	91.54%	8.22%	0.25%
From 5 to 9	138	1188	7908	175	316	9725	29.44%	11.25%	12.81%	85.30%	1.89%
10 or more	31	32	127	5350	72	5612	16.99%	14.36%	0.58%	2.31%	97.11%
Births	1170	1210	371	63	13726	16540					
TOT	5169	17045	9769	5662	15292	52937					
2014 – 2015											
No employees	2741	1669	270	44	445	5169	15.65%	40.96%			
From 1 to 4	1003	13747	1481	18	796	17045	51.60%	33.20%	90.17%	9.71%	0.12%
From 5 to 9	119	1008	8180	180	282	9769	29.57%	11.32%	10.76%	87.32%	1.92%
10 or more	27	18	111	5426	80	5662	17.14%	14.52%	0.32%	2.00%	97.68%
Births	799	1487	529	57	12420	15292					
TOT	4689	17929	10571	5725	14023	52937					
2015 – 2016											
No employees	2852	1175	136	21	505	4689	14.19%	40.06%			
From 1 to 4	900	14835	1427	24	743	17929	54.27%	33.88%	91.09%	8.76%	0.15%
From 5 to 9	94	1531	8514	183	249	10571	32.00%	11.37%	14.97%	83.24%	1.79%
10 or more	28	12	103	5525	57	5725	17.33%	14.69%	0.21%	1.83%	97.96%
Births	911	1301	343	35	11433	14023					
TOT	4785	18854	10523	5788	12987	52937					
2016 – 2017											
No employees	2815	1226	184	26	534	4785	14.49%	40.93%			
From 1 to 4	816	15778	1572	17	671	18854	57.07%	32.37%	90.85%	9.05%	0.10%
From 5 to 9	69	1156	8898	195	205	10523	31.86%	11.61%	11.28%	86.82%	1.90%
10 or more	17	17	93	5604	57	5788	17.52%	15.08%	0.30%	1.63%	98.07%
Births	852	1291	367	30	10447	12987					
TOT	4569	19468	11114	5872	11914	52937					

Source: Own calculation on AIDA's data 2010 – 2017

$^{+}$Source: Eurostat database "Business Demography by size class (from 2004 onward, NACE Rev. 2)"

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 75–88, DOI 10.21307/stattrans-2021-004
Received – 15.03.2019; accepted – 24.06.2020

Life expectancy in West African countries: Evidence of convergence and catching up with the north

OlaOluwa S. Yaya[1], **Oluwaseun A. Otekunrin**[2], **Ahamuefula E. Ogbonna**[3]

ABSTRACT

The article aims to investigate the possibility of the convergence and catching up of life expectancy values observed in West African countries with those noted in North African countries. Following the theory of time series convergence, documented in Bernard and Durlauf (1996) and Greasley and Oxley (1997), more robust unit root tests, based on the Fourier nonlinearity and instantaneous breaks proposed in Furuoka (2017), are used in investigating the convergence of each pair of a West African country and its North African counterpart. As no unit root in the differences of the pairs implies convergence, the results obtained by means of the new statistical approach quite outperform those produced by classical unit root tests. The results provide general evidence of the convergence of life expectancy values recorded in West Africa and North Africa.

Key words: Africa, convergence, Fourier function, life expectancy ratio, nonlinearity, unit root.

1. Introduction

Life expectancy is a statistical measure of the average number of years a person is expected to live. It is one of the indicators used for measuring the well-being of a population (Ortiz-Ospina, 2017). Factors contributing to life expectancy values include the educational level of individuals, access to the quality health system, economic empowerment, health behaviours among others (HPF, 2012).

Recent statistics show that Hong Kong has a life expectancy of 84.462 years, the highest in the world, followed by Japan with a value of 83.995 years. Furthermore, the

[1] Economic and Financial Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria & Honorary Research Fellow, ILMA University, Karachi, Pakistan. E-mail: os.yaya@ui.edu.ng; o.s.olaoluwa@gmail.com. ORCID: https://orcid.org/0000-0002-7554-3507.

[2] Statistical Design of Investigations Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria. E-mail: oa.otekunrin@ui.edu.ng. ORCID: https://orcid.org/0000-0002-4193-1413.

[3] Economic and Financial Statistics Unit, Department of Statistics, University of Ibadan, Ibadan, Nigeria & Centre for Econometric and Allied Research, University of Ibadan, Ibadan, Nigeria. E-mail: ae.ogbonna@cear.org.ng. ORCID: https://orcid.org/0000-0002-9127-5231.

life expectancy values of France, Australia, Israel, Iceland, and the United States are 82.74, 82.91, 82.93, 83.06, and 79.50 years, respectively (World Population Review, 2017). The average life expectancy in Europe was 75 years for males and 82 years for females in 2018; Western Europe had 79 years for males and 84 years for females, while Eastern Europe had 69 years for males and 78 years for females. The average life expectancy in Africa in 2019 was 61 years for males and 65 years for females (Statista, 2019). Life expectancy values from the Northern Africa region are higher than those from other regions of Africa. Many African countries, especially West African countries, have low life expectancies because they are still grappling with high under-five mortality rates, extreme poverty, hunger, high level of illiteracy, lack of access to quality medical care, environmental hazards, HIV/AIDS, malaria, road accidents, conflicts, wars, lifestyle diseases, among others (Nkalu and Edeme, 2019; Otekunrin et al. 2019a; Otekunrin et al. 2019b; Uchendu, 2018; Mondal and Shitan, 2013; Wiysonge, 2018). Therefore, this study is motivated by the need to provide evidence-based results that would be beneficial to policymakers in West African countries as they strive towards improving the well-being of their populace.

Kontis et al. (2017) projected future life expectancy in 35 industrialized countries with a Bayesian model ensemble. One of their results showed a 90% probability that life expectancy at birth will be greater than 86.7 years in 2030, among South Korean women. Nmeth and Missov (2018) presented the Gamma-Gompertz-Makeham model as a better alternative to existing techniques for calculating life expectancy values because of its ability to adequately handle right-censoring in the last open-ended age group of a life table, especially, where this group has the highest proportion of the population. They showed the applicability of this model in revising life expectancy trends of historical populations usually used for mortality forecasts. Using high correlations in life expectancy values between females and males over time, and among different countries, Pascariu et al. (2018) proposed the double-gap life expectancy forecasting model to predict life expectancy values and compared their results with two popular approaches.

Stevens et al. (2019) developed a technique for computing a reference life expectancy metric using mathematical simulations. The developed model was able to give accurate predictions of the life expectancy of the United States population. Barthold Jones et al. (2018) developed formal demographic measures for studying relationships between the shared life expectancy of two birth cohort peers, the proportion of their lives expected to overlap, and longevity. Their results showed that almost all changes to mortality schedules that result in higher life expectancies also result in higher proportions of life shared. van Baal et al. (2016) extended the Li-Lee model originally developed for coherent mortality forecasts for a group of populations

by adapting it to forecasts of life expectancy for different educational groups within a population. Using the population aged 65 and above in the Netherlands, the results of the analyses implied an increase in life expectancy for all educational groups coupled with widened differences in life expectancy between educational groups.

Since there are variations in the life expectancy globally based on geographical locations, the present paper, therefore, investigates the possibility of convergence of life expectancy values in West African countries to those of North Africa, utilizing the time series approach (see Bernard and Durlauf, 1996; Greasley and Oxley, 1997; and Cuñado and Pérez de Gracia, 2006; among others). These authors proposed convergence in the time series whenever the difference of natural logs of bigger and smaller magnitude time series is stationary. In this case, the bigger magnitude time series is the life expectancy of countries in North Africa, while the smaller magnitude time series is the life expectancy of countries in West Africa. We then conduct unit root tests on the log differences to establish convergence, in this context. The rejection of the null hypothesis in a given life expectancy difference series implies the impossibility of life expectancy of the West Africa country to meet up with that of the corresponding North Africa in the West and North Africa country pair.

Following Yaya, Ogbonna and Atoi (2019) and Yaya, Ogbonna and Mudida (2019) study, we adopt a battery of ADF-based unit root testing frameworks. We adopt the classical Augmented Dickey-Fuller [ADF] (Dickey and Fuller, 1979) unit root test as a base testing framework, which is known to have limitations in the presence of structural breaks of any form and nonlinearity. Also, the ADF test result is inconsistent in a small sample, since the lag augmentation component is required in the implementation, and such requires a long time series. Furthermore, other robust unit root testing frameworks applied herein include those established in Perron and Vogelsang (1992), Enders and Lee (2012a,b) and Furuoka (2017). These are the ADF with structural break (ADF-SB) test, Fourier ADF (FADF) and FADF with breaks (FADF-SB) tests, respectively. The Fourier ADF induces smooth breaks, unlike the instantaneous break induced by the ADF-SB test of Perron and Vogelsang (1992) ADF-SB test. Our consideration of unit root testing frameworks that incorporate each of Fourier function, structural breaks and both, is informed by our interest to account for plausible nonlinearity and structural breaks. Details of the FADF-SB test and its variant subsets are described in Section 2 that follows.

Following the introductory section, the rest of the paper is structured as follows: Section 2 provides a detailed description of the data and methods employed in the study. Section 3 presents some empirical results and the interpretation of the findings, while Section 4 concludes the paper with some health policy implications.

2. Data and Methods

The data used in this work are the annual life expectancy at birth for newborn infants in North and West African regions, spanning a period between 1960 and 2016. These estimates were obtained in 2017 by the United Nations Population Division[4]. The sampled countries in the Northern African region are Algeria (ALG), Libya (LBY), Egypt (EGY), Morocco (MOR) and Tunisia (TUN); while those West African countries are Benin (BEN), Burkina Faso (BFA), Cote D'Ivoire (CIV), Mali (MLI), Niger (NER), Nigeria (NGA) and Senegal (SEN). These countries were selected based on geographical locations and disparities, as we reduced the number of paired differences. Meanwhile, North Africa has seven countries in which five countries were selected, while the West African region has 17 countries in which seven countries were selected. The selected African countries are paired - one from the Northern African region and the other from the Western African region.

Consequently, given that there are five (5) selected North African countries and seven (7) selected West African countries, our sample, therefore, comprises a total of thirty-five (35) of such paired samples. The differences between the log-transformed annual life expectancy of the paired African countries (one North African country and one West African country) are obtained. The time series of these obtained differences in life expectancy between the African country pairs are thereafter examined for the presence of unit root, as a test for the plausibility of convergence of the annual life expectancy of the West African counties to their Northern counterparts, which have been earlier shown to have higher life expectancies.

The methodology for convergence of life expectancy is similar to that of income convergence defined in Bernard and Durlauf (1996) and Greasley and Oxley (1997). This framework is widely applied in convergence and catching up theory, where the difference in life expectancy is given as,

$$\lim_{K\to\infty} E(y_{i,t+k} - y_{j,t+k}|I_t) = \lim_{K\to\infty} E(IG_{ij,t+k}|I_t) = 0 \qquad (1)$$

where E is the expectation operation on the difference, $IG_{ij,t}$ between the two time series, i and j in the year t; $y_{i,t}$ is the logarithm of life expectancy of a North African country, while $y_{j,t}$ is the life expectancy of a country in the West African region. I_t is information available up to time t. The null hypothesis of convergence of life expectancy of any West African country to that of a North African country is rejected if the long-term predicted life expectancy gap has a unit root, otherwise, the hypothesis of possible convergence is not rejected.

[4] World Population Prospects: 2017 Revision.

Empirically, to test for a unit root in the differences defined in (1), we adopt an unrestricted ADF-based model with Fourier functions proposed by Furuoka (2017), which has also been applied to several times series in extant literature (see Yaya, Ogbonna and Atoi, 2019; Yaya, Ogbonna and Mudida, 2019; among others). This is the Fourier Augmented Dickey-Fuller with structural break (FADF-SB) test, given by the testing regression,

$$\Delta y_t = \mu + \beta t + \gamma_1 \sin\left(2\pi kt/N\right) + \gamma_2 \cos\left(2\pi kt/N\right) + \delta DU_t + \theta D\left(T_B\right)_t + \left(\rho - 1\right) y_{t-1} + \sum_{i=1}^{p} c_i \Delta y_{t-i} + \varepsilon_t \tag{2}$$

where $\Delta = \left(1 - B\right)$ and $\pi = 3.1416$; μ represents the constant term; β and ρ are, respectively, the slope parameters for the trend term t and the lagged dependent variable y_t, with $\rho = 1$ indicating unit root; γ_1 *and* γ_2 are the Fourier function slope parameters; k is the Fourier frequency; N is the number of observations; T_B indicates the point of observed structural break; δ and θ are, respectively, the slope parameters for the structural break dummy $\left(DU_t\right)$ and the one-time break dummy $\left(D\left(T_B\right)\right)$. We define $DU_t = 1$ if $t > T_B$ and $DU_t = 0$, otherwise; and $D\left(T_B\right)_t = 1$ whenever $t = T_B$ and $D\left(T_B\right)_t = 0$, otherwise. From the test regression in (2), in the absence of structural break dummies (where the structural break is not significant), the FADF-SB test reduces to the FADF test, and similarly in the testing regression. Also, whenever the Fourier parameters γ_1 and γ_2 are not significantly different from 0, the FADF test further reduces to the classical ADF test. In the FADF-SB test, insignificancy of Fourier parameters alone calls for testing ADF-SB regression in judging unit root. Thus, we have three other variants of FADF-SB tests: FADF, ADF-SB and ADF unit root tests. For details about each of these tests, readers are referred, respectively, to Enders and Lee (2012a, 2012b), Perron and Vogelsang (1992) and Dickey and Fuller (1979).

The preference for a test over the others is not arbitrarily or trivially implied on the basis of rejecting the null of a unit root. Preference for model adequacy is, however, made in a more formal way using the F-test as suggested by Furuoka (2017). This approach entails a comparison of a restricted model with an unrestricted model, and is given by:

$$F = \frac{\left(SSR_0 - SSR_1\right)/q}{SSR_1/\left(T - r\right)} \tag{3}$$

where SSR_0 and SSR_1 represent the sum of squares residuals (SSR) from the restricted and the unrestricted models, respectively; q and r are, respectively, the number of restrictions in the restricted model and the number of regressors in the unrestricted model. By this, the ADF regression model is perceived as the restricted ADF-SB model whenever the series is not characterized by the presence of structural breaks. In the same vein, the ADF regression model is a restricted FADF model whenever the Fourier function parameters are not significant. For the FADF-SB model case, the ADF regression is a restricted FADF-SB model whenever both nonlinearity and structural breaks are absent. The FADF model, in a similar manner, is a restricted FADF-SB model when there is no structural break in the series. Finally, the ADF-SB could be considered a restricted FADF-SB model whenever nonlinearity functional form is absent. These combinations of restricted and unrestricted models result in five pairs, which are tested using the F-test. They include $F_{(FADF,\ ADF)}$, $F_{(ADF-SB,\ ADF)}$, $F_{(FADF-SB,\ ADF)}$, $F_{(FADF-SB,\ FADF)}$ and $F_{(FADF-SB,\ ADF-SB)}$ tests, and serve as robustness checks in this paper.

3. Empirical Findings

We commence the unit root test of the differences of the log-transformed series for each of the thirty-five (35) paired life expectancy samples from the classical ADF test viewpoint. This is examined under three different ADF model structures – model with no regressors, model with constant only and model with constant and trend; with automatic lag selection option. We, however, are only interested in the model with the most negative significant t-statistics among the three model structures, and thus, interpret the result of the same. In line with the aforementioned criteria, the model with constant and trend seems to be mostly preferred except in the cases of LBY-BFA, EGY-NER, and TUN-NGA, where the model with constant only is preferred, and MOR-NER, TUN-BFA, TUN-MLI and TUN-SEN, where the model with no regressor is preferred. The classical ADF test (see Table 1) suggests that most of the paired differences have unit roots.

Table 1. Classical ADF tests

Differences	No regressor	Constant only	Constant with trend
ALG-BEN	0.6208[6]	-1.5726[6]	-2.4617[6]
ALG-BFA	-0.3450[6]	**-4.6970[3]**	**-4.8085[5]**
ALG-CIV	1.1854[7]	**-4.6255[5]**	**-5.1855[5]**
ALG-MLI	-1.7767[6]	0.2517[6]	-2.8899[9]
ALG-NER	-1.4155[3]	-1.7293[3]	-1.9979[3]
ALG-NGA	-0.6069[3]	**-3.8014[5]**	**-3.9154[5]**
ALG-SEN	-1.0990[6]	-1.2432[6]	**-4.5188[10]**

Table 1. Classical ADF tests (cont.)

Differences	No regressor	Constant only	Constant with trend
LBY-BEN	-1.9699[3]	-2.1341[3]	**-4.0889[3]**
LBY-BFA	-1.4873[3]	-1.8121[3]	-0.9751[10]
LBY-CIV	-0.4854[3]	-2.2175[3]	-2.2347[3]
LBY-MLI	**-2.0600[4]**	-0.0651[4]	-2.6865[4]
LBY-NER	**-2.5402[5]**	-2.6562[5]	-2.8803[5]
LBY-NGA	-1.1824[3]	-1.9195[3]	-1.5517[3]
LBY-SEN	-1.9024[4]	-0.3177[4]	**-4.3651[3]**
EGY-BEN	**-2.3632[6]**	-0.0756[6]	**-4.1869[3]**
EGY-BFA	-1.2552[6]	-0.1678[6]	**-6.0723[3]**
EGY-CIV	-0.1175[4]	-2.1008[4]	**-6.0627[3]**
EGY-MLI	**-2.0720[6]**	0.4590[6]	**-5.2433[3]**
EGY-NER	-0.7945[5]	**-3.5876[3]**	-2.3110[3]
EGY-NGA	0.3194[5]	**-4.1738[3]**	**-7.2086[3]**
EGY-SEN	**-2.1840[6]**	-2.5185[3]	**-5.5084[3]**
MOR-BEN	-0.2056[3]	-2.7751[3]	-3.0619[3]
MOR-BFA	-1.0178[6]	**-2.9752[3]**	**-3.9615[5]**
MOR-CIV	0.9926[6]	**-4.0205[3]**	**-5.7309[3]**
MOR-MLI	**-2.0190[6]**	-0.8639[6]	**-5.0506[3]**
MOR-NER	**-2.4056[3]**	-0.9971[3]	-2.1025[3]
MOR-NGA	-0.4574[3]	-2.8659[3]	**-4.0597[3]**
MOR-SEN	**-2.3161[6]**	**-3.0854[6]**	**-4.9771[5]**
TUN-BEN	-0.5022[3]	**-3.1105[3]**	-2.5215[3]
TUN-BFA	-0.8243[3]	-1.8103[3]	-0.1028[3]
TUN-CIV	-0.8385[3]	-2.2856[3]	-3.1636[3]
TUN-MLI	-1.7446[3]	-0.5202[3]	-0.8679[6]
TUN-NER	-1.8710[3]	**-3.4057[4]**	**-4.4992[4]**
TUN-NGA	-1.4505[3]	-2.3456[3]	-1.3866[3]
TUN-SEN	-1.3240[6]	0.2893[6]	-1.2203[6]

Note: In bold denotes rejection of the null hypothesis of unit root at 5% level.

Contrasting the classical ADF test results (Table 1) with the ADF test, where the lag specification is restricted to unity (see the second column in Table 2), we find that the stance of the latter differs markedly from the former, as only ALG-BFA was found to have a unit root. This suggests the dependence of the ADF unit root test on the lag specification and its sensitivity to the choice of lag, which limits the power of the conventional classical ADF test. In a bid to overcome this limitation, the FADF, ADF-SB, and FADF-SB tests are employed. The FADF test, which incorporates the ADF test in a Fourier framework, suggests stationarity of the log-transformed series in all cases except in ALG-BFA. However, both ADF-SB (which incorporates structural breaks in the ADF test framework) and FADF-SB (which accounts for structural breaks and

incorporates a Fourier function in the ADF testing framework) agree on the stationarity stance of the paired differences in all the cases considered. Except for the case of ALG-BFA, all four unit root tests suggest stationarity of all the paired differences. It appears that the incorporation of structural breaks in the unit root testing framework further improves the power of tests for both ADF and FADF tests.

Table 2. Fourier ADF Break tests

Differences	ADF	FADF		ADF-SB			FADF-SB			
	t stat.	*k*	*t* stat.	T_B	λ_B	*t* stat.	T_B	λ_B	*k*	*t* stat.
ALG-BEN	**-4.451**	2	**-5.000**	2002	0.75	**-5.495**	1968	0.16	1	**-7.014**
ALG-BFA	-2.532	1	-4.278	2015	0.98	**-4.606**	2015	0.98	1	**-5.206**
ALG-CIV	**-4.646**	1	**-5.375**	2011	0.91	**-5.999**	2011	0.91	1	**-6.370**
ALG-MLI	**-5.156**	2	**-5.718**	1992	0.58	**-7.124**	1992	0.58	2	**-7.957**
ALG-NER	**-4.725**	2	**-5.063**	2006	0.82	**-5.545**	1972	0.23	1	**-7.853**
ALG-NGA	**-4.869**	1	**-5.845**	1987	0.49	**-7.160**	1987	0.49	1	**-7.228**
ALG-SEN	**-4.480**	2	**-4.968**	2001	0.74	**-5.617**	2001	0.74	2	**-5.825**
LBY-BEN	**-4.043**	1	**-5.468**	2014	0.96	**-6.397**	2014	0.96	1	**-6.760**
LBY-BFA	**-4.527**	2	**-4.776**	1996	0.65	**-7.368**	1996	0.65	2	**-7.230**
LBY-CIV	**-4.874**	2	**-5.536**	1991	0.56	**-6.863**	1991	0.56	2	**-7.146**
LBY-MLI	**-4.741**	2	**-5.131**	2005	0.81	**-5.507**	1971	0.21	1	**-7.310**
LBY-NER	**-4.872**	1	**-5.841**	1986	0.47	**-6.642**	1986	0.47	1	**-6.857**
LBY-NGA	**-4.913**	2	**-5.578**	2000	0.72	**-6.328**	2000	0.72	2	**-6.658**
LBY-SEN	**-4.569**	1	**-5.989**	2012	0.93	**-6.713**	2012	0.93	1	**-7.041**
EGY-BEN	**-4.842**	2	**-5.270**	1995	0.63	**-7.709**	1995	0.63	2	**-7.818**
EGY-BFA	**-4.977**	1	**-5.242**	2009	0.88	**-6.258**	1975	0.28	1	**-6.670**
EGY-CIV	**-5.059**	2	**-5.442**	2004	0.79	**-5.876**	1970	0.19	1	**-7.757**
EGY-MLI	**-5.451**	1	**-6.086**	1985	0.46	**-6.622**	1988	0.51	2	**-6.832**
EGY-NER	**-4.697**	2	**-5.393**	1999	0.70	**-6.180**	1999	0.70	2	**-6.508**
EGY-NGA	**-4.872**	1	**-6.120**	2011	0.91	**-6.350**	1979	0.35	1	**-6.616**
EGY-SEN	**-4.540**	2	**-4.923**	1994	0.61	**-7.135**	1994	0.61	1	**-7.420**
MOR-BEN	**-4.548**	1	**-4.823**	2008	0.86	**-5.599**	1974	0.26	1	**-6.789**
MOR-BFA	**-5.043**	2	**-6.083**	1989	0.53	**-6.735**	1989	0.53	1	**-6.879**
MOR-CIV	**-4.828**	1	**-6.098**	1987	0.49	**-6.088**	1987	0.49	2	**-6.655**
MOR-MLI	**-4.679**	2	**-5.259**	1998	0.68	**-6.685**	1998	0.68	2	**-6.896**
MOR-NER	**-4.579**	1	**-5.587**	2005	0.81	**-5.840**	2012	0.93	1	**-6.380**
MOR-NGA	**-4.971**	2	**-5.510**	1993	0.60	**-7.157**	1993	0.60	2	**-8.102**
MOR-SEN	**-4.487**	1	**-4.759**	2007	0.84	**-5.575**	1973	0.25	1	**-6.879**
TUN-BEN	**-4.999**	1	**-6.082**	1988	0.51	**-7.071**	1988	0.51	1	**-7.475**
TUN-BFA	**-4.621**	2	**-5.198**	2002	0.75	**-5.669**	1968	0.16	1	**-7.143**
TUN-CIV	**-4.691**	2	**-5.065**	1997	0.67	**-6.710**	1997	0.67	1	**-6.937**
TUN-MLI	**-4.600**	1	**-5.649**	2009	0.88	**-6.029**	2011	0.91	1	**-6.615**
TUN-NER	**-5.408**	2	**-5.806**	1992	0.58	**-7.666**	1992	0.58	2	**-8.388**
TUN-NGA	**-4.691**	1	**-4.933**	2006	0.82	**-6.090**	1972	0.23	1	**-6.985**
TUN-SEN	**-4.811**	1	**-5.874**	1987	0.49	**-6.766**	1987	0.49	1	**-7.195**

In bold denotes rejection of the null hypothesis of unit root at a 5% level. For details about this test as well as critical regions, see Furuoka (2017).

Having shown that the paired differences in all cases considered are stationary, we further subject the contending unit root tests to some reliability tests, as a robustness check. This is by way of ascertaining the unit test that would be most appropriate for determining the stationarity stance of the paired differences. In this regard, we compare the performance of five pairs (restricted and unrestricted model constructs) of unit root tests using the F-statistics (see Table 3). $F_{(FADF,\ ADF)}$ tests whether the improvement of the FADF (unrestricted model construct) test over the ADF (restricted model construct) test is significant. We find statistically significant improvement of the FADF over the ADF unit root regression in only two (2) of the thirty-five (35) considered cases, and these include ALG_BFA and LBY-SEN. This is indicative of the relative similarity in the decision reached by both FADF and ADF unit root tests. On the other hand, $F_{(ADF-SB,\ ADF)}$ and $F_{(FADF-SB,\ ADF)}$ reveal the statistically significant improvement of ADF-SB and FADF-SB tests, respectively, over the ADF test, in all cases for the former and thirty-two (32) cases for the latter. Similarly, the FADF-SB unit root test is observed to be more reliable in comparison with FADF and ADF-SB, as it outperformed both in thirty-two (32) and thirty-one (31) cases, respectively. Both ADF-SB and FADF-SB are found to be more reliable than the ADF and FADF tests. It is evident here that the decision on the stationarity, or otherwise, of the paired differences based on the FADF-SB unit root test is more reliable (see Table 3). This outperformance over other conventional unit root tests is again upheld (see Furuoka, 2017; Yaya, Ogbonna and Mudida, 2019; among others), as it hinges on accounting for both nonlinearity and plausible presence of structural breaks. Adopting a battery of reliable unit root testing frameworks would also enhance researchers' decisions.

Table 3. Robustness checks

Differences	$F_{(FADF,\ ADF)}$	$F_{(ADF-SB,\ ADF)}$	$F_{(FADF-SB,\ ADF)}$	$F_{(FADF-SB,\ FADF)}$	$F_{(FADF-SB,\ ADF-S}$
ALG-BEN	2.557	**9.703**	**6.642**	**9.840**	**12.948**
ALG-BFA	**7.296**	**17.936**	**10.923**	**11.536**	2.709
ALG-CIV	3.039	**16.122**	**9.593**	**14.535**	**18.941**
ALG-MLI	2.583	**24.235**	**16.232**	**27.224**	**31.570**
ALG-NER	1.673	**8.937**	**8.600**	**14.632**	**16.406**
ALG-NGA	4.173	**23.622**	**13.990**	**20.600**	**26.858**
ALG-SEN	2.424	**10.279**	**6.089**	**8.993**	**9.936**
LBY-BEN	6.008	**12.488**	**7.603**	**7.634**	0.432
LBY-BFA	1.654	**25.071**	**12.816**	**22.578**	**25.490**
LBY-CIV	2.928	**22.370**	**13.815**	**22.260**	**27.107**
LBY-MLI	1.906	**7.308**	**6.784**	**10.921**	**12.679**
LBY-NER	4.119	**22.621**	**13.774**	**20.310**	**26.291**
LBY-NGA	2.986	**12.286**	**7.341**	**10.574**	**11.089**
LBY-SEN	**6.293**	**9.479**	**6.119**	**4.967**	**12.095**

Table 3. Robustness checks (cont.)

Differences	$F_{(FADF,\ ADF)}$	$F_{(ADF-SB,\ ADF)}$	$F_{(FADF-SB,\ ADF)}$	$F_{(FADF-SB,\ FADF)}$	$F_{(FADF-SB,\ ADF-S}$
EGY-BEN	2.555	**25.177**	**13.849**	**22.944**	**26.457**
EGY-BFA	1.354	**15.395**	4.469	**7.253**	**8.917**
EGY-CIV	1.878	**8.517**	**7.420**	**12.142**	**14.819**
EGY-MLI	3.017	**22.036**	3.561	3.777	6.931
EGY-NER	3.036	**13.503**	**7.943**	**11.589**	**15.294**
EGY-NGA	5.575	**6.366**	4.499	2.987	**8.900**
EGY-SEN	1.828	**26.046**	**16.761**	**29.642**	**32.872**
MOR-BEN	1.291	**13.936**	**5.821**	**9.900**	**11.622**
MOR-BFA	4.386	**16.212**	**9.956**	**13.395**	**16.305**
MOR-CIV	5.538	**5.522**	4.309	2.709	5.271
MOR-MLI	2.522	**18.042**	**10.000**	**15.995**	**19.826**
MOR-NER	4.293	**5.299**	**8.283**	**10.649**	**16.449**
MOR-NGA	2.528	**29.767**	**19.902**	**34.004**	**39.257**
MOR-SEN	1.357	**13.372**	**6.510**	**11.125**	**13.008**
TUN-BEN	4.671	**19.541**	**12.112**	**16.680**	**20.311**
TUN-BFA	2.663	**9.690**	**6.512**	**9.475**	**12.865**
TUN-CIV	1.884	**16.820**	**10.960**	**18.726**	**21.255**
TUN-MLI	4.300	**6.060**	**9.386**	**12.529**	**18.181**
TUN-NER	1.943	**28.841**	**18.389**	**32.440**	**35.624**
TUN-NGA	1.255	**16.276**	**6.246**	**10.757**	**11.711**
TUN-SEN	4.502	**20.916**	**13.416**	**19.129**	**25.502**

Note: In bold denotes the significance of F tests at a 5% level. For more details about these tests, see Furuoka (2017).

Consequent upon the reliability of the decision reached using FADF-SB and ADF-SB unit root tests, the stationarity stance of the paired differences in life expectancy between African countries is upheld. By implication, the paired differences exhibit mean-reverting characteristics and the absence of persistence. Consequently, the life expectancy of West African countries have a tendency to converge to those of the North African countries with time.

4. Conclusion

This study set out to examine the stationarity; and by extension, the convergence; of life expectancy in any given pair of a West African country and a Northern counterpart. The study draws from the concept of convergence and catching up theory as proposed by Bernard and Durlauf (1996) and Greasley and Oxley (1997). Consequently, a total of five (5) North African countries and seven (7) West African countries, which amounts to a total of thirty-five possible pairs, were considered. While the North African region is known to have a higher life expectancy, their West counterpart is observed to have a lower life expectancy. Therefore, the interest here is to examine the possibility of the life expectancy of the West African countries to

converge to those of the North African countries based on the current evolution of the life expectancy time series dynamics. The difference between the log-transformed life expectancy rates of the country pairs is subsequently subjected to a battery of unit root testing frameworks.

In achieving this, we consider a battery of unit root tests, which is good practice as suggested by several extant literature (Yaya, Ogbonna and Atoi, 2019 and Yaya, Ogbonna, and Mudida, 2019). This battery of unit testing frameworks include the classical ADF test, ADF with structural breaks (ADF-SB) and the Fourier-based ADF tests – FADF and FADF-SB. These were all applied on the obtained differences, which resulted in some interesting results. We find the classical ADF to reject the stance of a unit root in fewer cases, compared to the other unit root testing frameworks. A notable feature is the significant influence of accounting for structural breaks, especially, whenever they exist. We find FADF-BP and ADF-SB to consistently out-perform the classical ADF and FADF unit root tests that failed to account for structural breaks. However, we do not jump to draw rash decisions on these performances, as we further subject the results to some reliability tests. This draws from the F-statistics employed to compare the contending unit root testing model framework. Evidently, we find the FADF-SB to be most preferred, given its simultaneous incorporation of a Fourier function and accounting for structural breaks. Convincingly, we state here that the FADF-SB unit root testing framework is the most reliable in testing for the stationarity stance of the difference in life expectancy of any pair of countries.

Imperatively, the life expectancy of West African countries is most likely to converge to and catch up with the life expectancy of the North African countries. This stance is key for policymakers, as they make policy decisions that affect the lives of people in these countries. The information herein contained could spur governments of West African nations to pay greater attention to their economic growth and development, political stability, and security of life and property of their citizens. Also, the well-being of their citizens must be given utmost priority, through the provision of functional and efficient health and education systems, and the provision and maintenance of social and infrastructural facilities. Furthermore, public enlightenment campaigns on the dangers of risky health behaviours should be intensified and sustained. However, higher life expectancy may also have its own health implications, which need to be taken into consideration. These may include an increased risk of age-related diseases among the elderly, dementia, and physical disabilities. Policies should, therefore, be put in place to increase the quality of life of the elderly, so that achieving longevity does not become a burden on the health of the citizens.

References

ABORİGİNAL AND TORRES STRAİT ISLANDER HEALTH PERFORMANCE FRAMEWORK (HPF), (2012). Tier 1- Life expectancy and wellbeing, Available Online: https://www.health.gov.au, Accessed on March 2, 2019.

BARTHOLD JONES, J.A., LENART, A. and BAUDISCH, A., (2018). The complexity of the relationship between life expectancy and overlap of lifespans, PLoS ONE 13(7), e0197985, https://doi.org/10.1371/journal.pone.0197985.

BERNARD, A. B., DURLAUF, S. N., (1996). Interpreting tests of the convergence hypothesis. Journal of Econometrics, 71, pp. 161–173.

CUÑADO, J., PÉREZ DE GRACIA, F., (2006). Real convergence in Africa in the second-half of the 20th century. Journal of Economics and Business, 58(2), pp. 153–167, doi:10.1016/j.jeconbus.2005.07.002.

DICKEY, D. A., FULLER, W. A., (1979). Distribution of the Estimators for Autoregressive Time Series with a Unit Root. Journal of the American Statistical Association, 74, pp. 427–431.

ENDERS, W., LEE, J., (2012a). A unit root test using a Fourier series to approximate smooth breaks. Oxford Bulleting of Economics and Statistics, 74, pp. 574–599.

ENDERS, W., LEE, J., (2012b). The flexible Fourier form and Dickey-Fuller-type unit root tests. Economic Letters, 117, pp. 196–199.

FURUOKA, F., (2017). A new approach to testing unemployment hysteresis. Empirical Economics, 53(3), pp. 1253–1280.

GREASLY, D., OXLEY, S., (1997). Time-series based tests of the convergence hypothesis: Some positive results. Economic Letters, 56, pp. 143–147.

KONTIS, V., BENNETT, J. E., MATHERS, C. D., LI, G., FOREMAN, K. and EZZATI, M., (2017). Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble, Lancet, 389(10076), pp. 1323–1335, DOI: 10.1016/S0140-6736(16)32381-9.

MONDAL, N. I., SHITAN, M., (2013). Impact of Socio-Health Factors on Life Expectancy in the Low and Lower Middle Income Countries. Iranian Journal of Public Health, 42(12), pp. 1354–1362.

NMETH, L., MISSOV, T. I., (2018). Adequate life-expectancy reconstruction for adult human mortality data, PLoS ONE 13(6), e0198485, https://doi.org/10.1371/journal.pone.0198485.

NKALU, C. N., EDEME, R. K., (2019). Environmental Hazards and Life Expectancy in Africa: Evidence from GARCH Model, SAGE Open pp. 1–8, DOI: 10.1177/2158244019830500.

ORTIZ-OSPINA, ESTEBAN, (2017). "Life expectancy"- What does this actually mean? Our World in Data, Available Online: https://ourworldindata.org, Accessed on March 2, 2019.

OTEKUNRIN, O. A., OTEKUNRIN, O. A., MOMOH, S. and AYINDE, I. A., (2019a). How far has Africa gone in achieving the zero hunger target? Evidence from Nigeria. Global Food Security 22, pp. 1–12, https://doi.org/10.1016/j.gfs.2019.08.001.

OTEKUNRIN, O. A., MOMOH, S., AYINDE, I. A. and OTEKUNRIN, O. A., (2019b). How far has Africa gone in achieving sustainable development goals? Exploring African dataset, Data in Brief 27:104647, https://doi.org/10.1016/j.gfs.2019.08.001.

PASCARIU, M. D., CANUDAS-ROMO, V. and VAUPEL J. W., (2018). The double-gap life expectancy forecasting model. Insurance: Mathematics and Economics 78, pp. 339–350, https://doi.org/10.1016/j.insmatheco.2017.09.011.

PERRON, P., VOGELSANG, T. J., (1992). Non-stationarity and level shifts with an application to purchasing power parity. Journal of Business, Economics and Statistics, 10(3), pp. 301–320.

STATISTICA, (2018). Life Expectancy in Europe, Available Online: www.statista.com, Accessed March 2, 2019.

STATISTICA, (2018). Life Expectancy in Africa, Available Online: www.statista.com, Accessed June 22, 2020.

STEVENS, E. R., ZHOU, Q., TAKSLER, G. B., NUCIFORA, K. A., GOUREVITCH, M. and BRAITHWAITE, R. S., (2019). An alternative mathematical modeling approach to estimating a reference life expectancy. MDM Policy & Practice, pp. 1–12, DOI: 10.1177/2381468318814769.

UCHENDU, F. N., (2018). Hunger influenced life expectancy in war-torn Sub-Saharan African countries. Journal of Health, Population and Nutrition 37, 11, https://doi.org/10.1186/s41043-018-0143-3.

UNITED NATIONS. DEPARTMENT OF ECONOMIC AND SOCIAL AFFAIRS, POPULATION DIVISION, (2019). World Mortality 2019: Data Booklet (ST/ESA/SER.A/436), https://www.un.org/en/development/desa/population/publications/pdf/mortality/WMR2019/WorldMortality2019DataBooklet.pdf.

VAN BAAL, P., PETERS, F., MACKENBACH, J., and NUSSELDER, W., (2016). Forecasting differences in life expectancy by education. Population Studies, 70(2), pp. 201–216, DOI: 10.1080/00324728.2016.1159718

WIYSONGE, C. S. (2018). People are living longer in Africa but the rise of lifestyle diseases threatens progress. Quartz Africa, https://qz.com/africa/1475911/, Accessed June 22, 2020.

WORLD POPULATION REVIEW: LIFE EXPECTANCY BY COUNTRY 2017. Available Online: worldpopulationreview.com, Accessed on March 2, 2019.

YAYA, O. S., OGBONNA, A. E. and ATOI, N. V., (2019). Are inflation rates in OECD countries actually stationary during 2011–2018? Evidence based on Fourier Nonlinear Unit root tests with Break, The Empirical Economics Review, 9(4), pp. 309–325.

YAYA, O.S., OGBONNA, A.E. and MUDIDA, R., (2019). Hysteresis of Unemployment rate in Africa: New Findings from Fourier ADF test. Quality and Quantity International Journal of Methodology, 53(6), pp 2781–2795.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1, pp. 89–114, DOI 10.21307/stattrans-2021-005
Received – 01.05.2019; accepted – 16.06.2020

A latent class analysis on the usage of mobile phones among management students

Sunil Kumar[1], **Apurba Vishal Dabgotra**[2]

ABSTRACT

In the past few years, wireless devices, including pocket PCs, pagers, mobile phones, etc, have gained popularity among a variety of users across the world and the use of mobile phones in particular, has increased significantly in many parts of the world, especially in India. Cell phones are now the most popular form of electronic communication and constitute an integral part of adolescents' daily lives, as is the case for the majority of mobile phone users. In fact, mobile phones have turned from a technological tool to a social tool. Therefore, the influence of cell phones on young people needs to be thoroughly examined. In this paper, we explore the attitude of young adults towards cell phones and identify the hidden classes of respondents according to the patterns of mobile phone use. The Latent Class Analysis (LCA) serves as a tool to detect any peculiarities, including those gender-based. LCA measures the value of an unknown latent variable on the basis of the respondents' answers to various indicator variables; for this reason, a proper selection of indicators is of great importance here. In this work, we propose a method of selecting the most useful variables for an LCA-based detection of group structures from within the examined data. We apply a greedy search algorithm, where during each phase the models are compared through an approximation to their Bayes factor. The method is applied in the process of selecting variables related to mobile phone usage which are most useful for the clustering of respondents into different classes. The findings demonstrate that young people display various feelings and attitudes toward cell phone usage.

Key words: backward greedy search algorithm (BGSA), latent class analysis (LCA), AIC, BIC.

1. Introduction

With the advancement in technology, the number of users of mobile phones have increased rapidly in the entire world. It has now become a crucial part of majority of lives of youth. This is because there are so many applications available on cell phones these days like internet access, sending e-mails, games, access to social networking sites like Facebook, listening to music, playing radio, reading books, dictionary and so on. Therefore, youth use their mobile phones more often for their free time. It was stated in the report of TRAI (Jan 2018), that India has about 1.012 billion mobile phone connections, which makes India's telecommunication market, the world's second largest in terms of number of

[1]Department of Statistics, University of Jammu, J&K, India. E-mail: sunilbhougal06@gmail.com. ORCID: http://orchid.org/0000-0003-0249-8415.
[2]Department of Statistics, University of Jammu, J&K, India. E-mail: apurvavdabgotra@gmail.com. ORCID: https://orcid.org/0000-0002-8056-7239.

wireless connections after China. With youth population constituting half of the total population, India has become a fine breeding ground for highest cell connections. There exists a need to study and analyze the influence of cell phones on youth, because cell phones are responsible for modulating the thought process of any person especially that of the youngsters. In this paper, LCA is used for exploring the attitude of youth towards cell phones; to identify the hidden classes of respondents according to their mobile usage pattern and to arrive at the peculiarities in the cell phones usage, gender-wise, if any. The identification of most relevant variables related to mobile phone usage may aid in the evaluation of its user's attitude towards it and its impact on their lives in a way that usefully informs the role and vitality of cell phones in lives of today's youth.

LCA provides a tool for clustering and classification of individuals given the response pattern of a respondent to various questions or items of a questionnaire in a qualitative research. It is used for modelling and explaining the relationships between manifest or observed variable (may be dichotomous or polytomous) with respect to some unobserved or latent variables (may be dichotomous or polytomous) on the basis of data obtained in various kinds of surveys. LCA identifies unobservable (latent) subgroups within a population based on individuals' responses to different categorical observed variable. In addition to the above, LCA can also be used for the estimation of unknown parameters. The parameters of interests in any typical problem of latent class analysis are the unobserved proportion or size of the latent classes and the conditional item-response probabilities given the membership in a latent class. LCA will also provide estimate of the probabilities of respondents being misclassified by the questions. LCA was introduced in 1950 by Paul F. Lazarsfeld as a way of formulating latent attitudinal variables. Lazarsfeld performed LCA for building typologies (or clustering) based on dichotomous observed variables.

Then, Goodman (1974) proposed the estimation of LC model parameters using the maximum likelihood approach. Dempster et al. (1977) provided maximum likelihood estimation in the case of the observed and missing data involved in LCA. Haberman (1979) established the relationship between LC models and log-linear models for unknown cell counts frequency tables. Formann (1984, 1992) constructed a linear logistic LCA for dichotomous and polytomous variables. Mooijaart (1992) presented the application of EM algorithm in LCA in a very detailed way. Vermunt (2010) proposed the inclusion of the covariates in the LC models, which make it possible to predict the LC membership probabilities by covariates through a logistic link. Biemer (2010) discussed the use of LCA as a survey error evaluation technique. Many theoretical developments and applications of LCA have been proposed in the recent past years, to encourage the research studies in LCA, example: Lanza (2013), Boduszek et. al (2014), Kumar (2015, 2016, 2017), Porcu (2017), Petersen (2019) and Sapounidis (2019).

Since LCA measures the value of an unknown latent variable given the responses made by the respondents to the various indicator variables, so proper selection of relevant indicator variables out of the set of all possible manifest variables is required in order to get efficient estimates of the unknown parameters. We have used, Raftery and Dean (2010) greedy search algorithm for the selection of indicator variables. This algorithm is used for checking single variable for inclusion into/exclusion from the set of selected clustering variables. The "greedy" search checks the inclusion of each single variable not currently

selected into the current set of selected clustering variables. The variable that has highest evidence of inclusion is proposed and, if its clustering evidence is stronger than the evidence against clustering it is included. At every exclusion step the "greedy" search option checks the exclusion of each single variable in the currently selected set of clustering variables and proposes the variable that has lowest evidence of clustering. The proposed variable is removed if its evidence of clustering is weaker than its evidence against clustering.

Through greedy search algorithm, we have identified the indicators variables which measures the latent variables. Following Baumgartner and Steenkamp (2006), the most promising way of accounting for extreme response bias is the inclusion of statistical techniques in the data analysis. Based on qualitative nature of the data it is not possible to use the classical least squares theory. LCA calculate conditional probabilities using Bayesian methodology.

To understand the application of methodological development, we have considered a study on the behaviour of mobile phone users (especially, young users). The paper is ordered as follows: Section 2 describes the framework of the greedy search algorithm followed by LCA model. Section 3 presents the data description and section 4 provides the analysis of mobile users data using the software application LCAvarsel (Fop and Murphy, 2017) and poLCA (Linzer and Lewis, 2011), the most complete and user-friendly package for the variable selection and the estimation of the latent class models. Finally, in section 5, we summarize our findings and recognize the probable areas of further research.

2. Methodology

2.1. Variable Selection

The problem of selecting the relevant clustering variables, in LCA can be modified as a model selection problem. Different models are specified by the role allotted to the variables in relevance to their relationship with the clustering variable X. Then, these models are compared by means of a model selection criterion and relevant clustering variables are, thus chosen accordingly in order to form the best model. The framework was introduced by the work of Raftery and Dean (2006). The authors propose a procedure where, Y (the set of all manifest variables) is partitioned into the following subsets of variables:

1. Y^C, the set of current clustering variables;
2. Y^P, the variable(s) proposed to be added or removed from the set of clustering variables;
3. Y^{NC}, the set of other variables not relevant for clustering.

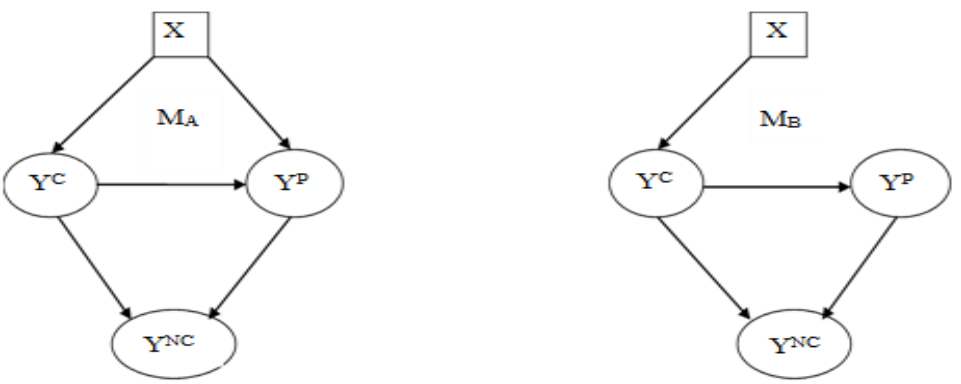
Given this partition of Y and the (unknown) clustering membership X, we can modify the question of effectiveness of Y^C for clustering as a question of model selection. The question, thus becomes to choose one of two different models, i.e. M_A which assumes that Y^P is useful for clustering and M_B which assumes that it is not. In model M_B, the set of variable(s) proposed to be added or removed from the set of clustering variables Y^P is conditionally independent of the cluster memberships X given the variables Y^C is already in the model. In model M_A, this is not the case. In both models, the set of other variables

considered Y^{NC} is conditionally independent of cluster membership X given Y^C and Y^P, but may be associated with Y^C and Y^P. Then, the decision for inclusion or exclusion of the variable belonging to Y^P is taken by comparing the models (Figure 1):

$$\begin{aligned} M_A : P(Y|X) &= P(Y^C, Y^P, Y^{NC}|X) \\ &= P(Y^C, Y^P|X)\, P(Y^{NC}|Y^C, Y^P), \qquad (1) \\ M_B : P(Y|X) &= P(Y^C, Y^P, Y^{NC}|X) \\ &= P(Y^C|X)\, P(Y^P|Y^C)\, P(Y^{NC}|Y^C, Y^P), \qquad (2) \end{aligned}$$

where , X is the (unobserved) set of cluster memberships. Model M_A implies that Y^P does provide information about clustering membership, beyond that given just by Y^C. Model M_B specifies that, given Y^C, Y^P is independent of the cluster memberships (defined by the unobserved variables X), i.e., Y^P gives no further information about the clustering. In model M_A , Y^P is useful for clustering and the joint distribution $P(Y^C, Y^P|X)$ corresponds to a Gaussian mixture distribution; on the other hand, M_B states that Y^P does not depend on the clustering variable X and the conditional distribution $P(Y^P|Y^C)$ corresponds to a linear regression.

Figure 1: Graphical Representation of Models M_A and M_B for Variable Selection.



An important feature of the framework formulation is that in M_B the irrelevant variables are not required to be independent of the clustering variables. This criterion allows to discard redundant variables related to the clustering ones but not to the clustering itself. Models M_A and M_B are compared via an approximation to the Bayes factor. The Bayes factor, B_{AB}, for M_A against M_B based on the data Y is given by

$$B_{AB} = \frac{P(Y|M_A)}{P(Y|M_B)}, \qquad (3)$$

where , $P(Y|M_i)$ is the integrated likelihood of model M_i (i =A,B), namely

$$P(Y|M_i) = \int P(Y|\Theta i, M_i) P(\Theta_i|M_i) d\Theta_i,$$

where , Θ_i is the vector-valued parameter of model M_i, and $P(\Theta_k|M_i)$ is its prior distribution (Kass and Raftery, 1995).

Let us now consider the integrated likelihood of model M_A, $P(Y|M_A) = P(Y^C, Y^P, Y^N C|M_A)$. From (1), the model M_A is specified by two probability distributions: the latent class model that specifies $P(Y^C, Y^P|\Theta_1, M_A)$, and the distribution $P(Y^{NC}|Y^C, Y^P, \Theta_1, M_A)$. We denote the parameter vectors that specify these two probability distributions by Θ_{11}, Θ_{12}, and we assume that their prior distributions are independent. Then, the integrated likelihood factors as follows:

$$P(Y|M_A) = P(Y^C, Y^P, Y^{NC}|M_A)$$
$$P(Y|M_A) = P(Y^C, Y^P|M_A)P(Y^{NC}|Y^C, Y^P, M_A),$$

where , $P(Y^{NC}|Y^P, Y^C, M_A) = \int P(Y^{NC}|Y^P, Y^C, \Theta_{12}, M_A)P(\Theta_{12}|M_A)d\Theta_{12}$
and $P(Y^C, Y^P|M_A) = \int P(Y^C, Y^P|\Theta_{11}, M_A)P(\Theta_{11}|M_A)d\Theta_{11}$.

Similarly, we obtain

$$P(Y|M_B) = P(Y^C|M_B)P(Y^P|Y^C, M_B)P(Y^{NC}|Y^C, Y^P, M_B).$$

The prior distribution of the parameter, Θ_{12},is assumed to be the same under M_A as under M_B. It follows that

$$P(Y^{NC}|Y^P, Y^C, M_A) = P(Y^{NC}|Y^P, Y^C, M_B).$$

We thus have

$$B_{AB} = \frac{P(Y^C, Y^P|M_A)}{P(Y^C|M_B)P(Y^P|Y^C, M_B)} \tag{4}$$

which has been greatly simplified by the cancellation of the factors involving the potentially high-dimensional Y^{NC}. The integrated likelihoods in (4) are still hard to evaluate analytically, so we approximate them using the BIC approximation given as:

$$BIC(Y|M_i) = 2 \times log(max.lnL) - (p) \times log(n)$$

where , i= A, B

max. lnL : maximum log- likelihood

p : number of parameters

n : number of observations

The above equation provides the BIC approximates for model M_A & model M_B and thus leading to the following criterion:

$$BIC_{diff} = BIC_A - BIC_B.$$

The clustering variables are selected using a stepwise algorithm of greedy search algorithm. At each stage it searches for the variable to add that most improves the clustering as measured by BIC, and then assesses whether one of the current clustering variables can be dropped. At each stage, the best combination of the number of groups and clustering model is chosen. For performing greedy search algorithm, we first have to choose the value of no. of classes for each of the latent variable, r=1,2,...,R; so that , our model is identifiable. For a given number of variables, not all the models specified by assigning different values to r

are identifiable. In fact, a necessary (though not sufficient) condition to the identifiability of a model with r latent classes is

$$\prod_{m=1}^{M} C_m > \left(\sum_{m=1}^{M} C_m - M + 1 \right) r, \tag{5}$$

with C_m the number of categories taken by variable X_m [Goodman (1974)]. Thus when selecting the number of classes, hereafter we will consider values of r for which this identifiability condition holds.

Set R (max. r), the maximum number of clusters to be considered for the data. Make sure that this number is identifiable for the data, i.e. R , the maximum number of latent classes should satisfy the identifiability condition in (5) for the set of variables currently taken into consideration in fitting the LCA model. If a latent class model on the set of all variables is identifiable for $R > 1$, we then estimate the model. For each category of each variable, we calculate the variance of its probability across groups. Then, for each variable, we add up these variances and rank the variables according to that sum. The reason behind this, is that the variables with high values of this sum have high between-group variation in probability, and hence they may be more useful for clustering. Given this ranking we choose the top k^* variables, where , k^* is the smallest number of variables that allow a latent class model with $R > 1$ to be identified. This will provide the starting Y^C. The other variables can be left in their ordering based on variability for future order of their inclusion in the algorithm. In fitting the LCA model we perform multiple runs with random starting values. Also, in this case the aim is to allows the search for the global maximum of the log-likelihood rather than a local one; then the model with the greatest log-likelihood is retained. The following are the inclusion and exclusion steps of a new variable:

Inclusion Step: We look at each variable in Y^{NC} as to be a new variable under consideration for inclusion into Y^P and the variable that has the highest evidence of inclusion is proposed. Then, calculate the difference in BIC for models M_A and M_B given the current Y^C. If the variable's BIC difference is:

– below 0, do not include the variable in Y^C and remove variable from Y^{NC};

– above 0, include variable in Y^C and stop inclusion step.

If we reach the end of the list of variables in Y^{NC}, the inclusion step is stopped.

Exclusion Step: We look at each variable in Y^C as to be a new variable under consideration for inclusion into Y^P (with the remaining variables in Y^C not including current Y^P now defined as Y^C in M_A and M_B) and the variable that has the lowest evidence of being in the set is proposed. Calculate the difference in BIC for models M_A and M_B. If the variable's BIC difference is:

– below 0, remove the variable from (the original) Y^C and place it under Y^{NC} and stop the exclusion step;

– above 0, do not remove the variable from (the original) Y^C.

If we reach the end of the list of variables in Y^{NC} the exclusion step is stopped.

If Y^C remains the same after consecutive inclusion and exclusion steps the greedy search algorithm stops because it has converged.

2.2. Latent Class Analysis

After identifying the variables for each latent variables, we next perform LCA for exploring the attitude of youth towards cell phones and to categories the respondents according to the pattern of their mobile phone usage. LCA has been used as a multivariate statistical tool for the study. LCA models comprises two types of probabilities which include

– the probability indicating the likelihood of a response by respondents in each of the classes and

– the probability representing the latent class size or the proportion of individuals who are members of a particular latent class.

The former one represents the probability of a particular responses to a manifest variable, conditioned on latent class membership and can be interpreted as factor loading for Factor Analysis, in which both the observed or latent variables are continuous. LCA provides a clustering of individuals in a population, based on the response patterns of individuals to the different observed variables.

The assumptions of the standard LC model (Biemer,2010) are as follows:

1. The sample can be treated as if it was a simple random sample without replacement from an infinite population, i.e. data is sampled without replacement from a large population units using SRS (Simple Random Sampling).

2. The indicators are locally independent within a latent class, means all the indicator variables have nothing in common except latent variable, i.e. after accounting for latent variable X, there is no association between indicator variables.

3. The response probabilities are homogeneous, i.e. the probabilities of selecting any two units (individuals) from the population are same.

4. The indicator variables are univocal, i.e. the indicator variables can measure one and only one latent variable.

Following the notation used by Linzer and Lewis (2011), suppose we have J polytomous categorical manifest variables (the observed variable) each of which contain K_j possible outcomes, for individuals $i = 1,2,3,...,N$. Let Y_{ijk} be the observed values of the J manifest variables such that

$$\left\{ \begin{array}{ll} Y_{ijk} = 1: & \text{if } i^{th} \text{ respondent give the } k^{th} \text{ response to the } j^{th} \text{ variable} \\ Y_{ijk} = 0: & \text{otherwise} \end{array} \right\}$$

where, j=1,2,...,J and k=1,2,...,K_j.

The LC models approximates the observed joint distribution of the manifest variables as the weighted sum of a finite number, R, of constituent cross-classification tables. Let π_{jrk} denote the cross-conditional probability that an observation in class r=1,2,...,R produces the k^{th} outcome on the j^{th} variable with

$$\sum_{k=1}^{K_j} \pi_{jrk} = 1.$$

Let p_r be the prior probabilities of latent class membership, as they represent the unconditional probability that an individual will belong to each class before taking into account

the responses Y_{ijk} provided on the manifest variables. The probability that an individual i in class r produces a particular set of J outcomes on the manifest variables, assuming conditional independence of the outcomes Y given class membership, is the product

$$f(Y_i;\pi_r) = \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jrk})^{Y_{ijk}}, \tag{6}$$

The probability density function across all classes is the weighted sum

$$f(Y_i|\pi,p) = \sum_{r=1}^{R} f(Y_i;\pi_r) = \sum_{r=1}^{R} P_r \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jrk})^{Y_{ijk}}. \tag{7}$$

The parameters P_r and π_{jrk} are estimated by the latent class model.

The unknown parameters of the LC models can be estimated by maximizing the Log likelihood function with respect to p_r and π_{jrk}, using the expectation-maximization (EM) algorithm (Dempster et al. (1977), McLachlan and Peel (2000) and Linzer and Lewis (2011)) some other algorithms can also be considered like a Newton-Raphson algorithm or a hybrid form of these two algorithms (McLachlan and Krishnan, 2008). In any case the algorithm is initialized through a set of randomly generated starting values and there is no guarantee of reaching the global maximum. For this reason, it is usually a good practice to run the procedure a number of times and select the best solution [Bartholomew et al. (2011)].
Given estimates $\hat{P}_r$ and $\hat{\pi}_{jrk}$ of P_r and π_{jrk} respectively, the posterior probability that each individual belongs to each class, conditional on the observed values of the manifest variables, is calculated by

$$\hat{P}(r_i|Y_i) = \frac{\hat{p}_r f(Y_i;\hat{\pi}_r)}{\sum_{q=1}^{R}\hat{p}_q f(Y_i;\hat{\pi}_q)'}, \tag{8}$$

where, $r_i \in (1,2,\ldots,R)$.

It is important that the condition $R\sum_j (K_j-1)+(R-1) \le n$ on the number of parameters should hold. Also, $R\sum_j (K_j-1)+(R-1) \le (3^{10}-1)$, i.e. one fewer than the total number of cells in the cross-classification table of the manifest variables, as then the latent class model will be unidentified. Under the assumptions of multinomial distribution, the log likelihood function can be given as

$$lnL = \sum_{i=1}^{n} ln \sum_{r=1}^{R} p_r \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jrk})^{Y_{ijk}}. \tag{9}$$

LCA not only builds a measurement or classification model but it also explains a relation of the class membership to explanatory variables by including covariates (single grouping variable or a combination of grouping variables) (Vermunt, 2010) in the model. These explanatory variables are referred to as covariates, predictors, external variables, independent variables, or concomitant variables. In a more explanatory study, one may wish to build a predictive or structural model for class membership whereas in a more descriptive study the aim would be to simply profile the latent classes by investigating their association with

external variables. Grouping variables can be used in LC models in order to model the unexplained heterogeneity in the data. In that case latent class membership probabilities are predicted by covariates through a logistic link.

Once the LC models are built then the next step is to select the optimum model as different LC models have a different number of latent classes. Usually, models with more parameters (i.e more latent classes) provide a better fit, and more parsimonious models tend to have a somewhat poorer fit. So, there is always very close agreement between goodness of fit and parsimony of the latent class models. We can test the goodness of fit of an estimated LCA models by the Pearson Chi-square (χ^2) or the Likelihood Ration Chi-square (L^2). However, the likelihood ratio Chi-square test, although extensively used in the statistical literature, has a number of important limitations. These limitations can be controlled by making use of several information criteria, such as the Akaike information criterion (AIC) (Akaike (1973)) and Bayesian information criterion (BIC) (Schwartz (1978)), each of which is designed to penalize models with larger numbers of parameters. AIC and BIC on the number of parameters in the model:

$$AIC = L^2 - 2 \times d.f. \quad and \quad BIC = L^2 - d.f. \times ln(n),$$

where, n is the sample size.

These information criteria are commonly used for selecting the optimal number of latent classes in a model. By comparing models with a different number of latent classes, a model with lower AIC and BIC is selected.

2.3. Latent Regression Models

The latent class regression model generalizes the basic latent class model by permitting the inclusion of covariates to predict individuals' latent class membership (Dayton and Macready, 1988; Hagenaars and McCutcheon, 2002). This is the so-called "one-step" technique for estimating the effects of covariates, because the coefficients on the covariates are estimated simultaneously as part of the latent class model. Covariates are included in the latent class regression model through their effects on the priors p_r. In the basic latent class model, it is assumed that every individual has the same prior probabilities of latent class membership. The latent class regression model, in contrast, allows individuals' priors to vary depending upon their observed covariates.

Denote the mixing proportions in the latent class regression model as p_{ri} to reflect the fact that these priors are now free to vary by individual. It is still the case that $\sum_r p_{ri} = 1$ for each individual. To accommodate this constraint, a generalized (multinomial) logit link function can be employed for obtaining the effects of the covariates on the priors (Agresti, 2002). Let X_i represent the observed covariates for individual i. First latent class is arbitrarily selected as a "reference" class and assumes that the log-odds of the latent class membership priors with respect to that class are linear functions of the covariates. Let β_r denote the vector of coefficients corresponding to the r^{th} latent class. With S covariates, the β_r have length S + 1; this is one coefficient on each of the covariates plus a constant. Because the first class is used as the reference, $\beta_r = 0$ is fixed by definition.

Then,

$$
\begin{aligned}
ln(p_{2i}/p_{1i}) &= X_i\beta_2 \\
ln(p_{3i}/p_{1i}) &= X_i\beta_3 \\
&\vdots \\
ln(p_{Ri}/p_{1i}) &= X_i\beta_R
\end{aligned}
$$

Following some simple algebra, this produces the general result that

$$p_{ri} = p_r(X_i;\beta) = \frac{e^{X_i\beta_r}}{\sum_{q=1}^{R} e^{X_i\beta_q}}. \tag{10}$$

The parameters estimated by the latent class regression model are the $R-1$ vectors of coefficients β_r and as in the basic latent class model, the class-conditional outcome probabilities π_{jrk}. Given estimates $\hat{\beta}_r$ and $\hat{\pi}_{jrk}$ of these parameters, the posterior class membership probabilities in the latent class regression model are obtained by replacing the p_r in Eq. 8 with the function $p_r(Xi;\beta)$ in Eq. 10:

$$\hat{P}(r|X_i;Y_i) = \frac{p_r(X_i;\hat{\beta})f(Y_i;\hat{\pi}_r)}{\sum_{q=1}^{R} p_q(Xi;\hat{\beta})f(Y_i;\hat{\pi}_q)}. \tag{11}$$

The number of parameters estimated by the latent class regression model is equal to $\sum_j^R (K_j-1)+(S+1)(R-1)$. The same considerations mentioned earlier regarding model identifiability also apply here. The parameters of the LC regression model are estimated in the same way as simple LC models, the only difference is that p_r is replaced by $p_r(X_i;\beta)$.

3. Data Description

The sample included management students at post-graduation level who were using smartphones for more than one year. The methodology evolved in the research was divided into three phases. The first phase was about understanding the role of smartphone in students' life with the identification of dimensions that influence their behaviour on day-to-day basis. That led to the development of an instrument based on the literature review of studies carried out in the past. The self-report instrument from this development is supplied to the students that makes the second stage of the study, i.e. data collection. The instrument was created with the help of Google form and thus, online data was collected. The survey was administered in a top-rank state University of Jammu, India, which boasts about its excellent technological infrastructure. The third and final phase was editing, coding and analysis of data. A total of 214 responses were used in analysis. Male respondents were 71.5% and female respondents were 28.5%. The analysis was carried out with statistical software SPSS-25 and R software.

The questionnaire contained a total of 28 seven-point Likert type scale item ranging from 'completely unimportant' to 'completely important'; 2 three-point Likert type scale items ranging from 'unimportant' to 'important' and 1 two-point scaling. The questionnaire also consists of Qualitative questions on Usage of cell phones; Necessity of cell phones in modern time; Cost efficiency of cell phones over landlines; Safety reasons for carrying cell phones; Reliance on the cell phones; Dependency on the cell phones; Non calling functionality of cell phones.

Variables on mobile phone usage behaviour viz. information access, personal safety, financial incentives, social interactions, parental contacts, time management, dependency, reputations, gender, brand importance, etc, are taken to be the observed or manifest variable for the analysis. All the manifest variables have polytomous (i.e. 7) response options except for 1 that have binary responses. Accordingly, 7 latent variables have been considered on the basis of different manifest (observed) variables. A detailed description of the variables used in this paper is in Appendix A.

A pilot study was also conducted, in which the reliability and validity of the questionnaire is evaluated using SPSS software and it shows an internal consistency reliability of about 0.89 (i.e. Cronbac alpha = 0.89), which means that the internal consistency of the questionnaire is good. Also, the construct validity of the questionnaire is also quite high.

4. Results

Our data set consists of 30 variables related to the different behaviour of youths over mobile usage and 1 variable related to their gender. Due to the different behaviour our population consists of highly heterogeneous variable, which results in the violation of LCA assumption of identifiablity. Thus, we formulated a different combination of variables in the form of 7 subsets on the basis of correlation. Thus, there may be 7 latent variables and the variable selection procedure has been performed for each of them, so as to get the relevant set of indicator variables for each of them. For selecting variables, we have performed analysis by using backward greedy search algorithm (BGSA). BGSA is performed in the following way:

First, we find the number of latent classes to be considered for the data, in order to make LCA models identifiable. Because, at each step, the BGSA considers only latent class analysis models for which the identifiablity condition described in equation (5) holds. Then, we choose the maximum possible number of latent classes for which the LC model is identifiable and perform BGSA using LCAvarsel package of R software. In this case the value of upper as well as lower is 0 by default. The individual step results of the variable selection procedure starting with a relevant set of variables are given in Table 1.

The greedy search algorithm starts with two successive removal steps, then it iterates alternating between removal and inclusion step. It stops when the set of relevant predictors remains unchanged after consecutive removal and inclusion steps. Column 2 of Table 1 shows the possible set of variables proposed for each of the latent variable in the greedy search algorithm. For latent variable X_1, initially a_7 (having highest BIC) is proposed for the removal and then the difference between the BIC of the two models is computed.

Table 1: Stepwise result of variable selection algorithm

Latent variables	Proposed variables	Step	Variable	BIC diff.	Decision	Selected variables
X_1	a_1, a_2, a_4, a_7, g	Remove	a_7	190.63	Accepted	a_1, a_2, a_4, g
		Remove	NA	NA	NA	
		Add	a_7	-135.90	Rejected	
X_2	a_3, a_5, a_7, g	Remove	g	-185.33	Rejected	a_3, a_5, a_7, g
X_3	$a_8, a_9, a_{10}, a_{11}, a_{12}, g$	Remove	a_8	115.34	Accepted	$a_9, a_{10}, a_{11}, a_{12}$ g
		Remove	a_{10}	-31.69	Rejected	
		Add	a_8	-148.60	Rejected	
X_4	$a_{13}, a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, a_{19}, a_{26}, g$	Remove	a_{26}	9.13	Accepted	$a_{13}, a_{14}, a_{15}, a_{16}, a_{17}, a_{18}, a_{19}, g$
		Remove	g	-6.98	Rejected	
		Add	a_{26}	-9.13	Rejected	
X_5	$a_{20}, a_{21}, a_{22}, a_{23}, g$	Remove	a_{23}	91.94	Accepted	$a_{20}, a_{21}, a_{22}, a_{23}, g$
		Remove	NA	NA	NA	
		Add	a_{23}	7.79	Accepted	
		Remove	NA	NA	NA	
X_6	$a_{24}, a_{25}, a_{26}, a_{27}, g$	Remove	a_{26}	2.41	Accepted	$a_{24}, a_{25}, a_{27}, g$
		Add	a_{26}	-86.82	Rejected	
X_7	a_{29}, a_{30}, g	Remove	a_{29}	52.68	Accepted	a_{29}, a_{30}, g
		Add	a_{29}	7.52	Accepted	

For the removal step the LCAversal computes the difference $BIC_B - BIC_A$, that is why its decision rule's inequalities are reversed, i.e. if the BIC difference comes out to be greater than 0, then we remove the variable. For latent variable X_1, it is envisaged from Table 1, that removal of a_7 is accepted as its $BICdiff. > 0$ and the variable a_7 is removed from the set of clustering variables Y^C and placed in Y^{NC}. When performing the stepwise selection, for some combinations of clustering variables and the number of classes, it could happen that a step of the variable selection procedure could not be performed because no latent class model is identifiable on any of the possible clustering sets. In such case, the step is not performed and a NA is returned. This is what happened at the next iteration of variable selection for X_1.

After two consecutive removal steps the inclusion step is performed and the removed variable a_7 is again proposed for the inclusion in Y^C, as it is the only variable in Y^{NC}. Then again the difference between BIC of the two models is computed and if it comes out to be greater than 0, we have to include the variable. But in this case, the value of $BICdiff. < 0$, so the decision of adding a_7 in Y^C is rejected. The selected indicator variables corresponding to each latent variable are shown in Table 1. Next, we perform Latent Class Analysis on the selected indicator variables for each latent variable in order to study the estimated heterogeneity of the population along with the number of latent classes.

From the results of BGSA, we have 7 latent variables, namely Work efficiency, Up to date, superiority over landlines, Safety/Security, Dependency, Negatives, Functionality; with selected, indicators variables. Further, we perform LCA on 7 latent variables in order to explain the heterogeneity among the sub-populations and conditional item-response probabilities, using poLCA (Linzer and Lewis, 2011) package of R. The selection of an

appropriate number of latent classes for each of the latent variable can be made by comparing the values of BIC or AIC, a suitable measure for variable selection. Models considered in our analysis are simple extensions of the basic LC models as proposed by Biemer and Wiesen (2002) with grouping variables. The use of grouping variables has been suggested by Hui and Walter (1980) to either ensure that the model is identifiable, or to improve the fit of the model or to reduce the effect of unobserved heterogeneity. Following Hui-Walter approach, we choose gender, brand and addiction as grouping variables. Table 2 shows the goodness of fit statistics and other statistics with optimal number of latent classes for each latent variable.

From Table 2 it is clear that the data set is best fitted for a 3-class model for all the latent variables except for X_7, which has 2-class best fitted model, as the corresponding BIC as well as AIC values for each of the latent variables are the lowest. The data were inconsistent with several other possible models for each of the latent variable.

Table 2: The Model diagnostics for different latent variables

Latent variables	**No. of opt. classes**	**Grouping variables**	**Residual d.f.**	**Max. log-likelihood**	χ^2	**AIC**	**BIC**
X_1	3	-	158	-1010.415	310.0607	2134.849	2355.075
		g	156	-1006.924	311.4976	2132.83	2349.325
		g&b	152	-1003.858	336.3663	2131.716	2340.407
X_2	3	-	158	-1072.475	267.0177	2279.204	2482.431
		g	156	-1068.602	279.4936	2266.949	2475.444
		g&b	152	-1064.097	312.6338	2252.195	2460.885
X_3	3	-	140	-1315.987	2585.898	2819.974	3169.056
		g	138	-1293.091	2695.378	2738.182	2993.996
X_4	3	-	86	-1807.442	696845.5	3870.883	4301.728
		g	84	-1801.077	1006308	3862.154	4299.73
X_5	3	-	140	-1316.358	1023.609	2780.716	3029.798
		g	138	-1310.409	1559.712	2772.819	3028.633
		g&a	134	-1300.755	1726.56	2761.51	3020.788
X_6	3	-	158	-1148.867	285.8787	2369.734	2578.228
		g	156	-1128.732	286.4188	2345.464	2558.691
		g&a	152	-1105.955	329.6206	2335.911	2544.601
X_7	2	g	8	-781.8273	50.3265	1593.655	1698.294
		g&b	4	-766.1268	45.0985	1585.254	1688.356
		g&a	4	-766.1388	51.1466	1585.278	1688.38
		g&b&a	16	-757.6595	94.3307	1579.319	1687.03

$d.f.$: *degrees of freedom*
χ^2 : *Chi − square value*
AIC : *Akaike information criterion*
BIC : *Bayesian information criterion*

Since, we have heterogeneous indicator variables measuring the different latent variables and different grouping variables, for this reason various models were considered and

tested for their identifiability. Model diagnostics are provided in tTable 2 and the path models of identifiable and efficient models are provided in Figure 2, 3 and 4, for each of the latent variable.

Figure 2: Path models for latent variables X_1, X_2 and X_3

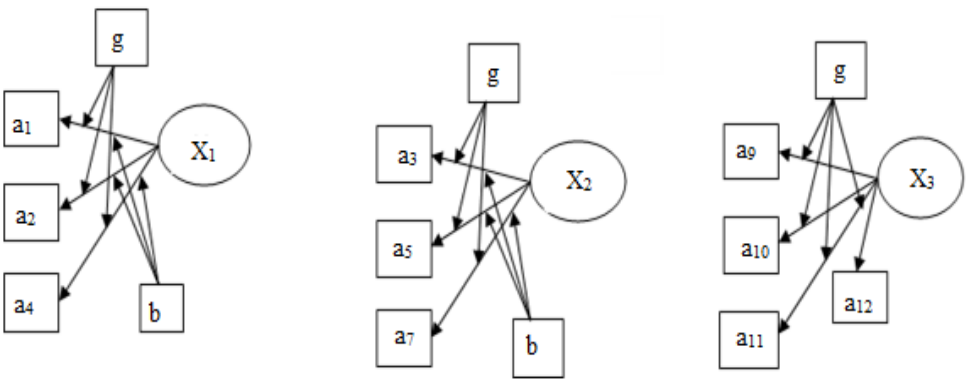
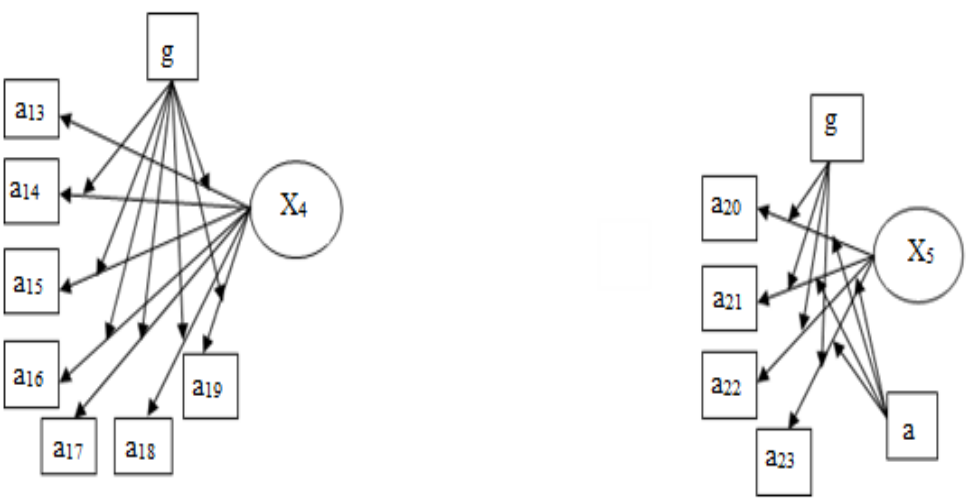


Figure 3: Path models for latent variables X_4 and X_5



From the results of Table 2, it is clear that the data set is best fitted for a 3-class model for all the latent variables except for X_7, which has 2-class for the best fitted model. Therefore, the underlying latent classes can be identified as "improve"(class 1), "same" (class 2) and "worsen" (class 3) for the latent variables X_1, X_2, X_3, X_4, X_5 *and* X_6. And for latent variable X_7, the underlying latent classes can be identified as "non-calling" (class1) and "calling" (class2).

Figure 4: Path models for latent variables X_6 and X_7

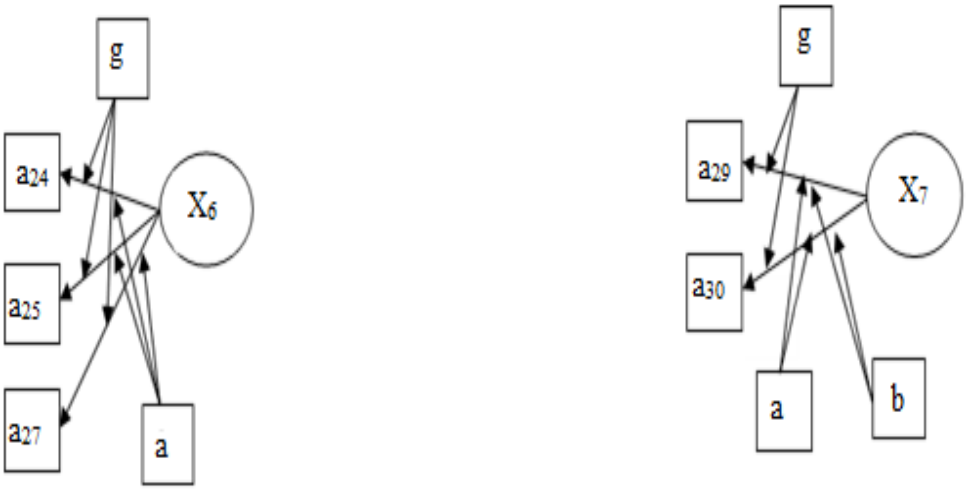


Table 3: Estimated class membership for different latent variables

Latent variables	Grouping variables	Latent class 1	Latent class 2	Latent class 3
Work efficiency (X_1)	-	0.1946	0.2987	0.5067
	g	0.2846	0.5149	0.2005
	g&b	0.4871	0.3453	0.1676
Up to date (X_2)	-	0.2514	0.4393	0.3093
	g	0.1472	0.5115	0.3785
	g&b	0.3724	0.1721	0.4555
Superiority over landlines (X_3)	-	0.3883	0.289	0.3227
	g	0.334	0.2344	0.4316
Safety/security (X_4)	-	0.1159	0.5639	0.3202
	g	0.1605	0.1577	0.6818
Dependency (X_5)	-	0.3645	0.1916	0.4439
	g	0.1663	0.5233	0.3104
	g&a	0.2919	0.2376	0.4705
Negatives (X_6)	-	0.4879	0.1743	0.3379
	g	0.1755	0.4936	0.3309
	g&a	0.3034	0.3236	0.373
Functionality (X_7)	-	0.374	0.626	-
	g	0.1962	0.4416	0.3622
	g&b&a	0.2828	0.7172	-

The estimated class membership probabilities from the latent class model are summarized in Table 3. For latent variable X_1, the estimated class membership probabilities with grouping variable gender are 0.2846 for class 1, 0.5149 for class 2 and 0.2005 for class 3, while the estimated class membership probabilities of the same with the inclusion of 1 more

grouping variable, i.e. brand, significantly increases to 0.4871 for class 1, 0.3453 for class 2and 0.1676 for class 3, which indicates that the work efficiency increases with the brand of mobile. And for latent variable X_2, the estimated class membership probabilities with single grouping variable gender are 0.1472 for class 1, 0.5115 for class 2 and 0.3785 for class 3, which again increases significantly with the inclusion of 1 more grouping variable, i.e. brand, to 0.3724, 0.1721 and 0.4555, respectively.

For latent variable X_3 and X_4, the estimated class membership probabilities increases significantly with the grouping variables gender. Further, for latent variable X_5 and X_6, the estimated class membership probabilities increase with two grouping variables, i.e. gender and addiction. And for latent variable X_7, the estimated class membership probabilities with single grouping variable, gender are 0.1962 for class 1, 0.4416 for class 2 and 0.3622 for class 3. Whereas the same probabilities increase to 0.2828 for class 1 and 0.7172 for class 2, with the inclusion of 2 more grouping variables, namely brand and addiction.

Thus, from Table 3 it can be seen that, for all the latent variables considered in the analysis, all are showing higher probabilities of being a member in the latent class 3 than in the latent class 2 or latent class 1, except for latent variable X_1, which has a higher probability of being in latent class 1 and X_7, which has a higher probability of being in latent class 2.

Table 4: Estimated conditional item response probabilities

	Indic-	Latent	Categories						
	tors	classes	1	2	3	4	5	6	7
		1	0.0000	0.0126	0.0844	0.1399	0.2886	0.4331	0.0413
	a_1	2	0.0000	0.0588	0.0000	0.0000	0.0000	0.3635	0.5777
		3	0.1395	0.1768	0.1728	0.0116	0.4994	0.0000	0.0000
		1	0.0000	0.0000	0.1296	0.2086	0.4016	0.2603	0.0000
X_1	a_2	2	0.0275	0.0918	0.0000	0.1096	0.0966	0.2689	0.4056
		3	0.3058	0.1732	0.0696	0.4228	0.0000	0.0000	0.0286
		1	0.0000	0.0096	0.0622	0.0468	0.2511	0.3619	0.2683
	a_4	2	0.0952	0.0143	0.0000	0.0226	0.0690	0.2050	0.5939
		3	0.0549	0.0263	0.0422	0.2913	0.4384	0.0593	0.0877
		1	0.0355	0.0137	0.0362	0.1368	0.1552	0.4503	0.1724
	a_3	2	0.0400	0.1061	0.1017	0.0721	0.4888	0.1914	0.0000
		3	0.0687	0.0000	0.0551	0.0456	0.1911	0.1341	0.5054
		1	0.0141	0.0374	0.0381	0.2422	0.2892	0.1950	0.1840
X_2	a_5	2	0.0000	0.1916	0.1117	0.1229	0.2973	0.2765	0.0000
		3	0.0706	0.0202	0.0805	0.0941	0.0000	0.2387	0.4959
		1	0.0223	0.0882	0.0919	0.2347	0.0787	0.2133	0.2709
	a_7	2	0.1598	0.1078	0.1269	0.0000	0.2871	0.1461	0.1723
		3	0.0342	0.0000	0.0000	0.0646	0.2786	0.3039	0.3187
		1	0.0974	0.1110	0.0899	0.0172	0.2832	0.2617	0.1396
	a_9	2	0.0369	0.0213	0.1569	0.4140	0.1566	0.0000	0.2144
		3	0.0129	0.0000	0.0293	0.0000	0.0206	0.1656	0.7716
		1	0.0879	0.1539	0.0559	0.1046	0.2215	0.2841	0.0921
	a_{10}	2	0.0488	0.0000	0.2393	0.4690	0.1022	0.1408	0.0000
X_3		3	0.0246	0.0000	0.0000	0.0000	0.0438	0.2342	0.6974

	Indic-	Latent	Categories						
	tors	classes	1	2	3	4	5	6	7
	a_{11}	1	0.0551	0.1119	0.1106	0.0840	0.2054	0.3647	0.0682
		2	0.0000	0.0000	0.0218	0.4234	0.2257	0.0000	0.3292
		3	0.0115	0.0000	0.0108	0.0298	0.0000	0.1617	0.7862
	a_{12}	1	0.1652	0.1306	0.1566	0.1324	0.1277	0.2874	0.0000
		2	0.0374	0.0923	0.1199	0.4178	0.2719	0.0245	0.0361
		3	0.0467	0.0112	0.0194	0.0279	0.0891	0.2298	0.5759
	a_{13}	1	0.0000	0.0000	0.0340	0.1150	0.4882	0.3015	0.0612
		2	0.0593	0.0603	0.0840	0.2334	0.1514	0.2136	0.1981
		3	0.0000	0.0066	0.0068	0.0080	0.0420	0.1263	0.8102
	a_{14}	1	0.0000	0.0873	0.0000	0.0000	0.1181	0.7481	0.0464
		2	0.0000	0.0000	0.1482	0.2964	0.2055	0.0922	0.2577
		3	0.0137	0.0000	0.0000	0.0000	0.0206	0.0493	0.9164
	a_{15}	1	0.0293	0.0582	0.0287	0.000	0.2635	0.5319	0.0884
		2	0.0000	0.1183	0.1190	0.321	0.2534	0.1552	0.0331
		3	0.0411	0.0069	0.0069	0.008	0.0301	0.1336	0.7734
X_4	a_{16}	1	0.0873	0.0291	0.1491	0.0280	0.5203	0.1862	0.0000
		2	0.0852	0.0896	0.1175	0.4252	0.1530	0.0991	0.0305
		3	0.0831	0.0272	0.0268	0.0870	0.1574	0.1663	0.4521
	a_{17}	1	0.0000	0.0574	0.0291	0.0000	0.1792	0.4605	0.2738
		2	0.0000	0.0342	0.0891	0.3966	0.1773	0.1848	0.1179
		3	0.0274	0.0540	0.0137	0.0454	0.0539	0.1024	0.7033
	a_{18}	1	0.0000	0.0584	0.0000	0.0300	0.2379	0.6445	0.0292
		2	0.1499	0.0904	0.1196	0.3279	0.1023	0.1529	0.0570
		3	0.0544	0.0476	0.0683	0.0199	0.1259	0.1967	0.4871
	a_{19}	1	0.0000	0.0291	0.0000	0.0000	0.0629	0.5629	0.3450
		2	0.0296	0.0000	0.0889	0.1484	0.3171	0.2772	0.1387
		3	0.0000	0.0000	0.0000	0.0342	0.0147	0.1872	0.7639
	a_{20}	1	0.0341	0.1000	0.1601	0.0343	0.2779	0.3065	0.0872
		2	0.0709	0.0000	0.0000	0.4495	0.3483	0.1313	0.0000
		3	0.0026	0.0075	0.0000	0.0000	0.0688	0.1309	0.7902
	a_{21}	1	0.0732	0.1212	0.1413	0.1166	0.2538	0.2486	0.0453
		2	0.1459	0.0478	0.0821	0.3367	0.2257	0.1367	0.0252
		3	0.0100	0.0000	0.0000	0.0357	0.0662	0.2535	0.6346
X_5	a_{22}	1	0.0000	0.1601	0.0250	0.0000	0.2837	0.2869	0.2443
		2	0.0585	0.0000	0.0000	0.4485	0.0560	0.2388	0.1982
		3	0.0201	0.0000	0.0441	0.0218	0.0440	0.1682	0.7018
	a_{23}	1	0.2360	0.1806	0.1702	0.1703	0.2429	0.0000	0.0000
		2	0.2494	0.0636	0.0000	0.4142	0.1439	0.1289	0.0000
		3	0.1249	0.0843	0.0931	0.1718	0.2534	0.1038	0.1688
	a_{24}	1	0.0738	0.0927	0.2772	0.1631	0.3185	0.0321	0.0426
		2	0.1474	0.1395	0.0000	0.0000	0.0000	0.2340	0.4791
		3	0.0000	0.0166	0.0000	0.2432	0.3924	0.3096	0.0382
	a_{25}	1	0.0000	0.1601	0.0774	0.3035	0.1927	0.1962	0.0701
X_6		2	0.2021	0.0269	0.0198	0.0000	0.1330	0.0943	0.5238
		3	0.0000	0.0219	0.0827	0.1665	0.2039	0.2973	0.2277

	Indic-tors	Latent classes	Categories						
			1	2	3	4	5	6	7
	a_{27}	1	0.0000	0.1997	0.1399	0.4699	0.1613	0.0000	0.0292
		2	0.1604	0.0582	0.1093	0.0699	0.0591	0.2006	0.3425
		3	0.0612	0.0000	0.0796	0.0207	0.3813	0.3772	0.0799
	a_{29}	1	0.7021	0.0372	0.0000	0.0085	0.0293	0.0289	0.1941
		2	0.1467	0.1417	0.0977	0.2507	0.1579	0.1515	0.0538
X_7	a_{30}	1	0.4618	0.0000	0.0236	0.0000	0.0074	0.000	0.507
		2	0.0590	0.1173	0.0950	0.2802	0.2121	0.202	0.0345

Table 4 provides the estimated conditional item response probabilities for each of the indicator variables corresponding to the different latent variables. The rows of Table 4 correspond to different latent classes of each latent variables and columns correspond to different categories of each of the indicator variable. For latent variable X_1 (work efficiency), the conditional probabilities $P[a_1 = 7|X_1 = 1] = 0.6466$ and $P[a_4 = 7|X_1 = 1] = 0.6393$, by considering grouping variable gender, envisaged that the respondents (students) are confused on the importance of cell phones for efficient time usage and multitasking given their work efficiency, respectively. While the fact that the same conditional probabilities by considering two grouping variables, namely gender and brand, reduces to 0.0413 and 0.2683, respectively, indicates that the brand of mobile phones plays an important role for students to make their work efficiently.

For latent variable X_2 (up to date), the conditional probabilities obtained by considering grouping variables, gender and brand, did not depict any peculiarities in the responses of the respondents. This means that the respondents believe that they remain up to date with the help of cell phones.

For latent variable X_3 (superiority over landlines), the conditional probabilities $P[a_9 = 7|X_3 = 1] = 0.8282$, $P[a_{10} = 7|X_3 = 1] = 0.7734$, $P[a_{11} = 7|X_3 = 1] = 0.7881$ and $P[a_{12} = 7|X_3 = 1] = 0.6037$, without considering any covariates, indicate that the respondents are not sure about their preference of cell phones over landlines. While the inclusion of the grouping variable gender, these conditional probabilities reduce to 0.1396, 0.0921, 0.0682 and 0.000, respectively, which indicates that the respondents have influence on their preference of cell phones over landlines.

For latent variable X_4 (safety/security), none of the conditional probabilities, when considered with the grouping variable, gender, indicates any sort of unexpected behaviour. This implies that respondents believed that carrying cell phones make them feel safe and secure.

Further, for latent variables X_5(dependency) and X_6 (negatives), the respondents believe that they are dependent on the cell phones which have negative impacts on their life. And for latent variable X_7 (functionality), the results signify that the respondents are comfortable for providing their opinion about the non-calling functionality of cell phones.

Table 5: Estimated coefficients on the covariates along with standard error

Latent variables	Latent classes	Covariates	Coefficient	Std. error	t-statistics	p-value
X_1	2	(Intercept)	-6.2089	2.2710	-2.734	0.007
		g	4.2355	1.7516	2.418	0.017
		b	2.3226	0.8163	2.845	0.005
		g:b	-1.6652	0.6313	-2.638	0.009
	3	(Intercept)	30.4070	0.4634	65.611	0.000
		g	-31.2964	0.5003	-62.555	0.000
		b	-10.0995	0.3870	-26.097	0.000
		g:b	10.1252	0.2639	38.367	0.000
X_2	2	(Intercept)	51.2779	4.3713	11.730	0
		g	-43.2907	4.4161	-9.803	0
		b	-18.1251	1.6642	-10.891	0
		g:b	14.7831	1.4530	10.174	0
	3	(Intercept)	46.5247	4.3529	10.688	0.000
		g	-39.9030	4.3566	-9.159	0.000
		b	-15.4059	1.4736	-10.454	0.000
		g:b	13.2461	1.4492	9.140	0.000
X_3	2	(Intercept)	-1.1287	0.7469	-1.511	0.133
		g	0.6227	0.5670	1.098	0.274
	3	(Intercept)	-0.7873	0.6032	-1.305	0.194
		g	0.8256	0.4564	1.809	0.073
X_4	2	(Intercept)	-15.3427	0.8012	-19.148	0
		g	14.9957	0.4724	31.743	0
	3	(Intercept)	-14.3230	0.6323	-22.65	0.000
		g	15.3336	0.5957	25.74	0.000
X_5	2	(Intercept)	-37.4900	0.5915	-63.379	0
		g	35.0006	0.6209	56.369	0
		a	12.8783	0.4421	29.126	0
		g:a	-12.0000	0.3340	-35.919	0
	3	(Intercept)	-5.8590	0.6810	-8.603	0.000
		g	1.6838	0.6976	2.414	0.017
		a	2.0573	0.3646	5.642	0.000
		g:a	-0.4166	0.2767	-1.505	0.135
X_6	2	(Intercept)	30.3755	0.4583	66.277	0
		g	-31.9703	0.4712	-67.842	0
		a	-10.3418	0.3147	-32.857	0
		g:a	10.9406	0.2395	45.670	0
	3	(Intercept)	29.2397	0.5782	50.568	0.000
		g	-31.2611	0.5784	-54.044	0.000
		a	-9.9287	0.3416	-29.063	0.000
		g:a	10.7219	0.2797	38.331	0.000

Latent variables	Latent classes	Covariates	Coefficient	Std. error	t-statistics	p-value
X_7	2	(Intercept)	90.1094	0.2250	400.318	0.000
		g	-42.1254	0.1856	-226.881	0.000
		b	-40.2853	0.1970	-204.410	0.000
		a	-30.1543	0.6356	-47.442	0.000
		g:b	24.9592	0.2027	123.114	0.000
		g:a	14.1658	0.5325	26.598	0.000
		b:a	13.4523	0.2485	54.115	0.000
		g:b:a	-8.2771	0.2005	-41.277	0.000

Table 5 provides the estimated coefficients along with standard errors. It is observed from Table 5 that all the covariates for the selected latent class model with 3 latent classes are highly significant at 5% level of significance. Here, the 1^{st} latent classes have been identified as "improve", 2^{nd} latent class has been named as "same" and finally the 3^{rd} latent class has been named as "worsen", for all latent variables, except for functionality for which the two classes are "non-calling" and "calling", respectively. It is also observed that with inclusion of significant covariates for each of the latent variable, the value of the standard errors and t- statistics significantly reduces. Next, we consider the latent regression model which permits the inclusion of covariates to predict individual latent class membership. The following table provides the log-ratio prior probability that a respondent will belong to the same group with respect to the 1^{st} group.

From appendix B, we can calculate the predicted prior probabilities on substituting the response of the i^th individual to the corresponding grouping variable. These probabilities will help to evaluate the estimated latent class membership of individuals with the inclusion of covariates.

5. Conclusions

This study considered the problem of selection of indicator variables for different latent variables and to identify different behavioural classes among latent variables. We adopt BGSA (Backward Greedy Search Algorithm) for indicator variables selection and LCA (Latent Class Analysis) for identifying the different behavioural classes. As an application we have investigated the different behaviours of youth towards the usage of mobile phones through questionnaire comprised of 31 items. The questionnaire consists of qualitative questions on usage of cell phones, necessity, cost factors over landlines, safety reasons for carrying cell phones, etc.

The different behaviours of mobile usage has been identified through BGSA as work efficiency, up to date, Superiority over landlines, Safety/Security, Dependency, Negatives and Functionality. Further, we have performed LCA to examine the heterogeneity among the population of youths. The latent classes have been identified as "improve", "same" and "worsen", for all latent variables, except for functionality, which includes two classes, namely "non-calling" and "calling", respectively.

It is envisaged that the different brands availability decreases their efficiency, 45% of students believe that they get updated with cell phones. Also, around 68% of the students

feel safe, while having cell phones and 47% of the students get addicted towards the usage of cell phones. Lastly, due to the non-calling functions of cell phones, 72% of the students gets addicted to cell phones because of the brand of cell phones.

Next, we constructed latent class regression models for each of the latent variable with covariates, which help us to predict the latent class membership of an individual. As the study is limited to the number of respondents with specific specialization, we recommend the proposed methodology to be used for a more diversified sample.

Acknowledgements

The authors are highly grateful to the learned referees for their constructive comments/ suggestions, which helped in the improvement of the paper.

References

AKAIKE, H., (1973). Information Theory and an Extension of the Maximum Likelihood Principle. $2^{(}nd)$ *International Symposium on Information Theory*, pp. 267–281.

BARTHOLOMEW, D., KNOTT, M. and MOUSTAKI, I., (2011). Latent Variable Models and Factor Analysis. *Wiley*.

BAUMGARTNER, H. and JAN-BENEDICT, E. M. STEENKAMP, (2006). Response Biases in Marketing Research. *Handbook of Marketing Research*, Thousand Oaks, CA: Sage, pp. 95–109.

BIEMER, P., (2010). Latent Class Analysis of survey error. *A John Willey and Sons*, Inc. publications.

BIEMER, P., WIESEN, C., (2002). Latent class analysis of embedded repeated measurements: An application to the National Household Survey on Drug Abuse. *Journal of the Royal Statistical Society*, Series A, 165(1), pp. 97–119.

BODUSZEK, D., O'SHEA, C., DHINGRA, K. and HYLAND, P., (2014). Latent Class Analysis of Criminal Social Identity in a Prison Sample. *Polish Psychological Bulletin*, 45(2), pp. 192–199.

DAYTON, C., MACREADY, G. , (1988). Concomitant-Variable Latent-Class Models. *Journal of the American Statistical Association*, 83(401), pp. 173–178.

DEAN, N., RAFTERY, A. E., (2010). Latent class analysis variable selection. *Annals of the Institute of Statistical Mathematics*, 62, pp. 11–35.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B., (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B (Methodological), 39(1), pp. 1–38.

FOP, M., SMART, K. M. and MURPHY, T. B., (2017). Variable selection for latent class analysis with application to low back pain diagnosis. *Annals of Applied Statistics*, 11(4), pp. 2085–2115.

FOP, M., MURPHY, T. B., (2017). LCAvarsel: Variable selection for latent class analysis R package version, https://cran.r-project.org/package=LCAvarsel.

FORMANN, A. K., (1984). Constrained latent class models: theory and applications. *British Journal Mathematical and Statistical Psychology*, 38, pp. 87–111.

FORMANN, A. K., (1992). Linear logistic latent class analysis for polytomous data. *Journal of American Statistical Association*, 87, pp. 476–486.

GOODMAN, L. A., (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61, pp. 215–231.

HABERMAN, S. J., (1979). Analysis of Qualitative Data. *New York: Academic Press 1979*; Vol. 2: New Developments.

HAGENAARS, J. A. and MCCUTCHEON, A. L., (2002). Applied Latent Class Analysis. *Cambridge University Press*, New York.

HUI, S. L., WALTER, S. D., (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, pp. 167–171.

KASS, R. E., RAFTERY, A. E., (1995). Bayes factors. *Journal of the American Statistical Association*, 90, pp. 773–795.

KUMAR, S., (2015). Diagnose response bias and heterogeneity: A Latent class approach on Indian household inflation expectation survey. *International Journal of Advances in Social Sciences*, 3(4), pp. 152–158.

KUMAR, S., (2016). Latent class analysis for reliable measure of inflation expectation in the Indian public. *European Journal of Economic and Statistics*, 1(1), pp. 9–16.

KUMAR, S., HUSAIN, Z., and MUKHERJEE, D., (2017). Assessing consistency of consumer confidence data using latent class analysis with time factor. *Economic Analysis & Policy*, 55, pp. 35–46.

LANZA, S. T., RHOADES, B. L., (2013). Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prevention science : The Official Journal of the Society for Prevention Research*, 14(2), pp. 157–168.

LAZARSFELD, P. F., (1950). The logical and mathematical foundation of latent structure analysis and the interpretation and mathematical foundation of latent structure analysis. *S.A. Stou er et al. (eds.), Measurement and Prediction*, pp. 362–472. Princeton, NJ: Princeton University Press.

LINZER, D. A., LEWIS, J., (2011). poLCA: Polytomous Variable Latent Class Analysis. *Annals of Applied Statistics*, 11(4), pp. 2085–2115, http://CRAN.R-project.org/package=poLCA.

MCLACHLAN, G. and PEEL, D., (2000). Finite Mixture Models. *John Wiley & Sons*, New York.

MCLACHLAN, G., KRISHNAN, T., (2008). The EM Algorithm and Extensions. *Wiley*.

MOOIJAART, A. B., (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika*, 57(2), pp. 261–269.

PETERSEN, K. J., QUALTER, P. and HUMPHREY, N., (2019). The Application of Latent Class Analysis for Investigating Population Child Mental Health: A Systematic Review. *Frontiers in Psychology*, Vol.10.

PORCU, M., GIAMBONA, F., (2017). Introduction to Latent Class Analysis with Applications. *The Journal of Early Adolescence*, 37(1), pp. 129–158.

RAFTERY, A. E., DEAN, N., (2006). Variable selection for model-based clustering. *The Journal of American Statistical Association*, 101, pp. 168–178.

SAPOUNIDIS, T., STAMOVLASIS, D. and DEMETRIADIS, S., (2019). Latent Class Modeling of Children Preference Profiles on Tangible and Graphical Robot Programming. *IEEE Transactions on Education*, 62(2), pp. 127–133.

TRAI, (2018). https://www.medianama.com/2018/03/223-india-had-1-012-billion-active-mobile-connections-in-january-2018-trai/.

VERMUNT, K., JEROEN, K., (2010). Latent class modelling with covariates: Two improved three-step approaches. *Political Analysis*, 18, pp. 450–469.

APPENDICES

6. Detailed description of variables

Variables	Descriptions
a1	A cell phone allows me to use my time efficiently
a2	I use my cell phone to make use of time that otherwise would be wasted
a3	We need a cell phone to be successful in the world today
a4	A cell phone allows me to do two things at once
a5	Those people who do not have a cell phone are out of touch with modern world
a7	I often use my cell phone to schedule/reschedule an appointment at the last minute
a8	It is financially beneficial to use a cell phone as opposed to a landline
a9	A cell phone is more affordable than a landline phone service
a10	If I had to choose, I would use a cell phone instead of a landline because a cell phone is cheaper
a11	A cell phone is a cheaper alternative for long distance calls than a landline
a12	I do not use landlines because having a cell phone is cheaper
a13	Having a cell phone makes me feel safe while I am walking alone at night
a14	My parent wanted me to have a cell phone so I can get in touch with her/him if necessary
a15	I use my cell phone to keep my parent from worrying about me
a16	Having a cell phone makes me feel safe while I am driving
a17	I got my cell phone to use in case of emergency
a18	My parent worries about me less because I have a cell phone
a19	With a cell phone I can keep in touch with my family members
a20	When I do not have my cell phone with me, I feel disconnected
a21	I feel lost when I leave my cell phone at home
a22	I always leave my cell phone on
a23	I feel upset when I miss a call to my cell phone
a24	A cell phone distracts me from being aware of my surroundings
a25	I feel embarrassed by my cell phone ringing at inappropriate times
a26	I am often distracted by my cell phone when driving
a27	I am tired of being accessible all the time
a29	I do not care to learn how to use non-calling functions on my cell phone
a30	I seldom use non-calling functions of my cell phone
g	Gender
b	The brand of a cell phone is important to me
a	A cell phone is addictive

7. Log ratio and Predicted prior probabilities for each of the latent variable

	r	Log ratio prior probability $ln(\frac{p_{ri}}{p_{1i}})$	Predicted prior probability p_{ri}
X_1	2	$-6.2089+4.23554\times g_i+2.3226\times b_i-1.6652\times(g_i:b_i)$	$\frac{\exp^{-6.2089+4.23554\times g_i+2.3226\times b_i-1.6652\times(g_i:b_i)}}{1+\exp^{-6.2089+4.23554\times g_i+2.3226\times b_i-1.6652\times(g_i:b_i)}+\exp^{30.4070-31.2964\times g_i-10.0995\times b_i+10.1252\times(g_i:b_i)}}$
	3	$30.4070-31.2964\times g_i-10.0995\times b_i+10.1252\times(g_i:b_i)$	$\frac{\exp^{30.4070-31.2964\times g_i-10.0995\times b_i+10.1252\times(g_i:b_i)}}{1+\exp^{-6.2089+4.2355\times g_i+2.3226\times b_i-1.6652}+\exp^{30.4070-31.2964\times g_i-10.0995\times b_i+10.1252\times(g_i:b_i)}}$
X_2	2	$51.2779-43.2907\times g_i-18.1251\times b_i+14.7831\times(g_i:b_i)$	$\frac{\exp^{51.2779-43.2907\times g_i-18.1251\times b_i+14.7831\times(g_i:b_i)}}{1+\exp^{51.2779-43.2907\times g_i-18.1251\times b_i+14.7831\times(g_i:b_i)}+\exp^{46.5247-39.9030\times g_i-15.4059\times b_i+13.2461\times(g_i:b_i)}}$
	3	$46.5247-39.9030\times g_i-15.4059\times b_i+13.2461\times(g_i:b_i)$	$\frac{\exp^{46.5247-39.9030\times g_i-15.4059\times b_i+13.2461\times(g_i:b_i)}}{1+\exp^{51.2779-43.2907\times g_i-18.1251\times b_i+14.7831\times(g_i:b_i)}+\exp^{46.5247-39.9030\times g_i-15.4059\times b_i+13.2461\times(g_i:b_i)}}$
X_3	2	$-1.1287+0.6227\times g_i$	$\frac{\exp^{-1.1287+0.6227\times g_i}}{1+\exp^{-1.1287+0.6227\times g_i}+\exp^{-0.7873+0.8256\times g_i}}$
	3	$-0.7873+0.8256\times g_i$	$\frac{\exp^{-0.7873+0.8256\times g_i}}{1+\exp^{-1.1287+0.6227\times g_i}+\exp^{-0.7873+0.8256\times g_i}}$
X_4	2	$-15.3427+14.9957\times g_i$	$\frac{\exp^{-15.3427+14.9957\times g_i}}{1+\exp^{-15.3427+14.9957\times g_i}+\exp^{-14.3230+15.3336\times g_i}}$
	3	$-14.3230+15.3336\times g_i$	$\frac{\exp^{-14.3230+15.3336\times g_i}}{1+\exp^{-15.3427+14.9957\times g_i}+\exp^{-14.3230+15.3336\times g_i}}$

	r	Log ratio prior probability $ln(\frac{p_{ri}}{p_{1i}})$	Predicted prior probability p_{ri}
X_5	2	$-37.4900+35.0006\times g_i+12.8783\times a_i-12.0000\times(g_i:a_i)$	$\frac{\exp^{-37.4900+35.0006\times g_i+12.8783\times a_i-12.0000\times(g_i:a_i)}}{1+\exp^{-37.4900+35.0006\times g_i+12.8783\times a_i-12.0000\times(g_i:a_i)}+\exp^{-5.8590+1.6838\times g_i+2.0573\times a_i-0.4166\times(\mathbf{g_i}:\mathbf{ab_i})}}$
	3	$-5.8590+1.6838\times g_i+2.0573\times a_i-0.4166\times(\mathbf{g_i}:\mathbf{ab_i})$	$\frac{\exp^{-5.8590+1.6838\times g_i+2.0573\times a_i-0.4166\times(g_i:ab_i)}}{1+\exp^{-37.4900+35.0006\times g_i+12.8783\times a_i-12.0000\times(g_i:a_i)}+\exp^{-5.8590+1.6838\times g_i+2.0573\times a_i-0.4166\times(g_i:a_i)}}$
X_6	2	$30.3755-31.9703\times g_i-10.3418\times a_i+10.9406\times(g_i:a_i)$	$\frac{\exp^{30.3755-31.9703\times g_i-10.3418\times a_i+10.9406\times(g_i:a_i)}}{1+\exp^{30.3755-31.9703\times g_i-10.3418\times a_i+10.9406\times(g_i:a_i)}+\exp^{29.2397-31.2611\times g_i-9.9287\times a_i+10.7219\times(g_i:a_i)}}$
	3	$29.2397-31.2611\times g_i-9.9287\times a_i+10.7219\times(g_i:a_i)$	$\frac{\exp^{29.2397-31.2611\times g_i-9.9287\times a_i+10.7219\times(g_i:a_i)}}{1+\exp^{30.3755-31.9703\times g_i-10.3418\times a_i+10.9406\times(g_i:a_i)}+\exp^{29.2397-31.2611\times g_i-9.9287\times a_i+10.7219\times(g_i:a_i)}}$
X_7	2	$90.1094-42.1254\times g_i-40.2853\times b_i-30.1543\times a_i+24.9592\times(g_i:b_i)+14.1658\times(g_i:a_i)+13.4523\times(b_i:a_i)-8.2771\times(\mathbf{g_i}:\mathbf{b_i}:\mathbf{a_i})$	$\frac{\exp^{90.1094-42.1254\times g_i-40.2853\times b_i-30.1543\times a_i+24.9592\times(g_i:b_i)+14.1658\times(g_i:a_i)+13.4523\times(b_i:a_i)-8.2771\times(g_i:b_i:a_i)}}{1+\exp^{90.1094-42.1254\times g_i-40.2853\times b_i-30.1543\times a_i}+24.9592\times(g_i:b_i)+14.1658\times(g_i:a_i)+13.4523\times(b_i:a_i)-8.2771\times(g_i:b_i:a_i)}$

where r: latent class (r=2,3) and
i: no. of respondents (i=1,2,...,N)

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 115–130, DOI 10.21307/stattrans-2021-006
Received – 07.01.2020; accepted – 23.12.2020

Analysis of Polish mutual funds performance: a Markovian approach

Dariusz Filip[1], **Tomasz Rogala**[2]

ABSTRACT

The aim of this study is to determine whether mutual funds provide benefits for their clients. The performance of Polish mutual funds has been evaluated in terms of their efficiency, including their potential inertia over time. Moreover, the use of the phenomenon of economies of scale resulting from assets inflow to the fund by means of the Markovian framework has been examined. The results are consistent with the efficient market hypothesis. When assessing the market-adjusted returns, underperformance was noticed in both small and large funds. The smart money effect, recognised in the literature, is not confirmed here; however, there are some noticeable investor reactions, such as the phenomenon of chasing performance.

Key words: Markov chain, smart money effect, effectiveness, performance inertia.

1. Introduction

One of the most discussed issues concerning financial markets in both scientific periodicals and specialized magazines is the efficiency of investment projects. In the area of capital asset management, extensive debates on the efficient market hypothesis have been held since the 1970s. It assumes that financial markets reflect the publicly available information accurately and efficiently. Moreover, no investment strategies based on past prices of financial instruments are capable of providing abnormal returns. It is also believed that there are no investors with access to confidential information permitting generation of abnormal returns operating in the market (Fama, 1970). The human capital theory, which is opposed to the abovementioned hypothesis, provides that portfolio managers might be able to gather, process, and use the data with which

[1] Cardinal Stefan Wyszynski University in Warsaw (UKSW), Faculty of Social and Economic Sciences, Department of Finance, Poland. E-mail: d.filip@uksw.edu.pl.
ORCID: https://orcid.org/0000-0002-6905-1004.

[2] Cardinal Stefan Wyszynski University in Warsaw (UKSW), Faculty of Mathematics and Natural Sciences, Institute of Mathematics, Poland. E-mail: t.rogala@uksw.edu.pl. ORCID: https://orcid.org/0000-0002-0817-4377.

other investors cannot familiarize themselves. Such an ability permits certain market participants to achieve a competitive advantage.

Investors who can be characterized by such skills include managers employed in investment fund companies. Managers' stock selection abilities should translate into both the achieved performance and its persistence. The literature on the subject matter has developed the term "abnormal return" and has called the phenomenon reflecting the tendency of achieving similar results in consecutive periods by financial intermediaries "performance persistence". Outperformance and performance persistence may have various sources. Apart from variables related directly to human factors, the relevant literature mentions also fund attributes, such as fund size, which can be based on inflow of assets to a fund. It is assumed that large funds, characterized by higher popularity among clients, can employ more skilled, better educated and more experienced managers, whom they will be able to pay more, and the hard-working managers will ensure persisting outperformance in exchange.

The study is an introduction to evaluating the effectiveness of funds operating in a developing market and provides a basis for further surveys and analyses in this area. The aim of this paper is to evaluate mutual funds' performance in the context of examining the efficiency, including its potential inertia over time, and the use of the phenomenon of economies of scale related to assets inflow to a fund. Generally, it is important to determine whether mutual funds are able to provide benefits to their clients and if their performance is a consequence of certain market circumstances. Hence, the analysis of the returns generated by collective investment institutions is particularly significant from the viewpoint of verifying the efficient market hypothesis. Additionally, a distinctive feature of this research is the application of an approach that has still been unpopular in the area of finance, consisting in construction of the Markov chain.

It needs to be also emphasized that the discussed subject matter brings utility values. Evidence of certain dependencies might influence the investment decisions made by individual investors by suggesting a probable potential of effective management of the assets entrusted to financial intermediaries. Mutual funds themselves could reproduce the information about having the appropriate attributes in the media in order to attract new clients.

This paper is composed of five sections. Part two presents a brief review of studies in the area of the discussed issues in the context of evaluating mutual funds' performance. Part three, which is a methodological section, describes the employed research approaches and data used in the analysis. In part four, empirical findings are reported. And the final section consolidates and summarizes the most significant results of the presented research.

2. Previous research

The literature review is focused on identifying a research gap in the area of capital allocation efficiency evaluation. The earliest studies include, for example, works by Sharpe (1964), Lintner (1965), and Mossin (1966), authors of the capital asset pricing model (CAPM). Successive researchers introduced modifications to this model in order to verify various hypotheses, including ones concerned with determining managers' skills as regards selecting securities for investment portfolios (e.g. Fama and French, 1993; Elton et al., 1996; Fung and Hsieh, 2004).

The published studies provided evidence in favour of the assumption that investment portfolio managers were incapable of generating abnormal returns. For instance, Jensen (1968) noted a certain predictability of investment returns. It concerned achievement of worse performance than the benchmark. As regards later studies, the research by Friend et al. (1970) as well as Henriksson (1984), who noticed the impossibility to obtain results exceeding a certain assumed benchmark, are worth mentioning. The results, in accordance with the efficient market hypothesis, could be observed with the use of risk-adjusted returns and, potentially, allowing for fee-adjusted returns. The publication that shed a new light on the findings of those days was the study by Grinblatt and Titman (1994). The authors emphasized that the evaluation of mutual funds' performance was extremely sensitive to the selection of a stock market index, treated as a benchmark.

More recent studies provided other performance measurement tools (e.g. Ferson and Schadt, 1996) or new research approaches, e.g. Bayesian methods (e.g. Huij and Verbeek, 2007) and bootstrap techniques (e.g. Huij and Derwall, 2008). One of the rare papers evaluating the performance of mutual funds by means of univariate and multivariate regime-switching models was the study by Ayadi et al. (2018). They applied the Markov chain procedure in the Treynor-Mazuy timing model in order to obtain reliable inferences on the market timing ability of Canadian fixed-income fund managers. The authors established that the regime-switching model was superior to univariate models due to the dynamic market conditions and cross-correlations of the funds' portfolios. Their conclusions regarding performance evaluation were multi-threaded.

Nevertheless, the main strand of the relevant literature at the turn of the 21st century was analyses dedicated to performance persistence. It can be defined as an increased propensity for relative repetition of mutual fund performance in consecutive periods. The empirical research carried out in the mid-1990s (e.g. Hendricks et al., 1993) was the first to suggest a relative stability of mutual funds' returns. Other studies additionally attempted to establish whether performance persistence was related to managers' characteristics (e.g. Du et al., 2009) or selection of portfolio components

(e.g. Grinblatt and Titman, 1992). Moreover, researchers asked the question if the mutual funds' performance persistence was a group phenomenon consisting in adopting a common winning investment strategy (cf. Goetzmann and Ibbotson 1994). For instance, Hendricks et al. (1993), who were mentioned above, identified the so-called hot hands effect concerning short-term performance persistence. They proved that, as a general rule, funds generating lower quarterly returns in one-year repeated performance below the benchmark in 4 successive quarters. In the case of winning funds, they found poor evidence for performance persistence in the next period. The persistence was noted also in the medium term, yet it was not as strong as one-year persistence.

More recent studies tried to engage stochastic methods for modeling the dynamics of risk-adjusted performance. One of them was the paper by Fenech et al. (2013), where investment rating migrations of Australian pension funds were measured by means of the Markov approach. The researchers investigated mobility matrices and found that the rating method mattered in terms of both statistics and investment decisions. In turn, Drakos et al. (2015), who also applied the Markov chain, examined whether there was a higher probability for mutual funds to remain in their initial ranking position compared to the probability of funds being characterized by a certain movement in ranking positions. They noted that there was a tendency for repeating performance in the post-ranking periods although the degree of mobility increased over time. Overall, the analyzed U.S. mutual fund market was characterized by a considerable degree of mobility. In summary, no straightforward conclusion was drawn. Performance persistence has been considered a market anomaly to this date.

The finance literature is also evidenced a significantly positive relationship between mutual fund flows and future performance. One of the first authors to discern this phenomenon was Gruber (1996), who noticed that investors had the ability to select funds which would be able to achieve superior performance in the next period. It means that mutual funds with net inflows outperform those with net outflows. As regards investors themselves, it was suggested that there might be informed investors capable of forecasting future investment results based on the information about past returns, who put their savings in funds with better future performance. Similarly, Zheng (1999) confirmed the relation and indicated that funds which received greater net flows outperform their less popular peers in the next period. Both these studies introduced the term "smart-money" effect to the relevant literature and defined it as mutual fund investors' ability to predict short-term performance and invest by moving money from underperformers to funds with better investment results (cf. Wermers, 2003; Sapp and Tiwari, 2004).

However, further studies (e.g. Frazzini and Lamont, 2006; Friesen and Sapp, 2007) noticed, contrary to what Gruber and Zheng argued, that a large group of investors

were less informed and less sophisticated than it would seem. Their activities in the form of investments in funds generated poor performance in the long run. The mentioned authors stated that fund net flows resulted in the so-called "dumb money" effect and investors themselves had low timing abilities, i.e. an average individual investor made wrong investment decisions most of the time. Teo and Woo (2001) also obtained evidence of the "dumb money" effect, which was reflected in high inflow funds underperforming low inflow funds over multi-year time periods. To the authors' best knowledge, there are very few papers in the smart money area to use stochastic methods. One of them is the study by Steffi Yang (2004), who developed a Markovian model of smart money chasing past winning funds. It was found that investors were sensitive to fund performance, in particular when funds could beat the market.

It should be highlighted that empirical investigations concerning mutual funds from the CEE countries where researchers use Markov-switching models are scare. The authors are familiar with only one paper in the relevant Polish (cf. Włodarczyk and Skrodzka, 2013) and one Romanian (Badea et al., 2019) literature analyzing efficiency of a limited number of local investment funds. Therefore, this study tries to fill the existing research gap and provides an opportunity to verify the potential market anomalies in developing economies, which might differ from the ones encountered in developed ones.

3. Data and methodological background

3.1. The scope and sources of data

The data used in this study was derived from a database which was created for the purpose of a research project conducted earlier and has still been updated. The data sources include publically available and specifically ordered information coming from reports prepared by the Chamber of Fund and Asset Management (IZFiA) and AnalizyOnline, respectively. Information about 46 Polish open-end mutual funds operating continuously between January 2010 and September 2018 was gathered for the purpose of verifying specific research hypotheses. The study was conducted for a homogeneous group of domestic equity funds with defined investment objectives. The starting point in this research was the values of the quarterly rates of return achieved by these entities. The funds' performance was compared to the values of the rates of return on benchmark being the local index of the securities market (the Warsaw Stock Exchange Index). It enables the calculation of market-adjusted returns. We decided to use market-adjusted returns instead of three- or four-factor models because of the lack of an appropriate data library or topical and downloadable databases containing fundamental factors (i.e. size, value, momentum) in Poland. The decision to resign from weekly or monthly data arose from a relatively high number of observations

usually indicating extremely small differences in the value between weekly or monthly rates of return of funds and benchmark (cf. Grinblatt and Titman, 1989). The annual values of assets under management also provided useful information, yet only with respect to the possibility to classify funds to the appropriate subsample in terms of fund size. Furthermore, the additional data included the quarterly values of net asset inflows to the portfolio. Due to the constraints on the volume of this paper, it was decided to omit the well-known formulas of the applied measures: the market-adjusted return and the inflow rate.

In order to capture the characteristics of the analyzed study sample, we decided to present the applied variables for the entire study in brief. Summary statistics across subsamples or in yearly periods are available from the authors of this paper at request. A description of the variables is shown in Table 1.

Table 1. Summary statistics for the applied variables

Variable	Mean	St. Dev.	Median	Min	Max	Q1	Q3
Return	-0.005	0.040	-0.007	-0.164	0.202	-0.028	0.014
Size*	410.5	727.7	158.4	2.5	5,089.5	76.5	411.7
Flow	1.017	0.217	0.984	0.403	3.531	0.923	1.061

Note: * values are expressed in PLN million.

As shown in Table 1, the sample of mutual funds is characterized by underperformance – the mean value of return equals -0.005. However, quite a large deviation from the mean is noticeable. The dominance of several entities whose asset values are disproportionately higher than those of the remaining funds in total can be noticed in the entire Polish mutual fund market. Hence, the distribution of the size variable is moderately or even highly positively skewed with a leptokurtic distribution. The last variable, flow, is partly affected by the size factor. Therefore, as the market developed over successive years, assets inflow was observable.

3.2. Hypotheses and methodological approach

There are many methods for the evaluation of fund performance in the finance literature. This article presents an analytical approach that can be used for the effectiveness evaluation of a consequence of certain market circumstances. Our research procedure consists of two parts. First, the efficiency of investment funds was analyzed against the return on benchmark. In this case, efficiency means the difference between the returns of a given fund in period *t* and the returns on the main local market index in the same period. The main Warsaw Stock Exchange Index (WIG) was employed as the benchmark. A similar approach was applied in some publications in finance. For instance, Abdymomunov and Morley (2011) examined time variation in multifactor models of asset returns, especially book-to-market and momentum

mimicking portfolios across stock market volatility regimes. As indicated previously, they employed market and portfolio returns using a two-state Markov-switching process.

However, the development of this approach served the purpose of determining whether the potential (in)efficiency was the domain of large or small funds. For this purpose, the entire sample was divided into two subsamples. The first of them, concerning large funds, was composed of 10% of the biggest entities in terms of assets held, whereas the second subsample covered the remaining 90% of entities, i.e. smaller funds. It was decided not to apply the traditional division into large vs. small with the use of the median or, for instance, the first and the last quartile due to a relatively significant asymmetry in the statistical distribution of the fund size's data. The approach consisting in examining fund efficiency among large and small entities separately was drawn from the study by Lee and Ward (2001), who investigated the relationship between past and present performance of UK real estate with the use of the Markov chain approach. This leads us to the initial two hypotheses:

Hypothesis 1: There is a tendency to uniformity of transition probabilities across states in regard to the obtained abnormal returns in two consecutive periods.

Hypothesis 2: The transition probabilities across states are the same for performance, regardless of the size of the fund (i.e. identical for large and small funds).

The second part of this research is dedicated to the issue discussed in the literature as the performance anticipation hypothesis (cf. Alves and Mendes, 2011). In this case, the authors attempted to answer the question whether abnormal returns were related to the mutual fund flow in the previous period. It was decided to analyze the flow–performance relationship by means of a four-state process. As was already mentioned, the authors found no papers in the smart money area using probabilistic methods and therefore they consider this as their contribution to the literature. Therefore, the final hypotheses read:

Hypothesis 3: There are equal transition probabilities of being visited across states in regard to prior net flows and the subsequent abnormal returns.

Hypothesis 4: There is no relationship between past net flows and future performance.

With regard to the first three hypotheses, it was decided to check the stationary distribution, which could be interpreted as a long run stability of the process. Therefore, the question of how many quarterly periods mutual funds need in order to attain the steady state will be crucial. Moreover, on the basis of observations concerning the issue in question, it is possible to determine whether the employed research procedures are helpful in explaining fund returns.

Taking into account the review of the applied empirical methods discussed in the literature, it was decided to adopt a probabilistic approach, which could be a natural and logical consequence of formulated research questions and the relevant hypotheses.

The chosen research approach was the Markovian framework (e.g. Kemeny and Snell, 1976). The Markov chain is defined as a special stochastic process with a countable state space and transitions at integer times. It could be said that a process $X = (X_t)_{t=1}^{\infty}$ is a Markov chain with the state space S if it takes value in set S and for every $n \in N$, for every $s_1, \ldots, s_n, s_{n+1}$, and for every $t \in \{n, n+1, n+2, \ldots\}$ we have that

$$P(X_{t+1} = s_{n+1} \mid X_t = s_n) = P(X_{t+1} = s_{n+1} \mid X_t = s_n, X_{t-1} = s_{n-1}, \ldots, X_1 = s_1). \quad (1)$$

From equation (1) we have an immediate interpretation of the Markov chain. Knowing the present state, it can be seen that the past of the process does not provide any further information about its future.

A crucial aspect in dealing with a Markov chain is its transition matrices, i.e. $\mathbf{P}_t$. Each element of transition matrix $\mathbf{P}_t$ corresponds to the estimated probability of transiting from state i to state j across states in t steps. Moreover, it can be said that a Markov chain with transition matrices $\mathbf{P}_t$ is homogeneous if $\mathbf{P}_t$ does not depend on t. More precisely, there exists a matrix $\mathbf{P}$ such that for every t we have that $\mathbf{P}=\mathbf{P}_t$.

A very important type of Markov chains is ergodic Markov chains. These are homogeneous Markov chains for which there exists the so-called stationary distribution, i.e. such a distribution π on S for which we have that

$$\pi = \pi \mathbf{P}. \quad (2)$$

The interpretation of equation (2), i.e. the meaning of an ergodic Markov chain, is very deep. In fact, for ergodic Markov chains, the dependence between being in a state and the initial probability (the choice of initial probabilities) decreases. In particular, there exists a limit of $\mathbf{P}^n$.

Ergodicity could be defined as the so-called metric transitivity, which is a property of indecomposable measure preserving transformations (cf. Poitras and Heaney, 2015). The abovementioned terms are related to the ergodic hypothesis (EH), which is applied in physics and thermodynamics. Ergodicity is a feature which helps us see that our process has a sort of stability in the long run. In particular, there is in ultimate loss of dependence on the initial states. More precisely, there exists such a limit

$$\mathrm{P} = \lim_{n \to \infty} \mathbf{P}^n$$

that

$$\mathrm{P} = \begin{bmatrix} \pi_1 & \pi_2 & \cdots & \pi_m \\ \cdots & \cdots & \cdots & \cdots \\ \pi_1 & \pi_2 & \cdots & \pi_m \end{bmatrix},$$

where π is a stationary distribution.

In our models, empirical distribution is used for calculating transition probabilities for every process. It is also assumed that the process under examination is homogeneous Markov chains. This assumption facilitates the study of the features of the returns generated by mutual funds. Moreover, it helps the authors find interesting properties of the returns since the applied approach is able to characterize more states of nature than the discussed dichotomies, e.g. outperformance vs. underperformance. The empirical distribution of funds' returns with respect to four subsequent quarters is taken into account.

However, restricting this study only to homogeneous Markov chains imposes major limitations on the authors. Firstly, it is assumed that – knowing the current rate of return of the mutual fund – it can be inferred that there is no significant dependence between past and future states of the analyzed processes. Secondly, the restriction imposed by homogeneity is the situation where there are the same probabilities of changing a state at every time moment. It should be also remembered that the employment of the Markov approach will not provide the information about the power or value of the transition from one state to another; it will only present the direction of the transition. Determining the probability of a change of a state informs the fund customer about what can be expected given specific initial assumptions.

4. Empirical results

As was mentioned, the aim of this study was to evaluate mutual funds' performance in the context of examining the efficiency, including its potential inertia over time, and the use of the phenomenon of economies of scale related to asset inflow to a fund. Hence, it was decided to divide this section into two respective parts. The relevant hypotheses will be verified in successive parts.

4.1. Difference between small and large mutual fund returns

This part presents the results of an analysis of performance in subsamples of small funds (see matrix A) and large funds (see matrix B). The authors try to determine the probabilities of obtaining abnormal returns in the two groups of funds and whether the probabilities differ.

Denote by X_t and Y_t the random variables which take values in the set {1, 2}. The events $\{X_t = 1\}$ and $\{Y_t = 1\}$ mean that small and large funds, respectively, outperform a benchmark. It is assumed that the processes $X = (X_t)_{t=1}^{\infty}$ and $Y = (Y_t)_{t=1}^{\infty}$ are homogeneous Markov chains with estimated transition matrices:

$$A=(a_{ij})_{i,j=1,2}=(P(X_t=i \mid X_{t-1}=j))_{i,j=1,2}=\begin{bmatrix} 0.477 & 0.523 \\ 0.351 & 0.649 \end{bmatrix}$$

and

$$B=(b_{ij})_{i,j=1,2}=(P(Y_t=i \mid Y_{t-1}=j))_{i,j=1,2}=\begin{bmatrix} 0.403 & 0.597 \\ 0.370 & 0.630 \end{bmatrix}.$$

The elements of transition matrices A and B correspond to the estimated probability of transiting from state i to state j across states (n = 2). The first state indicates outperforming funds and the second one – underperforming funds.

The results obtained for both subsamples point to a relative non-uniformity of transition probabilities across states. In the groups of both small and large entities, the probabilities that negative returns will persist are at similar levels (0.630-0.649). This may be related to the existence of the icy hands effect in the performance of Polish mutual funds, which consists in maintaining a portfolio generating a rate of return below the average in consecutive periods (cf. Urbański, 2017; Zamojska, 2011). Minor, yet observable, differences in the probabilities of transiting from the state of positive results to the state of negative market-adjusted returns were recorded among small funds (0.522) and, which was more visible, among large funds (0.597). The findings, termed as underperformance, correspond well with the efficient market theory. Regardless of whether a fund belonged to the group of large or small funds, having obtained outperformance, it definitely more frequently underperformed than repeated its superior returns in the subsequent period. Therefore, Hypothesis 1 needs to be rejected. Moreover, the values of transition probabilities for small and large funds are to a large extent comparable. Therefore, Hypothesis 2 should probably be confirmed.

These two Markov chains are ergodic and their stationary distributions are:

$$\pi^X = [\ \pi_i^X]_{i=1,2} = [0.402 \qquad 0.598],$$
$$\pi^Y = [\ \pi_i^Y]_{i=1,2} = [0.381 \qquad 0.619],$$

respectively.

The interpretation of the values of stationary distributions are as follows. In the long run, the probability that a small fund will be effective is 0.402. Moreover, 18 periods are needed to be near the steady state. For large funds, the situation is similar but the probability that it will be effective in the long run is slightly lesser, i.e. 0.381. Moreover, around 11 periods are needed to be near the steady state. In both cases, it can be seen that the probability that small or large funds will be effective in the long run is distinctly lower than 1/2.

4.2. Smart money effect

The second part of the study investigates whether inflow or outflow is related to beating the market. Precisely, the authors attempt to find the probability of achieving

abnormal returns, measured by means of a market-adjusted return. In this case, they build a Markov chain and try to find an indirect relation between the net flow and performance by computing adequate probabilities.

In the first step, we denote by X_t the random variable which takes values in the set {1, 2, 3, 4}. The events $\{X_t=1\}$, $\{X_t=2\}$, $\{X_t=3\}$ and $\{X_t=4\}$ mean: funds that were outperforming and registered inflows of money; outperforming funds that registered outflows of money; underperforming funds that registered inflows of money; and underperforming funds that registered outflows of money, respectively. We assume that the process $X=(X_t)_{t=1}^{\infty}$ is a homogeneous Markov chain with a transition matrix:

$$C=(c_{ij})_{i,j=1,2,3,4}=(P(X_t=i \mid X_{t-1}=j))_{i,j=1,2,3,4}=\begin{bmatrix} 0.326 & 0.229 & 0.286 & 0.159 \\ 0.134 & 0.270 & 0.344 & 0.251 \\ 0.143 & 0.148 & 0.300 & 0.409 \\ 0.147 & 0.274 & 0.168 & 0.411 \end{bmatrix}.$$

The elements of transition matrix C correspond to the estimated probability of transiting from state i to state j across states ($n = 4$).

The results show that transition probabilities are not uniform. In the case of funds characterized by net inflows, positive returns in the subsequent quarter coincide with a relevant customer response in the form of another asset inflow. For entities implementing such scenarios, the probability of positive performance persistence was 0.326. On the other hand, the probability that asset inflow after a worse period for funds does not mean that positive returns will be recorded later is similar (0.286). As regards funds with turbulence in asset inflow and outflow, i.e. increased redemptions in the previous period, but in the face of obtaining better results in the subsequent quarter, which could also result in a simultaneous inflow of new assets, the noticed probability of deteriorated performance in the subsequent period was 0.344. In the case of negative market-adjusted returns, in turn, regardless of whether there had been asset inflow or outflow, a definite (i.e. the probability was 0.411) outflow of assets from the fund and consistent persistence of negative returns was noticed. This finding can be supported with a higher number of observations assigned to the abovementioned states. These findings seem to be consistent with the results of the first part of the analysis, where funds were usually ineffective (e.g. Perez, 2012). Therefore, they do not permit a straightforward conclusion whether the smart money effect is present in the performance of Polish mutual funds. Nevertheless, certain regularities related to

a relatively strong sensitivity of fund customers to the investment performance achieved by the available forms of investment should be noticed.

The mentioned Markov chain is ergodic and its stationary distribution is as follows:

$$\pi = [0.177 \quad 0.231 \quad 0.265 \quad 0.327].$$

The interpretation of the stationary distribution of the above π is as follows. In the long run, there is no state which has a dominant probability. However, probability of the ineffectiveness of a fund is slightly greater than 1/2. Moreover, the probability that a fund will register outflow of money in the long run is also slightly higher than 1/2. The authors' calculations show that a fund needs about 21 periods to reach stationary distribution.

In order to verify Hypothesis 4, the relationship between past net flows and subsequent performance had to be checked by means of a probability matrix which is not a Markov chain. We denote by X_t the random variable which takes value 1 if a fund is outperforming at time t and -1 if it is underperforming, i.e. if its return is greater than the return of a market portfolio (WIG) and lower than that of WIG, respectively. Denote by Y_t the random variable which takes value 1 and -1 if the fund registers inflow and outflow of money at time t, respectively.

We assume that

$$D=(d_{ij})_{i,j=1,-1}=(P(X_t=i \mid Y_{t-1}=j))_{i,j=1,-1} = \begin{bmatrix} 0.400 & 0.600 \\ 0.407 & 0.593 \end{bmatrix}.$$

In particular, $d_{-1-1} = P(X_t = -1 \mid Y_t = -1) = 0.593$ and $d_{1-1} = P(X_t = 1 \mid Y_t = -1) = 0.600$ mean that if there is inflow or outflow of money at time t-1, then we will have that a fund return is worse than the market portfolio return with the probability of 0.6. In other words, neither purchases nor redemptions of unit shares were able to reverse this unfavorable regularity. Given such considerations, the results seem to be consistent with the effects of the examination of the smart money effect presented in matrix C, and at the same time they confirm the impossibility to generate abnormal returns, which is also consistent with the efficient market theory. Hence, Hypothesis 4 about the lack of the relationship between past net flows and subsequent performance might be confirmed. Nevertheless, the relationship should be also analyzed by means of other research methods.

5. Conclusions

The aim of this study was to evaluate the performance of Polish mutual funds in relation to the examination of efficiency, including its potential inertia over time, and the use of the phenomenon of economies of scale resulting from the net flow of assets.

The study sample consisted of 46 Polish domestic equity funds operating between January 2010 and September 2018. A Markovian framework was applied as the research approach. It was decided to use Markov chains in order to verify the formulated hypotheses. The results provide conclusions consistent with the efficient market theory. A certain inertia of returns concerning non-uniformity of transition probabilities across states, which results from the applied measure of return, could be observed. It is worth mentioning that the calculation of the market-adjusted returns revealed the lack of efficiency both in small and large funds with a slightly higher probability. Moreover, the discussed smart money effect was not detected in the present study, but the dominant funds were those which achieved poorer performance after asset inflow or outflow, regardless of the initial state. Hence, the issue of chasing performance by investors seems to be noticeable. In all cases, 11 and more periods were needed to reach stationary distribution.

In general, it was concluded that mutual funds operating in the analyzed developing market were unable to provide abnormal benefits to their clients. When market-adjusted returns were used, there was a higher probability of achieving underperformance than outperformance in relation to the initial state. Furthermore, the analyzed market circumstances, e.g. the smart money effect, which is recorded in the existing literature, was not confirmed here. The findings should be important for the theory of finance, especially from the viewpoint of verifying the efficient market hypothesis. The results could be also interesting to individual investors in the context of their investment decisions. At the same time, the utility value coming from the study does not seem to be very optimistic for mutual funds and their clients.

It should be emphasized that the study contributes to the current research by applying the Markovian framework. The constructed Markov chains are still unpopular in the field of finance, especially in relation to mutual funds operating in European markets. As was mentioned before, the authors made certain assumptions, e.g. a special stochastic process is homogeneous, which facilitated the study of the features of the rates of return. On the other hand, restricting this study only to homogeneous Markov chains imposed a number of limitations on the authors. Most importantly, past and future states of the examined processes are independent of each other. However, the employed approach proved helpful in explaining the returns generated by funds. The reasonability of its application in future studies in the discussed area also deserves a mention. One of the research perspectives that naturally come to mind in the first place could be that of a martingale approach in the evaluation of mutual fund performance. Subsequent studies should concentrate on the so-called stopping times, which – together with the Doob theorem – can help us calculate the probability that returns will reach fixed levels (cf. Devolder et al., 2012).

References

ABDYMOMUNOV, A., MORLEY, J., (2011). Time variation of CAPM betas across market volatility regimes, *Applied Financial Economics*, Vol. 21, pp. 1463–1478.

ALVES, C., MENDES, V., (2011). Does performance explain mutual fund flows in small markets? The case of Portugal, *Portuguese Economic Journal*, Vol. 10, pp. 129–147.

AYADI, M.A., LAZRAK, S., LIAO, Y., WELCH, R., (2018). Performance of fixed-income mutual funds with regime-switching models, *The Quarterly Review of Economics and Finance*, Vol. 69, pp. 217–231.

BADEA, L., ARMEANU, D.,S., PANAIT, I., GHERGHINA, S. C., (2019). A Markov Regime Switching Approach towards Assessing Resilience of Romanian Collective Investment Undertakings, *Sustainability*, Vol. 11, pp. 1–24.

DEVOLDER, P., JANSSEN, J., MANCA, R., (2012). *Stochastic Methods for Pension Funds*, ISTE Ltd, London and John Wiley & Sons, New York.

DRAKOS, K., GIANNAKOPOULOS, N., KONSTANTINOU, P., (2015). Investigating Persistence in the US Mutual Fund Market: A Mobility Approach, *Review of Economic Analysis*, Vol. 7, pp. 54–83.

DU, D., HUANG, Z., BLANCHFIELD, P. J., (2009). Do fixed income mutual fund managers have managerial skills? *The Quarterly Review of Economics and Finance*, Vol. 49, pp. 378–397.

ELTON, E. J., GRUBER, M. J., BLAKE, C., (1996). The Persistence of Risk-Adjusted Mutual Fund Performance, *Journal of Business*, Vol. 69, pp. 133–157.

FAMA, E. F., (1970). Efficient Capital Markets: A Review of Theory and Empirical Work, *Journal of Finance*, Vol. 25, pp. 383–417.

FAMA, E. F., FRENCH, K. R., (1993). Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, Vol. 33, pp. 3–56.

FENECH, J.-P., YAP, Y. K., SHAFIK, S., (2013). Brief Technical Note: A Markov Chain Approach to Measure Investment Rating Migrations, *Australasian Accounting, Business and Finance Journal*, Vol. 7, pp. 145–154.

FERSON, W., SCHADT, R. W., (1996). Measuring fund strategy and performance in changing economic conditions, *Journal of Finance*, Vol. 51, pp. 425–462.

FRAZZINI, A., LAMONT, O., (2006). Dumb money: mutual fund flows and the cross-section of stock returns. *NBER Working Paper* 11526.

FRIEND, I., BLUME, M. E., CROCKETT, J., (1970). Mutual Funds and Other Institutional Investors – A new perspective. New York: Mc Graw Hill Book Company.

FRIESEN, G., SAPP, T., (2007). Mutual fund flows and investor returns: an empirical examination of fund investor timing ability, *Journal of Banking & Finance*, Vol. 31, pp. 2796–2816.

FUNG, W., HSIEH, D. A., (2004). Hedge Fund Benchmarks: A Risk-Based Approach, *Financial Analysts Journal*, Vol. 60, pp. 65–80.

GOETZMANN, W. N., IBBOTSON, R. G., (1994). Do Winners Repeat? *Journal of Portfolio Management*, Vol. 20, pp. 9–18.

GRINBLATT, M., TITMAN, S., (1989). Mutual Fund Performance: An Analysis of Quarterly Portfolio Holdings, *Journal of Business*, Vol. 62, pp. 393–416.

GRINBLATT, M., TITMAN, S., (1992). The Persistence of Mutual Fund Performance, *Journal of Finance*, Vol. 7, pp. 1977–1984.

GRINBLATT, M., TITMAN, S., (1994). A Study of Monthly Mutual Fund Returns and Performance Evaluation Techniques, *Journal of Financial and Quantitative Analysis*, Vol. 29, pp. 419–444.

GRUBER, M., (1996). Another puzzle: The growth in actively managed mutual funds, *Journal of Finance*, Vol. 51, pp. 783–810.

HENDRICKS, D., PATEL, J., ZECKHAUSER, R., (1993). Hot hands in mutual funds: Short-run persistence of relative performance, 1974-1988, *Journal of Finance*, Vol. 48, pp. 93–131.

HENRIKSSON, R. D., (1984). Market Timing and Mutual Fund Performance: An Empirical Investigation, *The Journal of Business*, Vol. 57, pp. 73–96.

HUIJ, J., DERWALL, J., (2008). "Hot Hands" in bond funds, *Journal of Banking & Finance*, Vol. 32, pp. 559–572.

HUIJ, J., VERBEEK, M., (2007). Cross-sectional learning and short-run persistence in mutual fund performance, *Journal of Banking and Finance*, Vol. 31, pp. 973–997.

JENSEN, M., (1968). The Performance of Mutual Funds in the Period 1945-1964, *Journal of Finance*, Vol. 23, pp. 389–416.

KEMENY, J. G., SNELL, L. J., (1976). Finite Markov Chains. With a New Appendix "Generalization of a Fundamental Matrix". Springer-Verlag, New York-Berlin-Heidelberg-Tokio.

LEE, S.L., WARD, C. W. R., (2001). Persistence of UK Real Estate Returns: A Markov Chain Analysis, *Journal of Asset Management*, Vol. 1, pp. 279–291.

LINTNER, J., (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock, Portfolios and Capital Budgets, *The Review of Economics and Statistics*, Vol. 47, pp. 13–37.

MOSSIN, J., (1966). Equilibrium in a Capital Asset Market, *Econometrica*, Vol. 34, pp. 768–783.

PEREZ, K., (2012). Persistence in performance of Polish mutual funds (in Polish), *Finanse*, Vol. 1, pp. 81–113.

POITRAS, G., HEANEY, J., (2015). Classical Ergodicity and Modern Portfolio Theory, *Chinese Journal of Mathematics*, Article ID 737905, pp. 1–17.

SAPP, T., TIWARI, A., (2004). Does stock return momentum explain the 'smart money' effect? *Journal of Finance*, Vol. 59, pp. 2605–2622.

SHARPE, W. F., (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, *Journal of Finance*, Vol. 19, pp. 425–442.

STEFFI YANG, J.-H., (2004). The Markovian Dynamics of "Smart Money", *Far Eastern Meetings* from Econometric Society, No 797.

TEO, M., WOO, S.-J., (2004). Style effects in the cross-section of stock returns, *Journal of Financial Economics*, Vol. 74, pp. 367–398.

URBAŃSKI, S., (2017). Short-, medium- and long-run performance persistence of investment funds in Poland, *Bank i Kredyt*, Vol. 48, pp. 343–374.

WERMERS, R., (2003). Is Money Really "smart"? New Evidence on the Relation Between Mutual Fund Flows, Manager Behavior, and Performance Persistence. Working paper, University of Maryland.

WŁODARCZYK, A., SKRODZKA, W., (2013). Modelling Decision-Making Processes on The Mutual Funds Market Using Switching Treynor-Mazuy Model (in Polish), *Zarządzanie i Finanse*, Vol. 4, pp. 211–226.

ZAMOJSKA, A., (2011). Empirical verification of persistence performance of Polish equity fund (in Polish), *Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu*, No. 183, pp. 482–490.

ZHENG, L., (1999). Is Money Smart? A study of Mutual Fund Investors' Fund Selection Ability, *Journal of Finance*, Vol. 54, pp. 901–933.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 131–144, DOI 10.21307/stattrans-2021-007
Received – 25.01.2020; accepted – 22.11.2020

The construction and analysis of repeated measurement designs (RMD) using the trial and error method

Shakeel A. Mir[1], **Immad A. Shah**[2]

ABSTRACT

Repeated measurement designs prove broadly applicable in almost all branches of bio-sciences, including agriculture, animal husbandry, botany, zoology. Unbiased estimators for elementary contrasts among direct and residual effects are obtainable in this class of designs, which is considered their important property. In this paper, an attempt was made to provide a new method of overcoming a drawback in the construction method developed by Afsarinejad (1983), where one or more treatments may occur more than once in certain sequences causing the constructed designs to no longer remain uniform in the examined periods. Nine designs were constructed and presented jointly with their corresponding mathematical analyses.

Key words: residual effects, order effects, balanced minimal RMD.

1. Introduction

In many horticultural experiments, for example in high density plantation trials, it is imperative that some treatments like fertilizers, insecticides, etc., applied to the crop are not fully utilized by the crop and the unutilized portion called the residual of these applied treatments results in a carry-over effect, which is an effect that "carries over" from one experimental condition to another. Whenever subjects perform in more than one condition (as they do in within-subject designs), there is a possibility of order effects. In such situations, the biggest drawbacks known as the order effects are caused by exposing the subjects to multiple treatments. Order effects are related to the order in which treatments are given but not due to the treatment itself. Order effects can interfere with the analysis and also challenge the ability to correctly estimate the effect of the treatment itself. To overcome the residual effect of the treatment, the experimenter can introduce a rest period and allow the experimental units to even up.

[1] Professor & Head, Division of Agricultural Statistics, SKUAST-Kashmir, India. E-mail: mir_98@msn.com. ORCID: https://orcid.org/0000-0001-7221-9835.

[2] Ph.D. Statistics/SRF, Division of SAF, Faculty of Forestry, SKUAST-Kashmir, India. E-mail: immad11w@gmail.com. ORCID: https://orcid.org/0000-0003-2761-5112.

Many researchers such as Cochran and Cox (1986), Sheeh and Bross (1961),Westlake (1974), Afsarinejad (1989) and Afsarinejad (1990) have pointed out that if residual effects exist, then the methods applicable to conventional designs are not valid. The experimenter has to design the experiment such that the unbiased estimator for elementary contrasts among direct and residual effects can be obtained. In such situations the repeated measurement designs help us to estimate the order effects, thereby increasing the precision of estimates. Patterson (1952) introduced the concept of using the differences for constructing repeated measurement designs and used the method to construct the designs suggested by Williams (1949). In another attempt Bradley (1958), and Sheehe and Bross (1961) constructed the designs for odd and even number of treatments by using easy algorithms. Gorden (1961) used the group theory to form balanced RMD's. Hedayat and Afsarinejad (1975) defined repeated measurement designs balanced for direct and residual effects. They gave construction methods for balanced minimal RMD. In 1978, they gave the concept of a universal optimal design, completely symmetric design, uniform design and balanced uniform design. Pigeon and Raghavarao (1987) introduced a class of crossover designs known as control balanced residual effect design. They gave their structure in detail and also described the method of construction of these designs. Kenward and Jones (1987) proposed a method for the construction of log-linear models for binary data from cross-over designs. Repeated measurement designs have application in various fields of science such as agricultural science, animal science, medical science and engineering, etc. In intensive agricultural systems it is imperative that some treatments like fertilizer, pesticides, etc., applied to a crop are not fully utilized by the crop and the unutilized portion called the residual of those applied treatments is utilized by the subsequent crop. In such cases, the RMD plays a significant role to study the residual effect. Similarly, in the case of animal experiments, the severe restriction of numbers of animals leaves few degrees of freedom for error, thereby reducing statistical power drastically or preventing multivariate analysis entirely. In such cases, the primary benefit of a repeated measurement design (RMD) is statistical power relative to the sample size, which is important in many real research studies. In clinical trials the data on the patients is recorded more than once. In such situations, using the standard ANOVA procedures is not appropriate as it does not consider dependencies between observations within subjects in the analysis. To deal with such types of study data repeated measurement designs are used.

1.1. General Model of Analysis

In RMD certain terms like "direct effect" and "residual effect" are used. Suppose we have't' treatments, which are to be tested and studied via n experimental units. Each experimental unit is used in p periods resulting in $r_i \geq 1$ observations for i^{th} treatment,

i.e. $r_1+r_2+r_3+ \ldots + r_t = np$. Let D denote the set of all such arrangements to which we shall refer as design in D, then let $d_{(i,j)}$ denote the treatment assigned by D in the i^{th} period to the j^{th} experimental unit. Let y_{ij} be response under $d_{(i,j)}$ treatment applied to the j^{th} unit in the i^{th} period and to analyse this set of observations $(y_{ij's})$ we take the model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \tau_{d(i,j)} + \rho_{d(i-1,j)} + e_{ij},\ i=1,2,\ldots,p\ ;\ j=1,2,\ldots,n$$

where

y_{ij} = the observation in the i^{th} period on j^{th} unit
μ = general effect
α_i = i^{th} period effect
β_j = j^{th} unit effect
$\tau_{d(i,j)}$ = direct effect of treatment $d_{(i,j)}$
$\rho_{d(i-1,j)}$ = first order residual effect of treatment $d_{(i-1,j)}$
e_{ij} = uncontrolled random errors which are normally distributed with mean zero and variance σ^2 (i.e. Homocedastic).

1.2. Mathematical Analysis of RMDs

In vector notation np response under the above model can be written as:

$$E(\underline{Y}) = X\left(\underline{\gamma}\right) = [X_1 X_2]\begin{bmatrix}\underline{\gamma}\\ \underline{\gamma}\end{bmatrix} = X_1\underline{\gamma}_1 + X_2\underline{\gamma}_2$$

Here, $\underline{\gamma}_1$ is the set of the parameters to be estimated and $\underline{\gamma}_2$ is the set of all other parameters in the model. The normal equations estimating the parameters are:

$$\begin{bmatrix}X_1'\\ X_2'\end{bmatrix}[X_1 X_2]\begin{bmatrix}\widehat{\underline{\gamma}_1}\\ \widehat{\underline{\gamma}_2}\end{bmatrix} = \begin{bmatrix}X_1'\underline{Y}\\ X_2'\underline{Y}\end{bmatrix}$$

$$\begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix}\begin{bmatrix}\widehat{\underline{\gamma}_1}\\ \widehat{\underline{\gamma}_2}\end{bmatrix} = \begin{bmatrix}X_1'\underline{Y}\\ X_2'\underline{Y}\end{bmatrix}$$

$$X_1'X_1\underline{\hat{\gamma}}_1 + X_1'X_2\widehat{\underline{\gamma}_2} = X_1'\underline{Y} \qquad (1)$$

$$X_2'X_1\underline{\hat{\gamma}}_1 + X_2'X_2\widehat{\underline{\gamma}_2} = X_2'\underline{Y} \qquad (2)$$

From normal equation (2) we get

$$\widehat{\underline{\gamma}_2} = (X_2'X_2)^{-1}\left[X_2'\underline{Y} - X_2'X_1\underline{\hat{\gamma}}_1\right] \qquad (3)$$

Putting (3) in equation (1) we get

$$\underline{\hat{\gamma}}_1 = [X_1'\{1 - X_2(X_1'X_2)^{-1}X_2'\}X_1]^{-1}\left[X_1'\{1 - X_2(X_2'X_2)^{-1}X_2'\}\underline{Y}\right] =$$
$$\left[C\left(\underline{\hat{\gamma}}_1\right)\right]^{-1}\left[Q\left(\underline{\hat{\gamma}}_1\right)\right]$$

where

$C\left(\underline{\hat{\gamma}}_1\right)$ is the information matrix associated with $\underline{\hat{\gamma}}$.

1.3. Definition due to Hedayat and Afsarinejad and the design constructed

Hedayat and Afsarinejad (1975), Hedayat and Afsarinejad (1978), Afsarinejad (1983) and Afsarinejad (1990) defined balanced repeated measurement design as: A *RMD (t, n, p)* based on t treatments, *n* experimental units each being used in *p* periods is said to be balanced with respect to the sets of direct treatment effects and the first order residual effects if:

a) Each treatment is tested equally frequently λ_1 times in each period.

b) In the order of application, each treatment is preceded by each other treatment equally frequently λ_2 times.

Clearly, in a balanced RMD (*t, n, p*), the following relation holds:

- $n = \lambda_1 t$
- $n(p-1) = \lambda_2 t(t-1)$

Afsarinejad (1983) gave a construction method for the balanced minimal repeated measurement designs for an odd number of treatments. RMD (15, 105, 3), $\lambda_2 = 1$ and RMD (21, 105, 5), $\lambda_2 = 1$, constructed by this method, have a drawback that treatments can occur more than once in the same sequences. A design constructed by the method given by Afsarinejad is given below in Table 1.

Table 1. RMD (15, 105, 3) given by Afsarinejad

	Initial sequences						
Treatments	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
	1	3	5	7	7	5	3
	15	13	11	9	11	13	15
	3	5	7	7	5	3	1

From the table, it is clear that the treatment "7" is occurring more than once in the 4th initial sequence of RMD (15, 105, 3), $\lambda_2 = 1$. Therefore, on development the final RMD (15, 105, 3), $\lambda_2 = 1$ will have t=15 sequences wherein the treatments will be occurring more than once. Keeping this drawback in view new repeated measurement designs are proposed wherein no sequence has any treatment occurring more than once and which are also suitable for an even number of treatments using the trial and error method of design construction.

1.4. Trial and Error Method

Trial and error is a fundamental method of problem-solving. It is neither a method of finding the best solution nor a method of finding all solutions. It is a technique that

is used simply to find a solution. In this method, a researcher tries the option that has the best possible chances of succeeding. If that didn't work, one can try the next best option until they find a good solution. It is characterized by repeated, varied attempts which are continued until success, or until the practiser stops trying. The trial and error approach is used most successfully with simple problems and in games, and it is often the last resort when no apparent rule applies. Trial and error method is solution-oriented i.e. it makes no attempt to discover why a solution works, merely that it is a solution. It is problem-specific and makes no attempt to generalize a solution to other problems. It is non-optimal and needs little knowledge.

2. New Designs Constructed by Trial and error Method.

The same design by Afsarinejad is constructed using the trial and error method and is a repeated measurement design (15, 105, 3), implying a design having 15 treatments, 105 experimental units and 3 periods. The construction is given in Table 2.

Table 2. RMD (15, 105, 3), $\lambda_2 = 1$

Treatments / Period	Initial sequences						
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**
I	1	3	5	7	2	4	6
II	15	13	11	9	14	12	10
III	3	5	7	10	4	6	8

From the table above it is clear that the repeated measurements design has 15 treatments, 105 experimental units and the experiment lasts for 3 periods. Each experimental unit receives one treatment during each period. As it stands, the design is a 3x105 array containing entries from t = { 1, 2, . . . , 15}. The design is developed cyclically starting from the initial sequences. It can be easily verified that the set of initial sequences given when developed in a cyclic manner, the RMD (15, 105, 3), so obtained is a balanced repeated measurement design according to the definition of Hedayat and Afsarinejad, and has no sequence wherein no treatment occurs more than once in a sequence as shown in Table 3.

Table 3. Design obtained on development sequence of RMD (15, 105, 3)

	Experimental Units														
PERIOD	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
II	15	1	2	3	4	5	6	7	8	9	10	11	12	13	14
III	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2
	16	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
	3	4	5	6	7	8	9	10	11	12	13	14	15	1	2
	13	14	15	1	2	3	4	5	6	7	8	9	10	11	12
	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4

Table 3. Design obtained on development sequence of RMD (15, 105, 3) (cont.)

	Experimental Units														
	31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
	5	6	7	8	9	10	11	12	13	14	15	1	2	3	4
	11	12	13	14	15	1	2	3	4	5	6	7	8	9	10
	7	8	9	10	11	12	13	14	15	1	2	3	4	5	6
	46	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
	7	8	9	10	11	12	13	14	15	1	2	3	4	5	6
	9	10	11	12	13	14	15	1	2	3	4	5	6	7	8
	10	11	12	13	14	15	1	2	3	4	5	6	7	8	9
	61	**62**	**63**	**64**	**65**	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1
	14	15	1	2	3	4	5	6	7	8	9	10	11	12	13
	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3
	76	**77**	**78**	**79**	**80**	**81**	**82**	**83**	**84**	**85**	**86**	**87**	**88**	**89**	**90**
	4	5	6	7	8	9	10	11	12	13	14	15	1	2	3
	12	13	14	15	1	2	3	4	5	6	7	8	9	10	11
	6	7	8	9	10	11	12	13	14	15	1	2	3	4	5
	91	**92**	**93**	**94**	**95**	**96**	**97**	**98**	**99**	**100**	**101**	**102**	**103**	**104**	**105**
	6	7	8	9	10	11	12	13	14	15	1	2	3	4	5
	10	11	12	13	14	15	1	2	3	4	5	6	7	8	9
	8	9	10	11	12	13	14	15	1	2	3	4	5	6	7

*Initial Sequences.

2.1. Other designs constructed by the proposed method

A repeated measurement design (10, 30, 3) implies a design having 10 treatments, 30 experimental units and 3 periods. The construction is given in Table 4. The subsequent designs are also given in the following tables.

Table 4. RMD (10, 30, 7), $\lambda_2 = 2$

Treatments / Period	Initial sequences		
	I	**II**	**III**
I	1	4	2
II	10	7	9
III	2	5	3
IV	9	6	4
V	3	1	7
VI	8	10	5
VII	4	2	1

Table 5. RMD (15, 105, 5), $\lambda_2 = 2, \lambda_1 = 7$

Treatments / Period	Initial sequences						
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**
I	1	13	6	1	12	6	1
II	15	4	10	14	4	9	11
III	2	12	7	2	11	7	5
IV	14	5	9	13	5	8	6
V	3	11	8	3	10	3	2

Table 6. RMD (22, 154, 7), $\lambda_1 = 7, \lambda_2 = 2$

Treatments / Period	Initial sequences						
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**
I	1	4	7	10	2	5	8
II	22	19	16	9	21	18	15
III	2	5	8	7	3	6	9
IV	21	18	15	8	20	17	14
V	3	6	9	11	4	7	10
VI	20	17	14	12	19	16	13
VII	4	7	10	14	5	8	11

Table 7. RMD (26, 130, 11), $\lambda_1 = 5, \lambda_2 = 2$

Treatments / Period	Initial sequences				
	I	**II**	**III**	**IV**	**V**
I	1	6	11	3	8
II	26	21	10	24	19
III	2	7	6	4	9
IV	25	20	3	23	18
V	3	8	1	5	10
VI	24	19	5	22	17
VII	4	9	7	6	11
VIII	23	18	12	21	16
IX	5	10	15	7	12
X	22	17	16	20	15
XI	6	11	17	8	13

Table 8. RMD (21, 105, 9), $\lambda_1 = 5, \lambda_2 = 2$

Treatments / Period	Initial sequences				
	I	**II**	**III**	**IV**	**V**
I	1	5	13	19	15
II	21	17	10	3	7
III	2	6	12	18	14
IV	20	7	11	4	8
V	3	15	21	17	13
VI	19	8	1	5	9
VII	4	14	20	16	12
VIII	18	9	2	6	10
IX	5	13	19	15	1

Table 9. RMD (22, 66, 13), $\lambda_2 = 2$

Treatments / Period	Initial sequences		
	I	**II**	**III**
I	1	8	4
II	22	15	19
III	2	9	5
IV	21	14	18
V	3	10	6
VI	20	13	7
VII	4	11	16
VIII	19	12	8
IX	5	1	15
X	18	22	9
XI	6	2	14
XII	17	21	10
XIII	7	3	3
XIV	16	20	11
XV	8	4	1

Table 10. RMD (28, 252, 7), $\lambda_2 = 2$

Treatments / Period	Initial sequences								
	I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**
I	1	4	7	10	13	27	24	8	11
II	28	25	22	19	16	2	5	20	17
III	2	5	8	11	14	26	23	9	12
IV	27	24	21	18	15	3	6	19	16
V	3	6	9	12	28	25	22	10	13
VI	26	23	20	17	1	4	21	18	15
VII	4	7	10	13	27	24	8	11	1

Similarly the above sequences can be developed cyclically to yield the balanced repeated measurement designs.

2.2. Analysis of the constructed repeated measurement designs

Let us consider a repeated measurement design in which 't' treatments are applied to 'n' experimental units during 'p' periods. Let the rows correspond to periods and columns to experimental units. For analysis of these *'pn'* observations, we assume the fixed effect model:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \tau_{d(i,j)} + \rho_{d(i-1,j)} + e_{ij} \qquad \text{(A)}$$

where *i=1,2,...,p and j=1,2,...,n*

Y_{ij} is the observation in the i^{th} period on the j^{th} experimental unit receiving $d(i,j)$ treatment.

μ is the general effect.

α_i is the i^{th} period effect.

β_j is the j^{th} experimental unit effect.

$\tau_{d(i,j)}$ is the direct effect of $d(i,j)$ treatment.

$\rho_{d(i-1,j)}$ is the first order residual effect of $d(i,j)$ treatment.

Assuming e_{ij} are homoscedastic with mean (0) and variance(σ^2). The information matrix associated with the entire set of parameters of model (A) can be rewritten as:

$$X'X = \begin{bmatrix} np & nE_{1p} & nE_{1p} & \underline{r}' & \underline{r}^{*\prime} \\ nE_{p1} & nI_p & M & L' & L^{*\prime} \\ pE_{n1} & M' & pI_n & N' & N^{*\prime} \\ \underline{r} & L & N & diag(\underline{r}) & S \\ \underline{r}^* & L^* & N^* & S' & diag(\underline{r}^*) \end{bmatrix}$$

Where E_{ab} is the a x b matrix with all entries as one.

I_p : identity matrix of order p

M : incidence matrix of order p x n for the period unit.

L : incidence matrix of order t x p for the direct effects-period.

N : incidence matrix of order t x n for the direct effectsunits.

L^* : incidence matrix of order t x p for the residual effects-period.

N^* : incidence matrix of order t x n for the residual effects-period.

S : incidence matrix of order t x t for the direct effects-residual effects.

$\underline{r}'$: $[r_1,\ r_2, \dots, r_t]$

$\underline{r}^{*\prime} = [r_1^*,\ r_2^*, \dots, r_3^*]$

$diag(\underline{r}) = diag(r_1, r_2, \dots, r_t)$

$diag(\underline{r}^*) = diag(r_1^*, r_2^*, \dots, r_t^*)$

r_i = No. of observations in which the i^{th} treatment was applied.

r_i^* = No. of treatments in which the i^{th} treatment was applied in the periods other than the last period.

The normal equations of the above model from the information matrix are as follows:

$$np\hat{\mu} + nE_{1p}\underline{\hat{\alpha}} + pE_{1n}\underline{\hat{\beta}} + \underline{r'\hat{t}} + \underline{r}^*\underline{\hat{\rho}} = G$$

$$nE_{p1}\hat{\mu} + nI_p\underline{\hat{\alpha}} + M\underline{\hat{\beta}} + L'\underline{\hat{t}} + L^{*\prime}\underline{\hat{\rho}} = \underline{A}$$

$$pE_{n1}\hat{\mu} + M'\underline{\hat{\alpha}} + pI_n\underline{\hat{\beta}} + N'\underline{\hat{t}} + N^{*\prime}\underline{\hat{\rho}} = \underline{B}$$

$$\underline{r}\hat{\mu} + L\underline{\hat{\alpha}} + N\underline{\hat{\beta}} + \left(diag\underline{r}\right)\underline{\hat{\tau}} + S\underline{\hat{\rho}} = \underline{T}$$
$$\underline{r^*}\hat{\mu} + L^*\underline{\hat{\alpha}} + N^*\underline{\hat{\beta}} + S'\underline{\hat{\tau}} + \left(diag\underline{r^*}\right)\underline{\hat{\rho}} = \underline{R}$$

Let

A_i : sum of all the observations in the i^{th} period.

B_i : sum of the observations on the j^{th} experimental unit.

T_k : sum of the observations of the k^{th} treatment.

R_p : sum of those observations on which the p^{th} treatment was applied in the immediately preceding period.

G : grand total of all 'np' observations.

Let

$$\underline{A} = \left[A_1, A_2, \ldots\ldots, A_p\right]'$$
$$\underline{B} = \left[B_1, B_2, \ldots\ldots, B_n\right]'$$
$$\underline{T} = \left[T_1, T_2, \ldots\ldots, T_t\right]'$$
$$\underline{R} = \left[R_1, R_2, \ldots\ldots, R_t\right]'$$

Under the condition that from each unit we have observations in each period (no observation is missing), giving rise to '***pn***' observations, then:

$$M = E_{pn}$$
$$E_{1p}M = \mathrm{p}E_{1n}$$
$$NE_{n1} = LE_{p1} = \underline{r}$$
$$NE_{n1} = LE_{p1} = \underline{r}$$
$$N^*E_{n1} = L^*E_{p1} = \underline{r}^*$$
$$E_{1t}N = pE_{1n}$$
$$E_{1p}\underline{A} = G$$
$$E_{1p}\underline{B} = G$$
$$E_{1t}\underline{T} = G$$
$$E_{tt}N\underline{B} = pE_{t1}G$$
$$E_{tt}N^*\underline{B} = (p-1)E_{t1}G$$
$$E_{1t}N^*\underline{B} = (p-1)G$$

From Equation: we have

$$\underline{\hat{\alpha}} = \frac{\left[\underline{A} - nE_{p1}\hat{\mu} - M\underline{\hat{\beta}} - L'\underline{\hat{\tau}} - L^{*\prime}\underline{\hat{\rho}}\right]}{n}$$

Put $\underline{\hat{\alpha}}$ in equation:

$$PE_{n1}\hat{\mu} + \frac{1}{n}M'\underline{A} - M'E_{p1}\hat{\mu} - \frac{1}{n}M'M\underline{\hat{\beta}} - \frac{1}{n}M'L'\underline{\hat{\tau}} - \frac{1}{n}M'L^{*\prime}\underline{\hat{\rho}} + pI_n\underline{\hat{\beta}} + N'\underline{\hat{\tau}} + N^{*\prime}\underline{\hat{\rho}} = \underline{B}$$

$$\left(pI_n - \frac{p}{n}E_{nn}\right)\underline{\hat{\beta}} + \left(N' - \frac{1}{n}M'L'\right)\underline{\hat{\tau}} + \left(N^{*'} - \frac{1}{n}M'L^{*'}\right)\underline{\hat{\rho}} = \underline{B} - \frac{1}{n}M'\underline{A}$$

$$\underline{\hat{\beta}} = \frac{1}{p}I_n\left(\underline{B} - \frac{1}{n}E_{np}\underline{A}\right) - \left(N' - \frac{1}{n}E_{np}L'\right)\underline{\hat{\tau}} - (N^* - \frac{1}{n}E_{np}L^{*'})\,\underline{\hat{\rho}}$$

From Equation substituting $\underline{\hat{\beta}}$ in equation:

$$\underline{\hat{\alpha}} = \frac{\underline{A}}{n} - E_{p1}\hat{\beta} - \frac{L'\underline{\hat{\tau}}}{n} - \frac{L^{*\prime}\underline{\hat{\rho}}}{n}$$

From the equations and substituting $\underline{\hat{\alpha}}$ and $\underline{\hat{\beta}}$ *in equation* gives:

$$\left[(diag.\underline{r} - \frac{1}{n}LL' - \frac{1}{p}NN' + \frac{1}{np}\,\underline{r}\,\underline{r}')\right]\underline{\hat{\tau}} + (S - \frac{1}{n}LL^{*'} - \frac{1}{p}NN^{*'} - \frac{1}{np}\underline{r}\,\underline{r}^{*'})\hat{\rho} = \underline{T} - \frac{1}{n}L\underline{A} - \left(\frac{1}{p}N - \frac{1}{np}\underline{r}E_{1n}\right)\underline{B}$$

From the equations and substituting $\underline{\hat{\alpha}}$ and $\underline{\hat{\beta}}$ *in equation* gives:

$$\underline{r}^*\hat{\mu} + \left(\frac{1}{n}L^* + \frac{1}{np}L^*E_{pp}\right)\underline{A} - L^*E_{p1}\hat{\beta} - \frac{1}{np}L^*E_{pn}\underline{\hat{\beta}} - \frac{1}{n}L^*L'\underline{\hat{\tau}} - \frac{1}{n}L^*L^{*\prime}\underline{\rho} + \frac{1}{p}N^*\underline{B} - \frac{1}{np}N^*E_{np}\underline{A} - \frac{1}{p}\left(N^*N' - \frac{1}{np}N^*E_{np}L'\right)\underline{\hat{\tau}} - \frac{1}{p}\left(N^*N^{*'} - \frac{1}{np}N^*E_{np}L^{*'}\right)\underline{\hat{\rho}} + S'\underline{\hat{\tau}} + \left(diag.\underline{r}^*\right)\underline{\hat{\rho}} = \underline{R}$$

or

$$(S' - \frac{1}{n}L^*L' - \frac{1}{p}N^*N' + \frac{1}{np}\underline{r}^*\underline{r}')\underline{\hat{\tau}} + \left(diag\underline{r}^* - \frac{1}{n}L^*L^{*'} - \frac{1}{p}N^*N^{*'} + \frac{1}{np}\underline{r}^*\underline{r}^{*'}\right)\underline{\hat{\rho}} = \underline{R} - \frac{1}{n}L^*\frac{A}{p} - (\frac{1}{np}N^* - r^*E_{1n})\underline{B}$$

Equations are the reduced normal equations for estimating the direct and residual effects of a general *RMD* (*t, n, p*). As a special case, let each treatment occurs *v* times in each row (period), then:

$$\underline{r} = pvE_{11}$$
$$\underline{r}^* = (pv - v)E_{t1}$$
$$\underline{r}\,\underline{r}' = p^2v^2E_{tt}$$
$$\underline{r}\,\underline{r}^{*'} = pv^2(p-1)E$$
$$E_{np}L^{*\prime} = (p-1)vE_{nt}$$
$$L^*\underline{A} = v(G - A_1)E_{t1}$$
$$L = vE_{tp}$$
$$L^* = v\left[O_{t,1}E_{t,p-1}\right]$$
$$LL' = pv^2E_{tt}$$

$$L^*L^{*\prime} = (p-1)v^2E_{tt}$$
$$E_{tp}L^{*\prime} = (p-1)vE_{tt}$$

Substituting these relations in equations respectively gives:

$$\underline{\hat{\alpha}} = \left(\frac{1}{n}+\frac{1}{np}E_{pp}\right)\underline{A} - E_{p1}\hat{\mu} - \frac{1}{np}E_{pn}\underline{B} - \frac{v}{n}E_{pt}\hat{\tau} - \frac{1}{n}L^{*\prime}\underline{\hat{\rho}}$$

$$\underline{\hat{\beta}} = \frac{1}{p}\underline{B} - \frac{1}{np}E_{np}\underline{A} - \left(\frac{1}{p}N' - \frac{v}{n}E_{nt}\right)\underline{\hat{\tau}} - \left(\frac{1}{p}N^{*\prime} - \frac{(p-1)v}{np}E_{nt}\right)\underline{\hat{\rho}}$$

$$\left(pvI_t - \frac{1}{p}NN'\right)\underline{\hat{\tau}} + \left(S - \frac{1}{p}NN^*\right)\underline{\hat{\rho}} = \underline{T} - \frac{1}{p}N\underline{B}$$

$$\left(S' - \frac{1}{p}N^*N'\right)\underline{\hat{\tau}} + \left[v(p-1)I_t - \frac{(p-1)v}{pt}E_{tt} - \frac{1}{p}N^*N'\right]\underline{\hat{\rho}} = \underline{R} - \frac{1}{p}N^*\underline{B} + \frac{1}{t}E_{t1}A_1 - \frac{1}{pt}E_{t1}G$$

Using the notation $C_{a/\beta,\gamma..}$ to denote the matrix for estimating the parameter vector of a linear model after eliminating the general effect and the parameter vectors, γ... and ignoring any other parameter in the model which are not listed in the subscript. From equations the coefficient matrices $C_{\tau/\alpha,\beta,..}$, $C_{\rho/\alpha,\beta,..}$ in the equation estimating the direct and residual effects are respectively:

$$C_{\tau/\alpha,\beta,\rho..} = C_{\tau/\alpha,\beta,..} - F_1\overline{C}_{\rho/\alpha,\beta}F_1'$$

where

$$C_{\rho/\alpha,\beta,\tau..} = C_{\rho/\alpha,\beta,..} - F_1'\overline{C}_{\tau/\alpha,\beta}F_1$$

$$C_{\tau/\alpha,\beta,..} = pvI_t - \frac{1}{p}NN'$$

$$C_{\rho/\alpha,\beta.} = v(p-1)I_t - \frac{(p-1)}{np}v^2E_{tt} - \frac{1}{p}N^*N^{*\prime}$$

And $$F_1 = S - \frac{1}{p}NN^{*\prime}$$

2.3. Inverse of circulant matrix

The estimate of $\underline{\hat{\gamma}}_1$is

$$[X_1'\{1 - X_2(X_1'X_2)^{-1}X_2'\}X_1]^{-1}\left[X_1'\{1 - X_2(X_2'X_2)^{-1}X_2'\}\underline{Y}\right] = \left[C\left(\underline{\hat{\gamma}}_1\right)\right]^{-1}\left[Q\left(\underline{\hat{\gamma}}_1\right)\right]$$

Where

$C\left(\underline{\hat{\gamma}}_1\right)$ is the information matrix associated with $\underline{\hat{\gamma}}$.

For estimating the parameter vector of a linear model we have:

$$C_{\tau/\alpha,\beta,..} = pvI_t - \frac{1}{p}NN'$$

$$C_{\rho/\alpha,\beta.} = v(p-1)I_t - \frac{(p-1)}{np}v^2E_{tt} - \frac{1}{p}N^*N^{*\prime}$$

And to subsequently estimate the parameters we need to find the inverse of $C\left(\underline{\hat{\gamma}}_1\right)$, which is circulant in nature. Hence, to find the inverse a direct method proposed by Lin Fuyong (2011) is used to get the inverse matrix of the circulant matrix. The elements of the inverse matrix are functions of zero points of the characteristic polynomial $g(z)$ and $g'(z)$of the circulant matrix.

3. Conclusion

The initial sequences for the constructed designs have been given in the form of tables. A desirable property of the designs constructed by trial and error is that in all the given sequences of treatments, no treatment occurs more than once in a sequence, whereas in the method of construction given by Afsarinejad (1983) for same parametric combinations in some sequences treatments occur more than once. A matrix approach to the mathematical analysis of the designs is given along with the coefficient matrices $C_{\tau/\alpha,\beta,..}$, $C_{\rho/\alpha,\beta,..}$ in the equation estimating the direct and residual effects. And to subsequently estimate the parameters we need to find the inverse of $C\left(\underline{\hat{\gamma}}_1\right)$, which is circulant in nature, a direct method proposed by Lin Fuyong (2011) is used. Although the elementary method to find the inverse can also be used, in the case of matrices of higher order, the method given by Lin Fuyong is a feasible one.

Acknowlegements

The authors acknowledge the Pakistan Journal of Statistics and Operational Research for publishing the previous research work titled "On Balanced repeated Measurement Designs", Vol. X No.3 2014, pp. 289-298.

References

FUYONG, L., (2011). The inverse of circulant matrix, Applied Mathematics and Computation, DOI: 10.1016/j.amc.2011.03.052.

AFSARINEJAD, K., (1983). Balanced repeated measurements designs, Biometrika, 70, pp. 199–204.

COCHRAN, W. G, COX, G. M., (1986). Experimental Designs, 2nd Ed. Wiley, New York.

WESTLAKE, W. J., (1974). The use of balanced incomplete block designs for experiments involving sequences of treatments. Biometrika, 39, pp. 32–48.

SHEEHE, P. R., BROSS, D. J., (1961). Latin squares to balance immediate residual and other order effects. Biometrics, 17, pp. 405–414.

AFSARINEJAD, K., (1989). Circular balanced repeated measurements design, J. Statist. Prob. Letters., 7, pp. 3985–4027.

AFSARINEJAD, K., (1990). Repeated measurements designs - A review Comm. Statist. Theory, Math., 19(11), pp. 187–189.

HEDAYAT, A., AFSARINEJAD, K., (1975). Repeated measurements design I. In a Survey of Statistical Design and Linear Models, ed. (J. N. Srivastava Ed.) Amsterdam, North Holland, pp. 229–242.

HEDAYAT, A., AFSARINEJAD, K., (1978). Repeated measurements design II, The Annals of Statistics, 6(3), pp. 619–628.

PATTERSON, H. D., (1952). The construction of balanced designs for experiments involving sequences of treatments, Biometrika, 39, pp. 32–48.

WILLIAMS, E. J., (1949). Experimental designs balanced for the estimation of residual effects of treatments, Australian J. Sci. Res. A., 2, pp. 149–168.

BRADLEY, J. V., (1958). Complete counterbalancing of immediate sequential effects in a Latin Square Design, J. Am. Statist. Ass., 53, pp. 525–528.

GORDEN, B., (1961). Sequence in groups with distinct partial product, Pacific J. Math, 11, pp. 1309–1313.

PIGEON, J.G., D. RAGHAVARAO, (1987). Cross-over designs for comparing treatments with a control, Biometrika, 74, pp. 321–328.

KENWARD, M. G., JONES. B., (1987). A log linear model for binary cross over data. Applied Statistics, 36, pp. 192–204.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 145–162, DOI 10.21307/stattrans-2021-008
Received – 08.04.2019; accepted – 07.09.2020

Testing for a serial correlation in VaR failures through the exponential autoregressive conditional duration model

Marta Małecka[1]

ABSTRACT

Although regulatory standards, currently developed by the Basel Committee on Banking Supervision, anticipate a shift from VaR to ES, the evaluation of risk models currently remains based on the VaR measure. Motivated by the Basel regulations, we address the issue of VaR backtesting and contribute to the debate by exploring statistical properties of the exponential autoregressive conditional duration (EACD) VaR test. We show that, under the null, the tested parameter lies at the boundary of the parameter space, which can profoundly affect the accuracy of this test. To compensate for this deficiency, a mixture of chi-square distributions is applied. The resulting accuracy improvement allows for the omission of the Monte Carlo simulations used to implement the EACD VaR test in earlier studies, which dramatically improves the computational efficiency of the procedure. We demonstrate that the EACD approach to testing VaR has the potential to enhance statistical inference in most problematic cases – for small samples and for those close to the null.

Key words: VaR backtesting, exponential autoregressive conditional duration, boundary of the parameter space, test size, test power.

1. Introduction

Value-at-Risk (VaR) and Expected Shortfall (ES) are two measures of market risk that dominate contemporary banking regulation. Since its original inception in business (JP Morgan, 1994) and incorporation to regulatory standards (Basel Committee on Banking Supervision, 1996), VaR has become an industry standard in market risk management. Its constantly widening range of applications include new types of risk and new markets. Despite its widespread use, however, it has several flows. It does not take account of losses beyond a designated threshold as well as lacks subadditivity, which means that diversification does not, necessarily, imply reduction of risk. Therefore, ES, which remedies this problems, seems to be emerging as a new standard. In the light of the major reform of global supervisory standards, pursued

[1] Department of Statistical Methods, University of Łódź, Poland. E-mail: marta.malecka@uni.lodz.pl. ORCID: http://orchid.org/0000-0003-4465-9811.

by the Basel Committee since 2012 (Basel Committee on Banking Supervision, 2012-2017), ES is recommended for reporting exposures to market risk. Nevertheless, ES fails to satisfy a different mathematical principle – elicitability. Although this criterion has been shown to be erroneously deemed essential to backtesting (Gneiting, 2011), the question about ES-based statistical tests remains open (Acerbi and Szekely, 2014, Chen, 2014, Fissler and Ziegel, 2015, Fissler et al., 2016). No consensus on relevant procedures has yet been reached, either in academic studies or in business practice. Therefore, evaluation of risk models still relies on VaR. In an attempt to include most extreme losses, the regulator has recommended testing VaR on two low coverage levels – 1% and 2.5%. These Basel regulations motivate academics to review, develop and enhance statistical methods of backtesting VaR.

VaR backtesting procedures commonly refer to two criteria: the postulate of unconditional coverage, which treats the overall fraction of VaR violations, and the postulate of conditional coverage, which addresses their serial dependence. Perhaps of greater practical importance is detecting serial correlation of VaR failures, for their clustering may result in a series of catastrophic losses occurring one by one. This, in turn, seriously increases the risk of bankruptcy of a financial institution. The Markov test, which embeds the iid Bernoulli hypothesis within a binary first-order Markov chain and utilizes the likelihood ratio framework, has become the industry standard for testing the conditional coverage property (Christoffersen, 1998). This standard test, however, has been shown to exhibit unsatisfactory power (Lopez, 1999, Christoffersen and Pelletier, 2004, Berkowitz et al., 2011, Pajhede, 2017), which boosted the debate on other possibilities of testing VaR conditional coverage. Among other directions, like spectral tests (Berkowitz et al., 2011, Gordy and McNeil, 2018) or multi-level tests (Berkowitz, 2001, Hurlin and Tokpavi, 2007, Colletaz et al., 2013, Leccadito, Boffelli and Urga, 2014, Wied, Wei and Ziggel, 2016, Kratz et al. 2018), the duration-based approach attracted much attention in the scientific community (Christoffersen and Pelletier, 2004, Candelon et al. 2011, Pelletier and Wei, 2016). In the duration-based framework the sequence of VaR violations is transformed into the duration series. The idea behind this approach follows from the observation that the time that has passed since a VaR violation (hit) should not contain any information about further duration of the no-hit sequence. This implies the memory-free property of the duration series. Within discrete distributions, this property characterizes the geometric distribution (Berkowitz et al., 2011), while the only memory-free continuous distribution is the exponential distribution. To test VaR by means of the exponential distribution it has been proposed to nest the memory free null in the exponential autoregressive conditional model (EACD model, Engle and Russel, 1998). The EACD VaR test has been shown to compare favourably, in terms of its power, to other duration-based tests like the Weibull of the gamma test, especially for small sample sizes (Christoffersen and

Pelletier, 2004, Małecka, 2018). This test, however, suffers from significant size distortions, which means that the asymptotic distribution does not guarantee the correct test level. To make up for this deficiency it has been proposed to use the Monte Carlo method to simulate the null distribution of the test statistic (Christoffersen and Pelletier, 2004). The Monte Carlo approach, however, while ensuring the correct test level, impedes practical implementation of the procedure.

Our work addresses applicability of the exponential autoregressive conditional model to testing for serial correlation in VaR failure series. The goals of the paper are twofold: firstly, we seek to handle the problem of EACD test size distortions without resorting to the use of Monte Carlo simulations and secondly we investigate its power in relation to the standard VaR backtesting procedure. To avoid p-value computation through simulations, we study the asymptotic properties of the test statistic. Exploiting the fact that, in the VaR testing framework, the null value of the parameter vector lies exactly at the boundary of the parameter space, we show that the test statistic does not converge to the standard likelihood ratio (LR) limiting distribution. Using results on asymptotic LR properties under non-regular conditions (Self and Liang, 1987), we suggest p-value computation from the mixture of two chi-square distributions. We experimentally demonstrate the size improvement obtained by the proposed approach.

Given improved accuracy of the EACD VaR test, we investigate its power properties. To mimic a typical VaR failure correlation scheme, we adopt a GARCH model. The comparative evaluation of the EACD test power is conducted in relation to the Markov procedure, which has, so far, won widest recognition in the industry. We indicate cases where the EACD approach allows for power gains, which gives guidance as to practical application of the examined procedures.

Our study is based on earlier works by Christoffersen and Pelletier (2004) and Małecka (2018). The results of Christoffersen and Pelletier are improved by using asymptotic LR properties under non-regular conditions and implementing the EACD VaR test with the limiting mixture distribution. Since this replaces Monte Carlo simulations, our approach improves computational effectiveness of the procedure and facilitates its practical implementation. The results of Christoffersen and Pelletier are also improved by replacing the historical simulation model in the power study with the GARCH-model-based experiment. In this way we obtain the realistic setting, which mimics the volatility clustering of real financial data. In this experiment the serial correlation of VaR failures, as in reality, results from the volatility clustering of the portfolio returns. The volatility clustering is measured by the correlation coefficient of the squared returns, which, in the model we use, can be calculated analytically. Therefore, we are able to study the power of the test as a function of a controlled parameter of the return distribution, which is not attainable with the historical simulation experiment.

We extend the study by Małecka (2018) with respect to the contemporary international regulations in banking supervision. In addition to the typical 5% VaR, we include evaluation of test properties for two lower VaR coverage levels, indicated in the Basel rules. We discuss test accuracy in the context of the coverage level. We also extend the earlier study by depicting powers of the test as a function of volatility clustering. The shapes of the functions, compared to the power function of the standard Markov test, indicate cases where the EACD approach allows for more effective detection of incorrect risk models.

The paper proceeds as follows. Section 2 introduces the notation and presents the duration-based approach to VaR backtesting in relation to the standard Markov procedure. It shows the applicability of the EACD model to testing VaR and discusses the asymptotic distribution of the test statistic. Section 3 provides the study of test properties. Firstly, it details the design of the Monte Carlo experiment, showing a way to control volatility clustering. Secondly, it addresses test accuracy and presents improvements obtained by the use of the asymptotic mixture distribution. Finally, it gives comparative evaluation of test power in relation to the Markov test. The final section summarizes and concludes.

2. Testing VaR Conditional Coverage: EACD vs. Markov-Chain Approach

Let $\{R_t\}$ be the asset or portfolio return process, for which VaR at time t, at the level of tolerance p, is defined as the $p-$quantile of the relevant return distribution:

$$P\left(R_t < VaR_t(p)\right) = p, \quad t = 1,...,T. \tag{1}$$

Then, the VaR evaluation framework is based on the stochastic process of VaR failures:

$$I_t = \begin{cases} 1, & R_t < VaR_t(p) \\ 0, & R_t \geq VaR_t(p) \end{cases}, \tag{2}$$

whose realization is referred to as a hit sequence.

The standard Christoffersen's (1998) Markov test of VaR failure independence uses the framework of the binary Markov chain with the transition matrix:

$$\begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}, \tag{3}$$

where π_{ij} denotes the probability of a single-step transition from state i to state j, $i, j \in \{0,1\}$. The null hypothesis of equal transition probabilities $H_0 : \pi_{01} = \pi_{11}$ implies

the iid Bernoulli process with probability of VaR violence $\pi_1 = \pi_{01} = \pi_{11}$. To verify the above parameter restriction it has been proposed to use the likelihood ratio statistics:

$$LR_{ind} = -2\log\frac{\hat{\pi}_1{}^{t_1}(1-\hat{\pi}_1)^{t_0}}{\hat{\pi}_{01}{}^{t_{01}}(1-\hat{\pi}_{01})^{t_{00}}\hat{\pi}_{11}{}^{t_{11}}(1-\hat{\pi}_{11})^{t_{10}}} \sim_{as} \chi^2_{(1)}, \tag{4}$$

where $\hat{\pi}_1 = \dfrac{t_1}{t_0 + t_1}$, t_0 is the number of non-exceptions, t_1 – the number of exceptions, $\hat{\pi}_{01} = \dfrac{t_{01}}{t_0}$, $\hat{\pi}_{11} = \dfrac{t_{11}}{t_1}$ and t_{ij} – the number of transitions form state i to state j.

The construction of the Markov test implies that it only allows for detecting cases where the hit sequence follows a simple first-order Markov chain. A duration-based approach was proposed as means to capture more general forms of dependence. The duration-based tests use the transformation of the underlying $\{I_t\}$ process into the duration series $\{V_i\}$ defined as:

$$V_i = t_i - t_{i-1}, \tag{5}$$

where t_i denotes the time of the i-th VaR violation. The independence of the $\{I_t\}$ process implies that the time that has passed since a VaR violation (hit) should not contain any information about the further duration of the no-hit sequence. This memory-free property of the duration series motivates the use of the exponential distribution. In the exponential autoregressive conditional test the memory free null is tested against the alternative of the exponential process with a conditional mean. Exploiting the fact that the serially dependent hit sequence is likely to produce an excessive number of relatively short no-hit durations and relatively long no-hit durations, the test checks the autoregression coefficient of the conditional mean of the duration. The EACD approach utilizes the regression of the form:

$$E_{i-1}(V_i) = a + bV_{i-1} \tag{6}$$

(Engle and Russel, 1998). It assumes the exponential distribution, which gives the following conditional pdf function of the duration V_i:

$$f_{EACD}(v_i) = \frac{1}{a + bv_{i-1}} e^{-\frac{v_i}{a+bv_{i-1}}}. \tag{7}$$

Under the null hypothesis $H_0 : b = 0$ the conditional distribution becomes the exponential distribution with a constant mean.

By using the regression of the durations on their past values this test incorporates the information about the ordering of VaR failures. This offers potential power gains over other duration based procedures like the Weibull test or the gamma test, that simply nest the exponential distribution in wider distribution families and verify relevant restrictions.

The EACD-based VaR test verifies the parameter restriction through the likelihood ratio statistic, which requires computation of the loglikelihood function for the unrestricted and restricted case. Taking account of possible presence of censored durations at the beginning and at the end of the series, the loglikelihood takes the form:

$$\begin{aligned}\log L(V,\theta) &= C_1 \log S(V_1) + (1-C_1)\log f(V_1) + \sum_{i=2}^{N-1} \log f(V_i) + \\ &+ C_N \log S(V_N) + (1-C_N)\log f(V_N),\end{aligned} \tag{8}$$

where C_i is 1 if the duration V_i is censored and 0 otherwise, S is the survival function of the variable V_i, N is the number of VaR failures and θ is the vector of parameters (Christoffersen and Pelletier, 2004).

Assuming parameter values in the interior of the parameter space, the likelihood ratio statistic for one parameter restriction has the chi-square distribution with one degree of freedom χ_1^2. However, if the tested parameter value lies at or near the boundary of the parameter space, the asymptotic convergence to the chi-square distribution ceases to hold true. This is the case with the EACD VaR test since the null hypothesis imposes the zero value of the autoregression coefficient, and, at the same time, the coefficient satisfies the nonnegativity condition. This means that the vector of ECAD model parameters $\theta = [a,b]$ belongs to the space $\Theta = (0,+\infty)\times[0,+\infty)$, which, under the null, reduces to $\Theta_0 = (0,+\infty)\times\{0\}$. In such a case statistical inference based on the asymptotic χ_1^2 may be inaccurate. To overcome the problem of potential size distortions, the EACD VaR test has been originally implemented with the use of the Monte Carlo simulated p-values. Instead, using asymptotic results on the likelihood ratio distribution under non-standard conditions (Self and Liang, 1987), we propose to compute the p-values from the 50:50 mixture of chi-square distributions, with zero and one degrees of freedom:

$$LR_{EACD} \sim as 0.5\chi_0^2 + 0.5\chi_1^2. \tag{9}$$

Using the fact that the chi-square distribution with zero degrees of freedom reduces to the distribution with all its mass cumulated at zero, we get that the value of the test with 50% probability takes the value of 0 and with 50% probability is drawn from the chi-square distribution with one degree of freedom χ_1^2.

3. Monte Carlo Study of Test Properties

The tests described in Section 2 verify the conditional coverage property of VaR failures referring to the Markov chain framework or, after the transformation of the hit sequence into durations, to the exponential autoregressive conditional duration model. Since the two tests exploit different approaches and make use of different variables, they are likely to differ in power properties. Moreover, as they rely on asymptotic distributions, their finite sample properties are unknown. In the present section, using a finite sample setting, we evaluate and compare the statistical properties of the two tests through the Monte Carlo study. The comparative analysis includes their size and power. We discuss practical implications of the power properties, presenting conclusions as to when to prefer which of the two tests and indicating cases when the two approaches may complement each other.

The finite-sample statistical properties of the tests are evaluated for sample sizes chosen to be realistic for applications in finance: $T = 250,\ 500,...,\ 1500$. Such samples roughly correspond to daily data covering periods from one year to six years. The size and the power of the tests are approximated by rejection frequencies under the null and under the alternative, respectively. The size study includes significance levels 0.01, 0.05 and 0.1. For powers of the tests, only rejection rates at 0.05 significance level are reported. The size and the power estimates are computed over 10000 Monte Carlo trials.

The size study examines test rejection probabilities when the risk model is correct. We refer to a test as accurate if, under the correct model, the rejection probability corresponds to the assumed level of significance (nominal test size). Therefore, the size study requires generating $\{I_t\}$ series under the correct model, i.e. under the assumptions of the true failure probability and independence of VaR violations. To this end we use the Bernoulli distribution with the probability of success p, equal to the assumed level of VaR tolerance.

The size estimates obtained from the Bernoulli experiment (Tables 1-3) show the accuracy improvement of the EACD test gained by replacing the χ^2_1 distribution by the mixture of distributions $0.5\chi^2_0 + 0.5\chi^2_1$. In the case of the χ^2_1 the procedure is very conservative with the true test level leaning towards zero. This size distortion indicates that practical application of this test should not be based on the asymptotic χ^2_1 distribution. Employment of the mixture $0.5\chi^2_0 + 0.5\chi^2_1$ has the effect that the true test level approaches the nominal size. The test still tends to underreject the null, however the discrepancies between the simulated and the nominal size markedly decrease and the simulated rejection frequencies seem to converge to the desired level with lengthening the sample. The improvement in the accuracy of the test is demonstrated through the fit of the asymptotic and the empirical distribution function, based on a 1500 observation sample (Figure 1).

Table 1. Size estimates for Markov and EACD 1% VaR tests*

Test	Significance level $\alpha = 0.01$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0111	0.0122	0.0116	0.0124	0.0128	0.0128
$LR_{EACD}^{Chi\ square}$	0.0000	0.0000	0.0009	0.0010	0.0018	0.0019
$LR_{EACD}^{Mixture}$	0.0000	0.0008	0.0024	0.0034	0.0051	0.0047
Test	Significance level $\alpha = 0.05$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0234	0.0248	0.0293	0.0269	0.0209	0.0210
$LR_{EACD}^{Chi\ square}$	0.0002	0.0016	0.0085	0.0110	0.0131	0.0146
$LR_{EACD}^{Mixture}$	0.0013	0.0077	0.0203	0.0248	0.0298	0.0358
Test	Significance level $\alpha = 0.1$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0294	0.0418	0.0474	0.0480	0.0425	0.0453
$LR_{EACD}^{Chi\ square}$	0.0013	0.0077	0.0203	0.0248	0.0298	0.0358
$LR_{EACD}^{Mixture}$	0.0080	0.0280	0.0487	0.0616	0.0683	0.0796

* $LR_{EACD}^{Chi\ square}$ denotes the cases when the LR_{EACD} test size was estimated under the χ_1^2 distribution, while $LR_{EACD}^{Mixture}$ – the cases when the size was estimated under the mixture distribution $0.5\chi_0^2 + 0.5\chi_1^2$.

Source: Own work.

Table 2. Size estimates for Markov and EACD 2.5% VaR tests*

Test	Significance level $\alpha = 0.01$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0265	0.0293	0.0310	0.0284	0.0274	0.0274
$LR_{EACD}^{Chi\ square}$	0.0002	0.0007	0.0011	0.0009	0.0019	0.0022
$LR_{EACD}^{Mixture}$	0.0008	0.0016	0.0024	0.0030	0.0042	0.0047
Test	Significance level $\alpha = 0.05$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0393	0.0447	0.0443	0.0448	0.0424	0.0430
$LR_{EACD}^{Chi\ square}$	0.0032	0.0052	0.0079	0.0097	0.0104	0.0119
$LR_{EACD}^{Mixture}$	0.0077	0.0126	0.0189	0.0234	0.0241	0.0253

Table 2. Size estimates for Markov and EACD 2.5% VaR tests* (cont.)

Test	Significance level $\alpha = 0.1$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0526	0.0635	0.0640	0.0685	0.0784	0.0951
$LR_{EACD}^{Chi\ square}$	0.0077	0.0126	0.0189	0.0234	0.0241	0.0253
$LR_{EACD}^{Mixture}$	0.0203	0.0329	0.0429	0.0526	0.0557	0.0552

* $LR_{EACD}^{Chi\ square}$ denotes the cases when the LR_{EACD} test size was estimated under the χ_1^2 distribution, while $LR_{EACD}^{Mixture}$ – the cases when the size was estimated under the mixture distribution $0.5\chi_0^2 + 0.5\chi_1^2$.

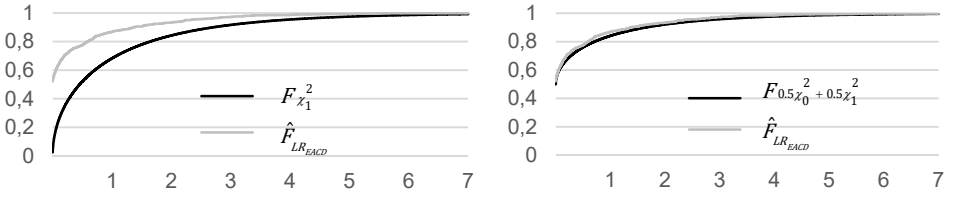
Source: Own work.

Table 3. Size estimates for Markov and EACD 5% VaR tests*

Test	Significance level $\alpha = 0.01$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0194	0.0293	0.0281	0.0340	0.0368	0.0400
$LR_{EACD}^{Chi\ square}$	0.0006	0.0008	0.0013	0.0014	0.0026	0.0017
$LR_{EACD}^{Mixture}$	0.0016	0.0036	0.0038	0.0046	0.0055	0.0061
Test	Significance level $\alpha = 0.05$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.0707	0.094	0.1199	0.1276	0.1233	0.1147
$LR_{EACD}^{Chi\ square}$	0.0054	0.0068	0.0105	0.0108	0.0139	0.0142
$LR_{EACD}^{Mixture}$	0.0115	0.0196	0.0258	0.0279	0.0301	0.0308
Test	Significance level $\alpha = 0.1$					
	Series length					
	250	500	750	1000	1250	1500
LR_{Ind}	0.1012	0.1797	0.2033	0.1792	0.1656	0.1654
$LR_{EACD}^{Chi\ square}$	0.0114	0.0197	0.0209	0.0265	0.0298	0.0289
$LR_{EACD}^{Mixture}$	0.0321	0.0483	0.058	0.062	0.0646	0.0675

* $LR_{EACD}^{Chi\ square}$ denotes the cases when the LR_{EACD} test size was estimated under the χ_1^2 distribution, while $LR_{EACD}^{Mixture}$ – the cases when the size was estimated under the mixture distribution $0.5\chi_0^2 + 0.5\chi_1^2$.

Source: Own work.



(a) Empirical cdf versus χ_1^2 (b) Empirical cdf versus $0.5\chi_0^2 + 0.5\chi_1^2$

Figure 1. Empirical cdf of the LR_{EACD} test statistic, based on 1,500 observations, versus χ_1^2 and $0.5\chi_0^2 + 0.5\chi_1^2$ asymptotic distributions.

Source: Own work.

Comparative evaluation of the size results in relation to the relevant outcomes for the benchmark Markov procedure shows that the EACD $LR_{EACD}^{Mixture}$ rejection frequencies tend to be closer to the nominal test size. The EACD test seems also more reliable as the relation of the estimated size to the chosen significance level is remarkably stable across significance levels and VaR coverage levels, especially for large samples. For the Markov LR_{Ind} this relation changes rapidly with both: chosen significance and VaR coverage. Contrary to the systematically undervalued but apparently convergent $LR_{EACD}^{Mixture}$ rejection frequencies, the LR_{Ind} test changes from being undersized to being oversized. Its rejection frequencies are much overvalued for 5% VaR, while for lower coverage levels they shift from being overvalued in tails to undervalued closer to the central area of the distribution. The differences between the estimated and the nominal size of LR_{Ind} range from minor – for 1% VaR and 0.01 significance – to large – for 5% VaR and 0.01 significance. In the last case the estimated size overvalues the nominal significance four times. Therefore, in the light of the results for all considered significance and coverage levels, the EACD approach to testing VaR offers accuracy improvement in comparison to the standard Markov-chain-based VaR test.

The size improvement attainable with the proposed method confirms that the asymptotic mixture distribution works well for the EACD test. However, this method does not solve the problem of the inaccurate size for small samples. Our results show that this problem cannot be handled by any of the considered approaches. In particular, none of the asymptotic approximations, including both standard likelihood ratio distribution and the non-standard mixture distribution, is relevant for daily data covering a one-year period. Therefore, if the sample size is limited, it seems recommendable to resort to the Monte-Carlo-based methods.

The comparison of the size results across 1%, 2.5% and 5% coverage levels shows that the EACD $LR_{EACD}^{Mixture}$ test performs best for 5% VaR. In this case, the observed size

distortions are lowest and convergence to the desired levels is fastest. This is particularly visible for popular significance levels 0.01 and 0.05. The recommended VaR coverage, for EACD-based VaR backtesting, is thus not in line with contemporary trends in banking supervision. The lower coverage levels, suggested by the Basel rules, lead to $LR_{EACD}^{Mixture}$ accuracy loss.

The power study investigates test performance under the assumption of an incorrect risk model. The test is regarded as more effective if its rejection frequencies, under the incorrect model, are higher. Since the considered procedures are aimed at checking the conditional coverage property, we study their ability to reject clustered VaR violations. Therefore, the simulation experiment in the power study is designed to reflect the volatility clustering of return data. The volatility clustering, in turn, implies the undesired serial correlation of VaR violations.

The $\{I_t\}$ series under the incorrect model is computed as the hit sequence from the GARCH(1,1) return process and the constant VaR level. The VaR level is set to the value of the unconditional p – quantile of the returns. This produces VaR failures with the tendency to correlate in time and, at the same time, guarantees the correct overall VaR failure rate. Through employment of the GARCH process we obtain the realistic setting, which mimics the volatility clustering of real financial data. The volatility clustering is measured by the correlation coefficient of the squared returns, which, under the specification we use, can be calculated analytically. This enables us to study the power of the test as a function of a controlled parameter of the return distribution. In order to calculate analytically the correlation coefficient of the squared returns, we use the GARCH model of the form:

$$\begin{aligned} R_t &= \sqrt{h_t} Z_t, \;\; Z_t : \;\; N(0,1), \\ h_t &= \omega + \alpha \varepsilon_{t-1}^2 + \beta h_{t-1}. \end{aligned} \tag{10}$$

Under specification (10) the correlation of the squared returns ρ is given by

$$\rho = \frac{\alpha^2 \beta}{1 - 2\alpha\beta - \beta^2}. \tag{11}$$

This is, however, subject to the restriction

$$(\alpha + \beta)^2 + 2\alpha^2 < 1 \tag{12}$$

If condition (12) does not hold, the correlations of the GARCH model are time-varying. In such a case, they have been shown to behave approximately as:

$$\rho = \alpha + \frac{\beta}{3}. \tag{13}$$

We choose the GARCH parameters on realistic levels $\omega = 0.01$, $\beta = 0.85$ and the correlation coefficient ρ to vary from 0.05 to 0.5. The α parameter is set to levels that

guarantee the desired value of ρ. Under the above parameter values the restriction (12) holds for correlations not higher than 0.4, thus a range of values from 0.05 to 0.4 is used.

Due to observed size distortions, in the power exercise we adopt the Monte Carlo test technique, which provides exact tests by replacing theoretical null distributions of test statistics by their sample analogues [6]. Through ensuring a correct test level we obtain comparability of the power results. Since estimating EACD model parameters requires at least three durations, which corresponds to at least two VaR violations, there are cases when the test is not feasible. Rejecting these cases constitutes a non-random sample selection rule. Therefore, we present effective power rates, which correspond to multiplying raw power by the rate of valid test runs. Referring to the results of the size study, in the power exercise we rely on 5% VaR. We report rejection frequencies for 0.05 significance level.

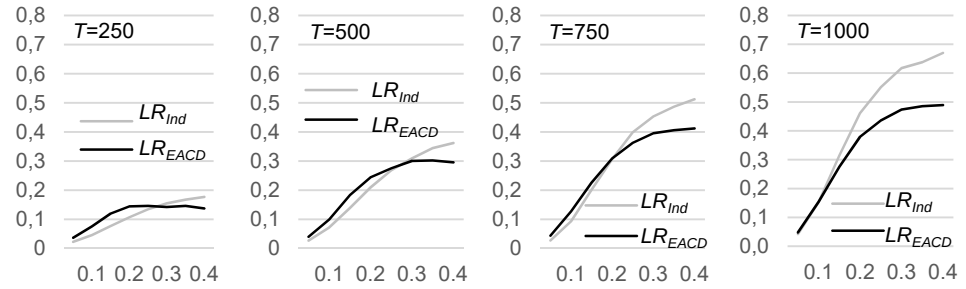
The power of the EACD LR_{EACD} test is evaluated in relation to rejection frequencies of the Markov LR_{Ind} test (Table 4). The results show superiority of the EACD procedure at short distances from the null. In the case of volatility clustering corresponding to 0.05 and 0.1 correlation of the squared returns, the EACD procedure exhibits higher power than the benchmark for all series lengths. Subsequent experiments show that this comparative advantage tends to vanish for a stronger correlation. However, it is observed relatively long for small samples.

Table 4. Power estimates for 5% VaR Markov and EACD tests on 0.05 significance level

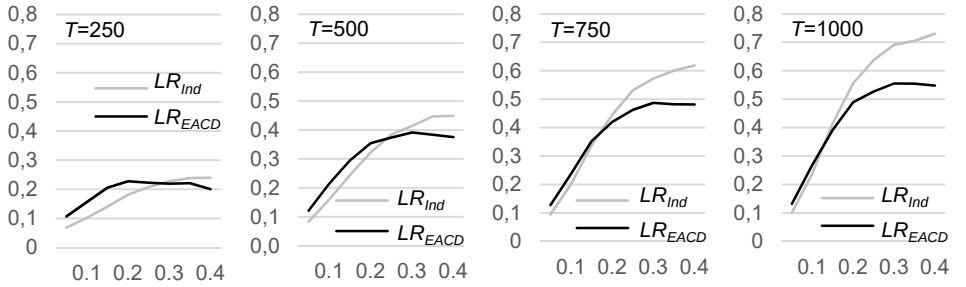
Test	Volatility clustering*	Series Length					
		250	500	750	1000	1250	1500
LR_{Ind}	0.05	0.068**	0.083**	0.095**	0.101**	0.102**	0.104**
	0.10	0.102**	0.160**	0.201**	0.239**	0.281**	0.316**
	0.15	0.141**	0.244**	0.337**	0.415	0.487	0.546
	0.20	0.181**	0.322**	0.445	0.557	0.633	0.708
	0.25	0.208**	0.384	0.531	0.637	0.722	0.803
	0.30	0.228	0.414	0.573	0.691	0.775	0.836
	0.35	0.239	0.447	0.600	0.705	0.787	0.859
	0.40	0.240	0.449	0.619	0.730	0.807	0.866
LR_{EACD}	0.05	0.107**	0.121**	0.128**	0.132**	0.137**	0.155**
	0.10	0.156**	0.215**	0.238**	0.269**	0.287**	0.321**
	0.15	0.206**	0.295**	0.353*	0.392	0.439	0.477
	0.20	0.228**	0.355**	0.420*	0.490	0.523	0.583
	0.25	0.223**	0.374	0.463	0.527	0.582	0.635
	0.30	0.220	0.392	0.487	0.555	0.618	0.657
	0.35	0.221	0.384	0.482	0.555	0.615	0.668
	0.40	0.201	0.375	0.482	0.548	0.614	0.662

*The volatility clustering in the simulated process is measured by the correlation coefficient of the squared returns ρ.

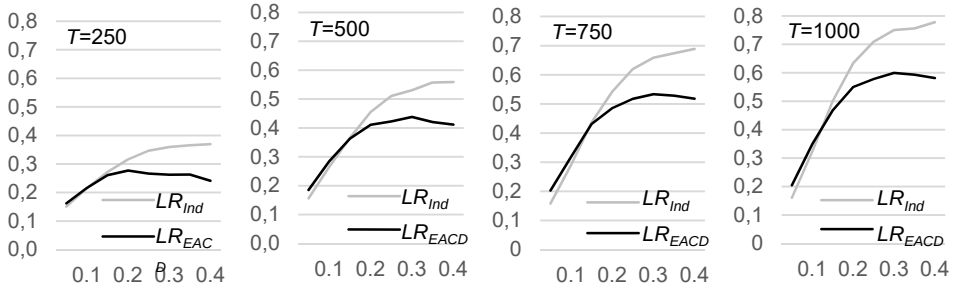
**Cases when the estimated power of LR_{EACD} exceeds that of LR_{Ind} are marked with double asterix.



(a) Significance level $\alpha = 0.01$



(b) Significance level $\alpha = 0.05$



(c) Significance level $\alpha = 0.1$

Figure 2. Power function estimates against volatility clustering for VaR tests, T=250,…,1000.

Source: Own work.

The relative performance of the tests is depicted by the power functions plotted against the strength of volatility clustering (Figure 2). The figures extend the power study, illustrating estimated powers for three significance levels: 0.01, 0.05 and 0.1. The sample sizes range from 250 to 1000, as for longer series the test performance is

relatively stable and follows the trends observed for 1000 sample. All plots confirm that the EACD power tends to grow faster than the Markov test power close to the null. Thus, this test is likely to outperform the benchmark procedure in detecting low-scale correlations. Its advantage in power against low-scale correlations is especially large for small samples. This suggests that the EACD approach has the potential to improve testing efficiency in cases when statistical inference is particularly troublesome – for small samples and close to the null.

The power functions illustrate also the downswing in the EACD test performance for largest correlations. This suggests that after exceeding some critical value of the correlation of the squared returns, the test power starts to deteriorate. Large correlation in squared returns is likely to produce VaR violations occurring one by one in turbulent periods, followed by long calm periods without any violation. This translates into a typical setup of a duration sequence with series of very short durations from turbulent time, interrupted by one long duration, corresponding to the calm period. The single outstanding duration replaces series of long durations. In such a setting the autoregressive model of durations tends to be insignificant. Thus, the excessive correlation of the squared returns works against the power of the test. This supports the practical conclusion that the EACD approach to testing VaR is particularly recommendable for detecting low-scale distortions from the null.

Combining the size and the power results, the EACD procedure seems complementary to the standard Markov test, as their relative performance depends on the distance from the null. At the null, the EACD test outperforms the benchmark procedure. This means that it is less likely to overreject the correct risk model. Of practical importance is the fact that its performance at the null is remarkably stable across significance levels and VaR coverage levels. Thus, its accuracy is only slightly influenced by the user's parameter choices. Close to the null, the EACD power grows quickly, which makes the test more sensitive to low-scale correlations. In practice it means that it is more likely to detect incorrect risk models when clustering of VaR failures is relatively small. On the other hand, the Markov test performs better at detecting a large correlation of VaR violations. Thus, in the light of their statistical properties, it is advisable that the procedures be employed simultaneously in testing conditional coverage property. Another practical guideline from our results is that the contrary decisions of the two tests may occur due to low-scale correlation rather than the type one error. Therefore, such an outcome of the backtesting procedure signals that the risk model should be recognized as incorrect.

4. Conclusion

The paper tackled the issue of evaluating risk models with respect to the contemporary changes in international banking regulation. In accordance with the Basel recommendations, we inquired into ways of assessing risk models based on the VaR measure. In this context we studied applicability of the EACD model. We considered the EACD test as means of testing conditional coverage property of VaR violations. We addressed the construction, asymptotic distribution as well as the finite sample size and power properties of the test.

With reference to the accuracy of backtesting, we sought to handle the problem of EACD test size distortions without resorting to the use of Monte Carlo simulations. Based on the observation that the conditional coverage property implies the parameter restriction that lies at the boundary of the parameter space, we suggested p-value computation from the mixture of chi-square distributions. In this way we obtained the procedure which is both accurate and computationally effective as it replaces the originally proposed Monte Carlo method. Since its construction is based on the duration series instead of the hit sequence, it also has the potential to exhibit power against more general forms of dependence than the standard VaR test, which operates within the framework of the first order Markov chain.

Via simulations we showed improvement in the test accuracy owned to replacing the asymptotic likelihood ratio distribution with a mixture of chi-square distributions. We confirmed the convergence of the true test level to the nominal size of the test. With the use of the GARCH model we designed the experiment, which enabled us to study the power of the tests against various levels of volatility clustering in return data. The estimated power functions showed that the EACD test outperforms the benchmark Markov procedure at the null and its power grows faster close to the null. Thus, this procedure may be useful to detect low-scale correlations and in this sense it may complement the standard Markov test. This comparative advantage of the EACD test turned out to be particularly large for shortest examined series lengths. Therefore, our results suggested that the EACD approach to VaR testing may aid statistical inference in most troublesome cases – for small samples and close to the null.

References

ACERBI, C., SZEKELY, B., (2014). Backtesting Expected Shortfall, *Risk*, November.

BASEL COMMITTEE ON BANKING SUPERVISION, (1996). Amendment to the capital accord to incorporate market risks: Technical document, available online: http://www.bis.org/publ/bcbs24.pdf (accessed June 4, 2018).

BASEL COMMITTEE ON BANKING SUPERVISION, (2012). Fundamental Review of the Trading Book: Technical document, available online: http://www.bis.org/publ/bcbs219.pdf (accessed June 4, 2018).

BASEL COMMITTEE ON BANKING SUPERVISION, (2013). Fundamental Review of the Trading Book: A Revised Market Risk Framework: Technical document, available online: http://www.bis.org/publ/bcbs265.pdf (accessed June 4, 2018).

BASEL COMMITTEE ON BANKING SUPERVISION, (2014). Fundamental Review of the Trading Book: Outstanding Issues: Technical document, available online: http://www.bis.org/bcbs/publ/d305.pdf (accessed June 4, 2018).

BASEL COMMITTEE ON BANKING SUPERVISION, (2015). Fundamental review of the trading book - interim impact analysis: Technical document, available online: http://www.bis.org/bcbs/publ/d346.pdf (accessed June 4, 2018).

BASEL COMMITTEE ON BANKING SUPERVISION, (2016). Minimum capital requirements for market risk: Technical document, available online: http://www.bis.org/bcbs/publ/d352.pdf (accessed June 4, 2018).

BASEL COMMITTEE ON BANKING SUPERVISION, (2017). High-level summary of Basel III Reforms: Technical document, available online: https://www.bis.org/bcbs/publ/d424_hlsummary.pdf (accessed June 4, 2018).

BERKOWITZ, J., (2001). Testing Density Forecasts with Applications to Risk Management, *J Bus Econ Stat*, Vol. 19(4), pp. 465–474, doi: https://dx.doi.org/10.1198/07350010152596718.

BERKOWITZ, J., CHRISTOFFERSEN, P., PELLETIER, D., (2011). Evaluating Value-at-Risk Models with Desk-Level Data, *Manage Sci*, Vol. 12(57), pp. 2213–2227, doi: https://dx.doi.org/10.1287/mnsc.1080.0964.

CANDELON, B., COLLETAZ, G., HURLIN, C., TOKPAVI, S., (2011). Backtesting Value-at-Risk: a GMM duration-based test, *J Financ Economet*, Vol. 9(2), pp. 314–343, doi: https://doi.org/10.1093/jjfinec/nbq025.

CHEN, J. M., (2014). Measuring market risk under the Basel accords: VaR, stressed VaR, and expected shortfall. Aestimatio, *The IEB International Journal of Finance*, Vol. 8, pp.184–201, doi: https://doi.org/10.2139/ssrn.2252463.

CHRISTOFFERSEN, P., (1998). Evaluating Interval Forecasts, *Int Econ Rev*, Vol. 39(4), pp. 841–862, doi: https://doi.org/10.2307/2527341.

CHRISTOFFERSEN, P., PELLETIER, D., (2004). Backtesting Value-at-Risk: A Duration-Based Approach, *J Financ Economet*, Vol. 2(1), pp. 84–108, doi: https://doi.org/10.1093/jjfinec/nbh004.

COLLETAZ, G., HURLIN, C., PERIGNON, C., (2013). The Risk Map: a New Tool for Risk Management, *J Bank Financ*, Vol. 37(10), pp. 3843–3854, doi: https://doi.org/10.1016/j.jbankfin.2013.06.006.

DUFOUR, J. M., (2006). Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics, *J Econometrics*, Vol. 133(2), pp. 443–477, doi: https://doi.org/10.1016/j.jeconom.2005.06.007.

ENGLE, R. F., RUSSEL, J. R., (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data, *Econometrica*, Vol. 66(5), pp. 1127–62, doi: https://doi.org/10.2307/2999632.

FISSLER, T., ZIEGEL, J. F., GNEITING, T., (2016). Expected shortfall is jointly elicitable with value at risk – Implications for backtesting, *Risk*, Vol. 29, pp. 58–61.

FISSLER, T., ZIEGEL, J. F., (2016). Higher order elicitability and Osband's principle, *Ann Stat*, Vol. 44(4), pp. 1680–707, doi: https://doi.org/10.1214/16-AOS1439.

GNEITING, T., (2011). Making and evaluating point forecasts, *J Am Stat Assoc*, Vol. 106(494), pp. 746–762, doi: https://doi.org/10.1198/jasa.2011.r10138.

GORDY, M. B., MCNEIL, A. J., (2018). Spectral Backtests of Forecast Distributions with Application to Risk Management, in: *Finance and Economics Discussion Series* 2018-021, Board of Governors of the Federal Reserve System, Washington.

HURLIN, CH., TOKPAVI, S., (2007). Backtesting value-at-risk accuracy: a simple new test, *J Risk*, Vol. 9(2), pp. 19–37, doi: https://doi.org/10.21314/JOR.2007.148.

KRATZ, M., LOK, Y. H., MCNEIL, A. J., (2018). Multinomial VaR backtests: A simple implicit approach to backtesting expected shortfall, *J Bank Financ*, Vol. 88, pp. 393–407, doi: https://doi.org/10.1016/j.jbankfin.2018.01.002.

LECCADITO, A., BOFFELLI, S., URGA, G., (2014). Evaluating the Accuracy of Value-at-Risk Forecasts: New Multilevel Tests, *Int J Forecasting*, Vol. 30(2), pp. 206–216, 014.doi: https://doi.org/10.1016/j.ijforecast.2013.07.

LOPEZ, J., (1999). Methods for Evaluating Value-at-Risk Estimates, *FRBSF Economic Review*, Vol. 2, pp. 3–17.

MAŁECKA, M., (2018). Exponential Autoregressive Conditional Duration Approach to Testing VaR, in: *ICoMS 2018: Proceedings of the 2018 International Conference*

on Mathematics and Statistics, ACM, New York, pp. 6–10, doi: https://doi.org/10.1145/3274250.3274254.

PAJHEDE, T., (2017). Backtesting Value-at-Risk: A Generalized Markov Test, *J Forecast*, Vol. 36(5), pp. 597–613, doi: https://doi.org/10.1002/for.2456.

PELLETIER, D., WEI, W., (2016). The geometric-VaR backtesting method, *J Financ Economet*, Vol. 14(4), pp. 725–745, doi: https://doi.org/10.1093/jjfinec/nbv015.

SELF, S. F., LIANG, K. Y., (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions, *J Am Stat Assoc*, Vol. 82(398), pp. 605–610, doi: https://doi.org/10.2307/2289471.

WIED, D., WEI, G. N. F., ZIGGEL, D., (2016). Evaluating Value-at-Risk forecasts: a new set of multivariate backtests, *J Bank Financ*, Vol. 72, pp. 121–132, doi: https://doi.org/10.1016/j.jbankfin.2016.07.014.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 163–178, DOI 10.21307/stattrans-2021-009
Received – 29.12.2019; accepted – 16.02.2021

On the generalisation of Quatember's bootstrap

Tomasz Żądło[1]

ABSTRACT

The problem of the estimation of the design-variance and the design-MSE of different estimators and predictors is considered. Bootstrap algorithms applicable to complex sampling designs are used. A generalisation of the bootstrap procedure studied by Quatember (2014) is proposed. In most of the cases considered in our simulation study it leads to more accurate estimates (or to very similar ones in remaining cases) of the design-MSE and the design-variance compared with the original algorithm and its other counterparts.

Key words: bootstrap for complex sampling designs, variance estimation, MSE estimation.

1. Introduction

Let the population of size N be denoted by Ω. The population is divided into D disjoint subpopulations (domains) Ω_d, each of size N_d, where $d = 1,2,...,D$. Let the sample be denoted by s and its size by n. The set of sampled elements of dth domain is denoted by s_d and its size by n_d. Let the values of the variable of interest observed in the sample be denoted by y_k $(k = 1,2,...,n)$. We additionally assume that vectors of auxiliary variables $\mathbf{x}_l$ $(l = 1,2,...,N)$ are known for all population elements. First and second order inclusion probabilities are denoted by π_k and π_{kl}, respectively. We consider the problem of estimation of the population (subpopulation) parameter θ (θ_d) using estimator $\hat{\theta}$ ($\hat{\theta}_d$). The key issue is the estimation of the design-variance and the design-MSE of $\hat{\theta}$ ($\hat{\theta}_d$). In official statistics, the design-based accuracy is of primary interest and hence model-based methods, where the prediction accuracy is assessed, are not widely used. What is more, the comparison of the accuracy of methods based on different approaches (e.g. design-based and model-based under different

[1] University of Economics in Katowice, Katowice, Poland, E-mail: tomasz.zadlo@ue.katowice.pl. ORCID: https://orcid.org/0000-0003-0638-0748.

superpopulation models) is not appropriate if MSE is estimated under different approaches too. Hence, the aim of the paper is to present:

- a proposal of a generalisation of Quatember (2014) bootstrap valid for complex sampling designs, which can be used to estimate the design-precision and the design-accuracy of any estimator or predictor,
- a simulation study of properties of our proposals and other bootstrap estimators of the design-variance and the design-MSE not only in the case of estimation of population parameters but also in the case of estimation and prediction of subpopulations characteristics.

2. Bootstrap methods for complex sampling designs

The classic Efron's bootstrap (Efron, 1979) procedure, where simple random samples are drawn with replacement from the original sample, is correct under independence of random variables. In the case of complex sampling designs appropriate modifications must be used.

According to Ranalli and Mecatti (2012), majority of bootstrap methods for complex sampling designs can be classified into one out of two approaches. The first one is called an ad-hoc approach and is usually based on iid resampling and rescaling sample data. They classify, inter alia, the rescaling bootstrap (Rao and Wu, 1988), the mirror-match bootstrap (Sitter, 1992) and the generalised weighted bootstrap (Beaumont and Patak, 2012) as methods belonging to this approach. Proposals presented by Antal and Tillé (2011, 2014) are also taken into account in this approach. The Authors use mixtures of several sampling designs for resampling to meet two conditions – firstly, the expectation over the bootstrap distribution of the Horvitz-Thompson (1952) (HT) estimator must be equal to the value of the HT estimator computed based on the original sample; secondly, the variance over the bootstrap distribution of HT estimator must be equal (or approximately equal) to the HT variance estimator (Horvitz and Thompson, 1952) or Sen-Yates-Grundy variance estimator (Sen 1953, Yates and Grundy 1953). The second approach is the plug-in approach. It is based on the concept of pseudopopulation, although in some methods the pseudopopulation is not physically generated. The basic idea is as follows:

1) We built a pseudopopulation $\Omega^* = \{1, 2, ..., k^*, ..., N^{pseudo}\}$, where pseudoelements are replications of elements observed in the original sample. The element k observed in the original sample is replicated w_k -times.
2) A bootstrap sample s^* of size n (original sample size) is drawn from Ω^* mimicking the original sampling design.

3) The value of estimator $\hat{\theta}$ is computed based on s^* and it is denoted by $\hat{\theta}^*$.
4) Steps b) and c) are iterated B times providing $\hat{\theta}_b^*$, where $b = 1, 2, ..., B$.

Bootstrap estimators of the design-variance and the design-bias are defined as follows (e.g. Rao and Wu 1988):

$$\hat{D}_{boot}^2(\hat{\theta}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*\right)^2, \tag{1}$$

$$\hat{B}_{boot}(\hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^* - \hat{\theta}, \tag{2}$$

where $\hat{\theta}$ is the value of the considered estimator based on the original sample.

The algorithm presented above allows for different definitions of weights w_k. One of the first proposals was presented by Holmberg (1998), who defined it as follows: $w_k = \lfloor \pi_k^{-1} \rfloor + \epsilon_k$, where $\lfloor \pi_k^{-1} \rfloor$ is rounded down value of π_k^{-1}, ϵ_k is generated from Bernoulli distribution with probability $\pi_k^{-1} - \lfloor \pi_k^{-1} \rfloor$.

Other solutions include Barbiero and Mecatti (2010) 0.5 bootstrap, where inverses of inclusion probabilities are rounded to the nearest integer. Barbiero and Mecatti (2010) consider two x-balanced methods, where inverses of first order inclusion probabilities are rounded down and additional pseudoelements are included in the pseudopopulation to reach the minimum absolute difference between total values of an auxiliary variable in the real population and the pseudopopulation. Barbiero, Manzi and Mecatti (2015) define w_k as calibration weights rounded to the nearest integer.

There are two possible limitations of the above algorithms. Firstly, we require generation of the pseudopopulation of size (approximately) equal to the original population size, which may be problematic in the case of large real populations. Secondly, the number of replications w_k must be integer. The first problem is solved by Ranalli and Mecatti (2012) by directly re-sampling from the sample using appropriate sampling designs where n out of n elements are drawn at random, mimicking the original sample design, where N out of n elements are selected. The Quatember (2014) bootstrap omits both of the limitations but it is proposed only for simple random sampling without replacement and for probability proportional to size sampling.

Let us present the idea of the Quatember (2014) bootstrap. Although the pseudopopulation is not created, the process of sampling from the pseudopopulation is mimicked in the procedure of selecting a bootstrap sample of size n out of n elements observed in the original sample with appropriate probabilities by modification of the original sampling scheme. Firstly, let us present the algorithm of drawing bth ($b = 1, 2, ...B$) bootstrap sample of size n for simple random sampling without replacement. Quatember (2014) assumes that the number of replications of sample

element k in the pseudopopulation, which is not physically created, equals its (possibly non-integer) design-weight - the inverse of the first order probability: Nn^{-1}. After draw $j-1$ the number of remaining replications of element k in the pseudopopulation equals: $Nn^{-1}-h_{k,j-1}$, where $h_{k,j-1}$ is the number of replications of element k selected in the bootstrap procedure in the first $j-1$ draws. What is more, the probability of selecting a population element from the pseudopopulation of size N in the j th draw equals $(N-j+1)^{-1}$. Finally, element k is drawn from the original sample in the j th draw ($j=1,2,...,n$) of the bootstrap algorithm with probability:

$$\left(Nn^{-1}-h_{k,j-1}\right)\times(N-j+1)^{-1}. \tag{3}$$

Secondly, we present the algorithm of drawing bth ($b=1,2,...B$) bootstrap sample of size n for probability proportional to size sampling. Quatember (2014) assumes that the number of replications of sample element k in the pseudopopulation, which is not physically created, equals its (possibly non-integer) design-weight given by: $t_x(x_k n)^{-1}$, where $t_x=\sum_{i\in\Omega}x_i$. After draw $j-1$ the number of remaining replications of element k in the pseudopopulation equals: $t_x(x_k n)^{-1}-h_{k,j-1}$, where $h_{k,j-1}$ is the number of replications of element k selected in the bootstrap procedure in the first $j-1$ draws. What is more, Quatember (2014) assumes the following probability of selecting an population element from the pseudopopulation of size N in the jth draw in his algorithm: $x_k\left(t_x-\sum_{i\in s_{bj-1}}x_i\right)^{-1}$, where s_{bj-1} is the subset of bth bootstrap sample after draw $j-1$. The drawback of the Quatember (2014) bootstrap is that the assumed probability does not lead to the first order inclusion probabilities proportional to the values of the auxiliary variable (as they should be for probability proportional to size sampling). Finally, element k is drawn from the original sample in the jth draw ($j=1,2,...,n$) of the bootstrap algorithm with probability:

$$\left(t_x(x_k n)^{-1}-h_{k,j-1}\right)\times x_k\left(t_x-\sum_{i\in s_{bj-1}}x_i\right)^{-1}. \tag{4}$$

3. The proposed bootstrap method

The idea of the proposed bootstrap results from motivating simulations studies where we usually observed properties of the design-variance estimators based on the original Quatember (2014) bootstrap better than that of competitors, but problems with

estimation of the design-MSE of some estimators and predictors using auxiliary information. To improve the method we propose to change the number of replications of sampled elements assumed by Quatember (2014) to be equal inverses of first order inclusion probabilities. Although these weights seem to be a natural choice, the choice is not the only and the best one – similarly to the choice between the Horvitz-Thompson estimator (using these weights to estimate the population total) and other estimators or predictors using different weighting systems, which usually lead to more accurate estimates than the Horvitz-Thompson estimator. Hence, below we propose to replace inverses of first order inclusion probabilities in the algorithm presented by Quatember (2014) by some calibration weights summing up to the population size, but other weighting systems are also possible.

To clarify considerations presented below, let us introduce the idea of the calibration estimator of the population total. It is given by (Deville, Särndal 1992):

$$\hat{\theta}^{CAL} = \sum_{k \in s} w_k y_k \ , \tag{5}$$

where weights w_k are solutions of:

$$\begin{cases} f_s(w_k, \pi_k^{-1}, q_k) \to \min \\ \sum_{k \in s} w_k \mathbf{x}_k = \sum_{l \in \Omega} \mathbf{x}_l \end{cases}, \tag{6}$$

where $f_s(w_k, \pi_k^{-1}, q_k)$ is some distance measure between weights of the calibration estimator w_k and the inverses of the first order inclusion probabilities π_k^{-1}, where for more generality additional known weights q_k can be included. The minimization in (6) leads to the approximate design-unbiasedness of the calibration estimator. The equality in (6) is the condition of model-unbiasedness of the estimator (5) under the linear model. If in (6) we additionally assume that:

$$f_s(w_k, \pi_k^{-1}, q_k) = \sum_{k \in s} \frac{\left(w_k - \pi_k^{-1}\right)^2}{\pi_k^{-1} q_k}, \tag{7}$$

then the resulting calibration estimator is called a generalised regression estimator (GREG) (Deville, Särndal 1992; Särndal, Swensson, Wretman 1992, p. 232; Rao, Molina 2015, p. 13). Deville and Särndal (1992) prove under some conditions that calibration estimators and the generalised regression estimator of the population total are asymptotically equivalent. But their values are very similar even for small sample sizes, as shown by Singh and Mohl (1996) and Stukel, Hidiroglou and Särndal (1996).

Our proposal of the bootstrap algorithm for simple random sampling without replacement is as follows. In the b th bootstrap sample ($b = 1, 2, ..., B$) element k is

drawn from the original sample in the j th draw ($j=1,2,...,n$) with probability (compare with (3)):

$$\left(w_k - h_{k,j-1}\right)\times(N-j+1)^{-1}, \tag{8}$$

where w_k's are some calibration weights such that $\sum_{i\in\Omega} w_k = N$ (e.g. calibration weights considered by Deville and Särndal (1992)).

Our proposal of the bootstrap algorithm for probability proportional to size sampling is as follows. In the b th bootstrap sample ($b=1,2,...,B$) element k is drawn from the original sample in the j th draw ($j=1,2,...,n$) with probability (compare with (3)):

$$\left(w_k - h_{k,j-1}\right)\times x_k\left(t_x - \sum_{i\in s_{bj-1}} x_i\right)^{-1}, \tag{9}$$

where w_k's are some calibration weights such that $\sum_{i\in\Omega} w_k = N$ (e.g. calibration weights considered by Deville and Särndal (1992)).

Of course, the choice of w_k's in the proposed algorithms is ambiguous (similarly to the choice of weights used in estimation). In the simulation studies, presented in the next section, we will consider four arbitrary chosen cases - calibration weights which fulfil four systems of calibration equations presented below. Firstly, we will consider weights w_{1k} ($k=1,2,...,n$) such that (Deville and Särndal (1992)):

$$\sum_{k\in s} w_{1k}\mathbf{x}_k = \sum_{l\in\Omega}\mathbf{x}_l \wedge \sum_{k\in s} w_{1k} = N \wedge L_k \le w_{1k} \le U_k, \tag{10}$$

where in simulation studies, to avoid negative and extremely large calibration weights, we will assume that $\forall_k L_k = 0$ and $\forall_k U_k = 10\pi_k^{-1}$. Secondly, we will consider weights w_{2k} ($k=1,2,...,n$) defined similarly to (10) but for domains:

$$\forall_d \sum_{k\in s_d} w_{2k}\mathbf{x}_k = \sum_{l\in\Omega_d}\mathbf{x}_l \wedge \forall_d \sum_{k\in s_d} w_{2k} = N_d \wedge L_k \le w_{2k} \le U_k, \tag{11}$$

where L_k and U_k are defined as in (10). Thirdly, we will consider weights w_{3k} ($k=1,2,...,n$), which leads to quantile calibration (similarly to Barbiero, Manzi and Mecatti 2015):

$$\sum_{k\in s} w_{3k} I(\mathbf{x}_k \le \mathbf{x}_p) = Np \wedge \sum_{k\in s} w_{3k} = N \wedge L_k \le w_{3k} \le U_k, \tag{12}$$

where $\mathbf{x}_p$ denotes the vector of population quantiles of auxiliary variables of order $p = \{0.25, 0.5, 0.75\}$, L_k and U_k are defined as in (10). Fourthly, we will consider weights w_{4k} $(k = 1, 2, ..., n)$ defined similarly to (12) but for domains:

$$\forall_d \sum_{k \in s_d} w_{4k} I(\mathbf{x}_k \leq \mathbf{x}_{dp}) = N_d\, p \wedge \forall_d \sum_{k \in s_d} w_{4k} = N_d \wedge L_k \leq w_{4k} \leq U_k\,, \qquad (13)$$

where $\mathbf{x}_{dp}$ denotes the vector of domain quantiles of auxiliary variables of order $p = \{0.25, 0.5, 0.75\}$, L_k and U_k are defined as in (10).

In cases (10) and (12) calibration equations are solved based on the whole sample, which may be a good solution in the case of estimation of population parameters. We hope that taking into account information on auxiliary variables in building pseudopopulation will give better properties of the design-variance and the design-MSE bootstrap estimators than in case of the algorithm proposed by Quatember (2014). What is more, in cases (11) and (13) calibration equations are solved based on samples in domains, taking into account domain-specific information on auxiliary variables, which should additionally lead to better results in the case of estimation of domain parameters.

4. Simulation study

We present results of a design-based simulation study conducted in R (R Development Core Team 2019). We use real data on $N = 281$ Swedish municipalities (Särndal, Swensson and Wretman 1992). We assume a relatively large sample size $n \approx 0.15N$ to show clearly differences between properties of different variance and MSE estimators. Revenues from 1985 municipal taxation (in millions of kronor) are the variable of interest, 1975 population (in thousands) – the auxiliary variable. We consider two subpopulations – the first one of size $N_1 = 104$, which consists of municipalities belonging to regions 1, 2 and 3; and the second of size $N_2 = 177$, which consists of municipalities belonging to regions 4-8. Large domains sizes will allow us to compare properties of estimators of design-variances and design-MSEs of direct and indirect estimators and predictors of domain totals. We consider probability proportional to size sampling using Brewer sampling scheme (Brewer 1975, Brewer and Hanif 1983). It is known to be a fast algorithm that does not cause problems in the case of asymmetry of the auxiliary variable as it can happen in the case of Rao-Sampford sampling scheme. However, in this sampling scheme there is a problem with computation of joint inclusion probabilities – a recursive formula is required and it implies a complete exploration of the splitting tree (Tillé 2006, p. 113).

In the simulation study we consider the problem of estimation of design-variances and design-MSEs of the following estimators and predictors:

- the Horvitz-Thompson (1952) estimator of the population total (which will be denoted by: HT) and of domains totals (HTd1, HTd2),
- the generalised regression estimator (e.g. Deville and Särndal 1992) of the population total (GREG) and of domains totals (GREGd1, GREGd2),
- the modified generalised regression estimator (e.g. Särndal 1981) of domains totals (MGREGd1, MGREGd2),
- the best linear unbiased predictor (e.g. Royall 1976) of domains totals (BLUPd1, BLUPd2) under the following model $Y_k = \beta_1 x_k + \beta_1 + \xi_k$, where $\xi_k \sim iid(0, \sigma^2)$.

We consider the following estimators of design-variances and design-MSEs of the above listed estimators and predictors:

- based on the Holmberg (1998) bootstrap (which will be denoted by H),
- based on the Antal and Tillé (2011) bootstrap (AT),
- based on the Quatember (2014) bootstrap (Q),
- the proposed generalised Quatember (2014) bootstrap with weights fulfilling calibration equations (10) (GQ1),
- the proposed generalised Quatember (2014) bootstrap with weights fulfilling calibration equations (11) (GQ2),
- the proposed generalised Quatember (2014) bootstrap with weights fulfilling calibration equations (12) (GQ3),
- the proposed generalised Quatember (2014) bootstrap with weights fulfilling calibration equations (13) (GQ4).

In the case of all bootstrap methods the number of bootstrap iterations equals $B = 1000$. Additionally, we consider classic design-variance estimators of the Horvitz-Thompson estimator and the GREG estimator (in both cases denoted by cl), where only first order inclusion probabilities are used. It results from the problems with computations of second order inclusion probabilities in Brewer sampling scheme described above. We use the following design-variance estimator of the Horvitz-Thompson estimator of the population total (Antal and Tillé 2011, p. 536):

$$\hat{D}^2(\hat{\theta}^{HT}) = \sum_{k=1}^{n} c_k \left(y_k \pi_k^{-1} - \sum_{k=1}^{n} c_k y_k \pi_k^{-1} \left(\sum_{k=1}^{n} c_k \right)^{-1} \right)^2, \tag{14}$$

where we use $c_k = n(1-\pi_k)(\mathrm{n}-1)^{-1}$ proposed by Hájek (1981), which gives efficient and only slightly biased design-variance estimator (Antal and Tillé 2014, p. 1348).

To estimate the design-variance of the GREG estimator we use the following one based on the Deville's method (Deville 1993):

$$\hat{D}^2(\hat{\theta}^{GREG}) = \left(1 - \sum_{k=1}^{n} a_k^2\right)^{-1} \sum_{k=1}^{n} (1-\pi_k)\left(e_k \pi_k^{-1} - A\right)^2, \tag{15}$$

where $a_k = (1-\pi_k)\left(\sum_{k=1}^{n}(1-\pi_k)\right)^{-1}$, $A = \sum_{k=1}^{n} a_k e_k \pi_k^{-1}$, $e_k = y_k - \mathbf{x}_k^T \mathbf{B}$, g_k - g-weights of GREG (see Deville and Särndal 1992), $\mathbf{B} = \sum_{k=1}^{n}\left(g_k \pi_k^{-1} \mathbf{x}_k \mathbf{x}_k^T\right)^{-1} \sum_{k=1}^{n}\left(g_k \pi_k^{-1} \mathbf{x}_k y_k\right)$. In the case of (14) and (15) replacing y_k with $a_{dk} y_k$, where $a_{dk} = 1$ if $k \in s_d$ and 0 otherwise, gives estimators of design-variances of estimators of domain totals.

In the simulation study we compute:

– the relative biases of the estimators of the design-variance of different estimators as

$$100\% \cdot V^{-1} \frac{1}{B} \sum_{r=1}^{R}\left(\hat{V}_r - V\right), \tag{16}$$

– the relative biases of the estimators of the design-MSE of different estimators as

$$100\% \cdot MSE^{-1} \frac{1}{B} \sum_{r=1}^{R}\left(\hat{MSE}_r - MSE\right), \tag{17}$$

– the relative RMSEs of the estimators of the design-variance of different estimators as

$$100\% \cdot V^{-1} \sqrt{\frac{1}{B} \sum_{r=1}^{R}\left(\hat{V}_r - V\right)^2}, \tag{18}$$

– the relative biases of the estimators of the design-MSE of different estimators as

$$100\% \cdot MSE^{-1} \sqrt{\frac{1}{B} \sum_{r=1}^{R}\left(\hat{MSE}_r - MSE\right)^2}, \tag{19}$$

where $\hat{V}_r$ and $\hat{MSE}_r$ are estimators of the design-variance and the design-MSE, respectively, obtained in the rth Monte Carlo iteration $r = 1, 2, ..., R$, whereas V is the simulation design-variance given by $V = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_d^r - \frac{1}{B}\sum_{r=1}^{R}\hat{\theta}_d^r\right)^2$, MSE is the simulation design-MSE given by $MSE = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}_d^r - \theta_d\right)^2$, $\hat{\theta}_d^r$ is the value of the estimator of the subpopulation total (or its special case – the estimator of the population

total denoted by $\hat{\theta}^r$) computed in the rth iteration, θ_d is the value of the subpopulation total (or its special case – the population total denoted by θ), the number of samples drawn in the Monte Carlo simulation study equals $R = 1000$.

Firstly, we would like to present design-based properties of the considered estimators and predictors. The Horvitz-Thompson estimator is design-unbiased and hence we will consider only its design-variance estimators. GREG is asymptotically design-unbiased estimator (Deville and Särndal 1992), MGREG is approximately p-unbiased if the overall sample size increases even if the domain sample size is small (Molina and Rao 2015, p. 22) – for these estimators usually only design-variance is estimated. Although their relative design-biases obtained in the simulation study are small (see Table A1 in Appendix) we also analyze properties of estimators of their design-MSEs. We also consider best linear unbiased predictors for which prediction-MSEs (not design-MSEs) are usually estimated. Although in our simulation study, their design-biases and design-MSEs are not large (see Table A1 in Appendix), including them will allows us to check properties of the proposed design-MSE estimators not only for design-unbiased or approximately design-unbiased statistics.

Secondly, we present main results of the simulation study. RRMSEs of estimators of design-variances and design-MSEs are presented in Tables 1-3 below, their design-biases in Tables A2-A4 in Appendix. If we compare relative design-biases (see Table A2 and Table A3 in Appendix) and RRMSEs (Table 1 and Table 2) of our proposals of design-variance estimators with bootstrap competitors, we see that usually the best results are obtained for one of the proposed methods or the results for our method are very close to the best one (except results for the HT estimator). Among four proposals (GQ1-GQ4) the GQ1 method is the best choice in most of the cases. If we compare RRMSEs (see Table 3) of our proposals of design-MSE estimators with bootstrap competitors, we obtain similar conclusions – results for GQ1 are usually the best or close to the best.

Table 1. RRMSEs in % of bootstrap estimators of design-variances – part 1

Method	HT	HTd1	HTd2	GREG	GREGd1	GREGd2
cl	27.6	9.8	7.6	26.7	13.7	12.5
H	27.6	10.7	9.0	38.4	17.7	16.8
AT	28.1	10.9	8.6	46.8	22.4	21.9
Q	29.6	11.9	9.6	32.0	11.7	10.7
GQ1	31.3	12.5	10.2	27.5	10.5	9.3
GQ2	31.8	13.5	9.7	28.2	12.2	10.6
GQ3	32.8	12.6	10.1	28.4	10.8	9.7
GQ4	34.4	14.1	10.1	30.1	13.7	12.9

Table 2. RRMSEs in % of bootstrap estimators of design-variances – part 2

Method	MGREGd1	MGREGd2	BLUPd1	BLUPd2
H	40.4	33.9	43.6	36.9
AT	44.1	39.9	45.0	37.4
Q	41.5	30.7	38.7	32.0
GQ1	43.2	29.2	36.0	25.6
GQ2	44.9	29.8	36.9	25.7
GQ3	42.5	30.1	35.7	23.0
GQ4	46.4	31.4	38.6	25.9

Table 3. RRMSEs in % of bootstrap estimators of design-MSEs

Method	GREG	GREGd1	GREGd2	MGREGd1	MGREGd2	BLUPd1	BLUPd2
cl	26.4*	13.8*	12.6*	- **	- **	- ***	- ***
H	39.0	17.6	16.9	40.6	34.8	64.7	38.1
AT	47.2	22.8	22.6	44.2	40.5	64.5	35.3
Q	31.9	11.9	11.0	41.5	30.8	67.0	30.9
GQ1	27.4	10.6	9.3	44.7	32.8	67.3	34.6
GQ2	67.2	176.0	177.5	95.2	45.3	65.3	83.5
GQ3	77.0	42.7	42.5	75.2	112.6	65.4	60.6
GQ4	123.4	183.9	179.6	166.2	113.0	66.1	113.8

* - design-variance estimator (15) is used to estimate design-MSE

** - classic design-MSE estimator not available due to the lack of second order inclusion probabilities

*** - design-MSE estimator not available (prediction-MSE is usually estimated)

5. Conclusions

We present a generalisation of the bootstrap algorithm for complex sampling designs proposed by Quatember (2014), used to estimate the design-variance and the design-MSE. We study its properties in the case of estimation of population total using the HT and GREG estimators and in the case of estimation of subpopulation total using the HT, GREG, MGREG estimators and the BLUP. In the simulation study based on real data we show that our proposal gives more accurate design-MSE and design-variance estimators in most of cases (or of similar accuracy in other cases) for estimators and predictors which use auxiliary information compared with the original algorithm and other bootstrap methods considered in the paper.

Acknowledgements

This paper was presented at the MSA 2019 conference, which financed its publication. Organization of the international conference "Multivariate Statistical Analysis 2019" (MSA 2019) was supported from resources for popularization of scientific activities of the Minister of Science and Higher Education in the framework of agreement No. 712/P-DUN/202019.

References

ANTAL, E., TILLÉ, Y., (2011). A Direct Bootstrap Method for Complex Sampling Designs From a Finite Population, *Journal of the American Statistical Association*, Vol. 106, No. 494, pp. 534–543.

ANTAL, E., Tillé, Y., (2014). A New Resampling Method for Sampling Designs Without Replacement: The Doubled Half Bootstrap*, Computational Statististic, V*ol. 29, No. 5, pp. 1345–1363.

BARBIERO, A., MANZI, G., MECATTI, F., (2015). Bootstrapping probability-proportional-to-size samples via calibrated empirical population, *Journal of Statistical Computation and Simulation*, Vol. 85, No. 3, pp. 608–620.

BARBIERO, A., MECATTI, F., (2010). Bootstrap algorithms for variance estimation in πPS sampling, In *Complex Data Modeling and Computationally Intensive Statistical Methods* edited by P. Mantovan and P. Secchi, pp. 2019–2026. Springer-Verlag, Italia.

BEAUMONT, J. F., PATAK, Z., (2012). On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling*, International Statistical Review*, Vol. 80, No. 1, pp. 127–148.

BREWER, K. E. W., (1975). A simple procedure for sampling πpswor*, Australian & New Zealand Journal of Statistics*, *V*ol. 17, No. 3, pp. 166–172.

BREWER, K. E. W., HANIF M., (1983). *Sampling with unequal probabilities,* Springer, New York.

DEVILLE, J. C., (1993). *Estimation de la variance pour less enquêtes en deux phases.* Manuscript, INSEE, Paris.

DEVILLE, J. C., SÄRNDAL, C. E., (1992). Calibration estimators in survey sampling, *Journal of the American Statistical Association*, Vol. 87, pp. 376–382.

EFRON, B., (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics*, Vol. 7, pp. 1–26.

HÁJEK, J., (1981). *Sampling From a Finite Population*, Marcel Dekker, New York.

HOLMBERG, A., (1998). A bootstrap approach to probability proportional to size sampling, *Proceedings of Section on Survey Research Methods*, American Statistical Association, Washington, pp. 378–383.

HORVITZ, D.G., THOMPSON, D. J., (1952). A Generalization of Sampling Without Replacement From a Finite Universe, *Journal of the American Statistical Association*, Vol. 47, No. 260, pp. 663–685.

QUATEMBER, A., (2014). The Finite Population Bootstrap – from the Maximum Likelihood to the Horvitz-Thompson Approach, *Austrian Journal of Statistics,* Vol. 43, pp. 93–102.

R DEVELOPMENT CORE TEAM, (2019). *A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna.

RANALLI, M. G., MECATTI, F., (2012). Comparing Recent Approaches for Bootstrapping Sample Survey Data: A First Step Towards a Unified Approach, *Proceedings of Section on Survey Research Methods*, American Statistical Association, Washington, pp. 4088–4099.

RAO, J. N. K, MOLINA, I., (2015). *Small area estimation*. *Second edition*, John Wiley and Sons, Hoboken, New Jersey.

RAO, J. N. K., WU, C. F. J., (1988). Resampling Inference for Complex Survey Data, *Journal of American Statistical Association*, Vol. 83, pp. 231–241.

ROYALL, R. M., (1976). The Linear Least Squares Prediction Approach to Two-Stage Sampling, *Journal of the American Statistical Association*, Vol. 71, pp. 657–473.

SÄRNDAL, C. E, (1981). Frameworks for Inference in Survey Sampling with Applications to Small Area Estimation and Adjustment for Nonresponse, *Bulletin of the International Statistical Institute*, Vol. 49, pp. 494–513.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

SEN, A. R., (1953). On the estimate of variance in sampling with varying probabilities, *Journal of the Indian Society of Agricultural Statistics,* Vol. 5, No. 2, pp. 119–127.

SINGH, A. C, MOHL, C. A., (1996). Understanding calibration estimators in survey sampling, *Survey Methodology*, Vol. 22, pp. 107–115.

SITTER, R. R., (1992). A Resampling Procedure for Complex Survey Data, *Journal of the American Statistical Association*, Vol. 87, pp. 755–765.

STUKEL, D. M., HIDIROGLOU, M. A., SÄRNDAL, C. E., (1996). Variance estimation for calibration estimators: A comparison of jackknifing versus Taylor linearization, *Survey Methodology*, Vol. 22, pp. 177–125.

TILLÉ, Y., (2006). *Sampling algorithms,* Springer-Verlag, New York.

YATES, F., GRUNDY, P. M., (1953). Selection Without Replacement from Within Strata with Probability Proportional to Size, *Journal of the Royal Statistical Society*, *Ser. B*, Vol. 15, pp. 235–261.

APPENDIX

Table A1. Relative design-biases and design-RRMSEs in % of considered estimators and predictors

estimator/predictor	relative bias (in %)	RRMSE (in %)
HT	-0.079	1.836
HTd1	-0.215	14.654
HTd2	0.037	11.498
GREG	-0.172	1.790
GREGd1	-0.799	15.592
GREGd2	0.359	12.232
MGREGd1	-0.208	2.804
MGREGd2	-0.148	2.190
BLUPd1	-3.100	3.809
BLUPd2	0.544	2.935

Table A2. Relative biases in % of bootstrap estimators of design-variances – part 1

Method	HT	HTd1	HTd2	GREG	GREGd1	GREGd2
Cl	9.5	0.6	-1.1	1.6	-6.8	-8.0
H	8.1	-1.4	-2.9	9.1	-0.8	-1.5
AT	9.6	0.6	-1.0	16.8	4.1	3.3
Q	13.5	5.6	3.5	6.2	0.6	0.2
GQ1	14.6	5.8	3.6	3.4	-2.5	-3.1
GQ2	14.8	8.8	6.3	3.8	3.3	3.2
GQ3	14.9	5.5	3.3	3.2	-2.9	-3.7
GQ4	15.2	9.4	7.0	3.8	1.7	1.7

Table A3. Relative biases in % of bootstrap estimators of design-variances – part 2

Method	MGREGd1	MGREGd2	BLUPd1	BLUPd2
H	3.2	1.8	7.8	9.3
AT	9.2	10.3	10.3	10.8
Q	7.6	-1.1	1.0	1.4
GQ1	8.6	-3.8	-1.7	-3.8
GQ2	11.4	-3.2	-1.7	-5.7
GQ3	7.0	-4.5	-1.8	-3.8
GQ4	10.2	-2.9	-2.5	-7.0

Table A4. Relative biases in % of bootstrap estimators of design-MSEs

Method	GREG	GREGd1	GREGd2	MGREGd1	MGREGd2	BLUPd1	BLUPd2
cl	0,6*	-7.1*	-8.1*	- **	- **	- ***	- ***
H	10.1	2.3	2.0	4.5	3.8	-63.0	9.6
AT	16.4	4.3	3.7	9.1	10.4	-62.8	7.1
Q	5.8	0.9	0.7	8.0	-1.0	-65.8	0.5
GQ1	2.9	-2.5	-2.9	10.3	-1.8	-66.1	2.6
GQ2	20.8	102.3	104.4	27.3	7.7	-60.9	18.1
GQ3	22.7	15.8	14.2	20.7	19.9	-63.1	11.2
GQ4	48.8	102.8	102.9	57.0	36.2	-51.5	43.5

* - design-variance estimator (15) is used to estimate design-MSE

** - classic design-MSE estimator not available due to the lack of second order inclusion probabilities

*** - design-MSE estimator not available (prediction-MSE is usually estimated)

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 179–195, DOI 10.21307/stattrans-2021-010
Received – 23.12.2019; accepted – 17.02.2021

Bankruptcy prediction of small- and medium-sized enterprises in Poland based on the LDA and SVM methods

Aneta Ptak-Chmielewska[1]

ABSTRACT

The impact the last financial crisis had on the small- and medium-sized enterprises (SMEs) sector varied across countries, affecting them on different levels and to a different extent. The economic situation in Poland during and after the financial crisis was quite stable compared to other EU member states. SMEs represent one of the most important segments of the economy of every country. Therefore, it is crucial to develop a prediction model which easily adapts to the characteristics of SMEs.

Since the Altman Z-Score model was devised, numerous studies on bankruptcy prediction have been written. Most of them involve the application of traditional methods, including linear discriminant analysis (LDA), logistic regression and probit analysis. However, most recent studies in the area of bankruptcy prediction focus on more advanced methods, such as case-based reasoning, genetic algorithms and neural networks. In this paper, the effectiveness of LDA and SVM predictions were compared. A sample of SMEs was used in the empirical analysis, financial ratios were utilised and non-financial factors were taken account of. The hypothesis assuming that multidimensional discrimination was more effective was verified on the basis of the obtained results.

Key words: discriminant analysis, support vector machines, bankruptcy prediction, SMEs.

1. Introduction

In Poland, the economic situation during the financial crisis 2009-2010 and after was quite stable compared to other European countries. Poland was a "green island" on the map of Europe.

The last financial crisis affected the SMEs sector in different countries at different levels and strength. Even among EU some economies suffered less compared to other ones. SMEs are the most important sector of the economy of every country. Therefore, we need a prediction model that easily adapts to SMEs characteristics.

Poland was the leader of growth among all OECD countries (Raport… 2012, p.9). Faster than ever we decreased the distance to Western Europe. We still kept high

[1] Warsaw School of Economics, Poland. E-mail: aptak@sgh.waw.pl. ORCID: https://orcid.org/0000-0002-9896-4240.

financial credibility, avoided the recession and dramatic currency crises or debt crises. Those phenomena deeply concerned other European countries including our central and eastern European neighbours. One of the pillars of this success was effectively operating Polish companies. Corporations were able to flexibly adjust the operations to the crises environment. Adjustment abilities among small companies were of course obvious but their expectations about finishing the recession were more visible. This behaviour is not the most effective from the company's point of view but is a visible sign of Polish transformation success. Our companies are doing as well as their highly capitalized competitors from abroad. They have the past history full of hard conditions of survival and now they are more resistant to the crisis. SMEs are mostly flexible and adjustable (due to low employment and expenses limitations) to changing economic conditions. Resistibility to crises depends also on low export level. SMEs do not use the external funding and are more conservative to expansion and in consequence are not involved in risky financial operations and big investment projects. Majority of Polish SMEs are involved in services sector which suffer the least from the crises. SMEs react on crises decreasing the employment. Orłowski et al. (2010) conducted the project to give the answer to the questions: is SME sector able to overcome this crisis by its own? Or any external help is needed? Can this crisis be helpful in improvement of enterprises functioning? Results show that for some SMEs this crisis is more a "medial" fact than real. They are more pessimistic about the influence of this crisis on the situation in the country and the sector than on an individual enterprise. Opinions are formulated not only by the enterprises but also by media and social opinions (pessimism). Only medium enterprises and exporters declare that they have to cope with real problems. Enterprises do cope with less number of orders but they do not feel lower profitability and delays in payments, which are more characteristic for the recession. Crisis is not the main threat for enterprises' functioning. More dangerous are typical difficulties like taxes, competition. Reaction among SMEs is more often passive than active like restructuring or new markets.

Since the Altman's Z-Score model (Altman 1968), many bankruptcy prediction studies have been written. Most of them use the traditional methods, like discriminant analysis, logistic regression (Back et al. 1996), probit analysis (Zmijewski 1984). Recent studies in this area are focused on more advanced methods, like case-based reasoning CBR (Bryant 1997; Yip 2006; Sartori, Mazzucchelli, Di Gregorio 2016), genetic algorithms GA (Back et al. 1996) and neural networks NN (Desai et al, 1996, Abdou et al. 2008, Derelioğlu and Gürgen, 2011, Blanco et al. 2013) or support vector machines SVM (Huang et al, 2004, Kim et al. 2010).

In their publication Sartori, Mazzucchelli, Di Gregorio (2016) used the case-based reasoning (CBR) method to forecast the bankruptcy and compared the results with the Z-Score model. The authors found that the CBR approach could be useful for clustering

enterprises according to similarity metrics. Another method used for bankruptcy prediction was Genetic algorithms (GAs). Gordini (2014) compared the genetic algorithms with two other methods, namely logistic regression (LR) and support vector machine (SVM). The results suggest that GAs is a quite effective and promising compared with LR and SVM, especially in small misclassification rate type II. In this paper the size of companies and the geographical area were analysed. Both characteristics seem to influence the accuracy of the models. The author built separate models for separate geographical areas. GAs prediction accuracy in each area was higher than that of the other models separately. Sohn, Kim and Yoon (2016) applied fuzzy logistic regression for the default prediction models. According to the authors the proposed approach outperforms the logistic regression in terms of discriminatory power. Similarly, Chaudhuri and De (2011) also used the fuzzy model but using support vector machines. Those models outperformed traditional bankruptcy prediction models. According to Psillaki, Tsolas, Margaritis (2010) non-financial indicators are also useful. They proved that management is an important indicator of enterprises' risk, mostly financial risk. More efficient firms and firms with more liquid assets are much less risky. Analogous to Kalak and Hudson (2016), who emphasized the differences between small and micro firms, also Gupta et al. (2015) investigated how the SMEs size can affect bankruptcy risk. Their research results suggest that separate models for micro firms are desired. In case of small and medium companies, there is no such a need as the determinants present a similar level of hazard. A two-stage genetic programming (2SGP) model was proposed by Huang et al. (2006). This approach achieves better results. Also, Berg (2007) used accounting-based models for bankruptcy prediction. The generalized additive models are more effective compared to models: linear discriminant analysis, neural networks and generalized linear models. Modina and Pietrovito (2014) identified that capital structure and interest expenses of SMEs play more important role than economic characteristics while specifying the determinants of company's default probability. The approach presented by Andreeva et al. (2014) combines the use of Generalized Extreme Value (GEV) regression. Additionally authors compared two different ways of treating the missing values, namely multiple imputation and the Weights of Evidence approach. According to the results obtained, the BGEVA approach outperforms the logistic regression, where in the case of missing values WoE showed better results. In order to identify defaulted SMEs, Calabrese et al. (2015) investigated a binary regression accounting-based model. Results obtained suggest that their approach outperformed the classical logistic regression model for different default horizons considered.

The literature of bankruptcy prediction in Poland is very reach. When first bankruptcies were registered in 1990 (economic transition) the researchers started to be interested in this subject. It is not possible to mention all the literature from this area.

Only some of the articles can be cited. One of the first authors who apply the Altman model for Polish enterprises was Mączyńska (1994). Researchers used financial ratios at that time and built national models for the bankruptcy prediction (Wędzki 2000; Stępień, Strąk 2003; Prusak 2005, Pogodzińska, Sojak 1995; Gajdka, Stos 1996; Hadasik 1998; Wierzba 2000). They used multivariate discriminant functions based on small samples of data. Next, analysis were expanded to more frequent samples and more advanced statistical models, logit models (Hołda 2001; Sojak and Stawicki 2000; Gruszczyński 2003; Michaluk 2003; Mączyńska 2004; Appenzeller, Szarzec 2004; Korol 2004; Hamrol et al. 2004; Wędzki 2004; Stępień, Strąk 2004; Prusak, Więckowska 2007; Jagiełło 2013; Pociecha et al. 2014; Karbownik 2017). Only recently more advanced data mining and survival models have been applied (Pociecha et al. 2014; Ptak-Chmielewska 2016, Korol 2010b; Gąska 2016; Zięba et al. 2016). Many models for different sectors and sizes of enterprises have been estimated (Jagiełło 2013; Brożyna et al. 2016; Balina, Bąk 2016; Karbownik 2017). Model that used macroeconomic variables and non-financial variables have been much rarer (Korol 2010a; Ptak-Chmielewska and Matuszyk 2017). Only one of the papers included the economic cycle and suggested that enterprise bankruptcy prediction models should include measures reflecting changes in economic conditions (Pociecha and Pawełek 2011). Prusak (2018) proposed a very original method of research using a literature review as the database for rating model. His database covered a long period (Q4 2016–Q3 2017) mostly from Google Scholar and ResearchGate and wide space of Central and Eastern European countries. He collected material including countries like: Poland, Latvia, Lithuania, Estonia, Russia, Ukraine, Hungary, Slovakia, Czech Republic, Bulgaria, Romania, Belarus. Based on this literature review, he proposed the ratings (Prusak 2018, p. 17) from Rating 0 — no studies in enterprise bankruptcy prediction in a given country, up to Rating 4 — most advanced methods are utilised in enterprise bankruptcy prediction in this country and researchers proposed new solutions that affect discipline development. According to this assessment, Poland got the highest grade equal to 4.0 (Prusak 2018, p.17). Only Czech Republic got a grade as high as Poland (4.0).

The main goal of this paper is to check and verify the effectiveness of multidimensional discrimination like support vector machines in comparison with traditional linear multivariate discrimination. Empirical verification was based on the sample of data for Polish enterprises. Discriminant analysis and support vector machines methods were applied as a research tool. Multivariate discriminant analysis is based on linear discrimination. Support vector machines are based on hyperplane discrimination and can be considered as more advanced and flexible discrimination compared do linear discriminant analysis. High accuracy of prediction bankruptcy in the small and medium enterprises (SME) sector is always in scope of interest not only

researchers but also policy makers and business. This paper contributes to research area on this topic.

2. Materials, Methods and Results

The database used in this paper covered 806 enterprises. The sample was quite balanced consisting of 311 bankruptcies and 495 firms in a good condition. Firms were mostly small and medium enterprises (SME) from the Polish market. The balanced sample enables to estimate the robust misclassification rates. The information about financial ratios came from financial statements covering the homogenous period of 3 years 2008-2010. The bankruptcy events were recorded for the period 2009-2012. Only standard 12-month period of observation was taken for analysis. The data sample (anonymous) was delivered by one of the consulting firm form Poland. Only selected 16 financial ratios were considered (see Table 1).

Table 1. Financial ratios utilised in the models (calculated at the date of FS)

Ratio	Name	Formula
X_1	current liquidity	$\frac{\text{current assets}}{\text{short term liabilities}}$
X_2	quick ratio	$\frac{\text{current assets} - \text{inventory} - \text{prepayments}}{\text{short term liabilities}}$
X_3	cash liquidity	$\frac{\text{cash}}{\text{short term liabilities}}$
X_4	capital share in assets	$\frac{\text{current assets} - \text{short term liabilities}}{\text{total assets}}$
X_5	gross margin	$\frac{\text{gross profit/loss on sale}}{\text{operating expenses}}$
X_6	operating profitability of the sales	$\frac{\text{profit/loss on operating activities}}{\text{total revenues}}$
X_7	operating profitability of the assets	$\frac{\text{profit/loss on operating activities}}{\text{total assets}}$
X_8	net profitability of the equity	$\frac{\text{net profit/loss}}{\text{equity}}$
X_9	assets turnover	$\frac{\text{total revenues}}{\text{total assets}}$
X_10	current assets turnover	$\frac{\text{total revenues}}{\text{current assets}}$

Table 1. Financial ratios utilised in the models (calculated at the date of FS) (cont.)

Ratio	Name	Formula
X_11	receivables turnover	$\frac{\text{total revenues}}{\text{receivables}}$
X_12	inventory turnover	$\frac{\text{total revenues}}{\text{inventory}}$
X_13	capital ratio	$\frac{\text{equity}}{\text{total liabilities}}$
X_14	coverage of the short-term liabilities by the equity	$\frac{\text{equity}}{\text{short term liabilities}}$
X_15	coverage of the fixed assets by the equity	$\frac{\text{equity}}{\text{fixed assets}}$
X_16	share of the net financial surplus in the total liabilities	$\frac{\text{net profit/loss} + \text{amortisation} + \text{interests}}{\text{total liabilities}}$

Financial ratios were only static for one period, not including dynamic ratios. Among non-financial variables we considered only five of them due to limited information. Non-financial variables are presented in Table 2. In our opinion the selection of variables is very important in analysis of bankruptcy. In most cases only financial ratios are considered (Du Jardin 2009), but very often it is not sufficient to achieve a very good prediction accuracy of the model. For our purpose the sample was partitioned in the proportion 70%:30%.

Univariate analysis was based on t-test for interval variables and chi-sqr test for binary variables (significance level 0.1). Additionally, the correlation analysis eliminated correlated ratios ($r>0.7$). For the model only 9 variables were selected. Financial ratios used in the models estimation were the following: current liquidity, gross margin, operating profitability of sales, assets turnover, current assets turnover, capital ratio, coverage of short-term liabilities by equity. Non-financial factors (binary) used in estimation were the following: region_low_risk and legal_form_group2.

Table 2. Non-financial factors utilized in the models

Name	Attributes/categories
Sector of activity	Equal proportion of companies from sectors: Production, Trade and Services. This variable was dichotomized and reference category was set to Services (lowest risk of bankruptcy).
Cluster of regions	16 regions grouped into 3 clusters according to bankruptcy rate ("low risk", "average risk", "high risk") and dichotomized. Reference category was set to "high risk" group. Clusters grouped by k-means clustering method based on bankruptcy rate.

Table 2. Non-financial factors utilized in the models (cont.)

Name	Attributes/categories
Legal form	group1: limited liability company and group2: joint stock company, limited partnership company, other (cooperative, association, etc.). Reference category was set to group1.
Age of the company	Interval variable (age in completed years at the start of the observation period).
Number of employees	Interval variable (number of employed workers on the date of FS).

One of the most frequently used methods in bankruptcy prediction is discriminant analysis. The main idea of this method is to classify all cases (individuals) into two (or more) classes. In this case into two classes: bankrupted and non-bankrupted enterprises. Discriminant (linear) function is used to classify using the training sample. To construct this function the explanatory variables are used. Those variables must be carefully selected, must follow the normality distribution and must not be correlated to each other. The classification and the function construction is based on the criterion of maximization the distance between groups. A very important assumption in this method is the assumption of equality of variances between subsamples. This assumption must be positively verified before drawing conclusions on the results received from analysis.

The estimation of misclassification errors is not biased when the sample is close to balanced (equal proportion of observations in both groups). In the classic approach the discriminant function is linear – the Fisher discriminant function (Ptak-Chmielewska 2012):

$$Z = a_0 + a_1X_1 + a_2X_2 + \ldots + a_nX_n,$$

where:

Z – dependent variable,

a_0 – intercept,

a_i, $i = 1, 2, \ldots, n$ – discriminative coefficients (weights),

$X_1, X_2, \ldots, X_n$ – explanatory variables (financial ratios).

After estimating the value of this discriminant function the cut-off point value must be set up to classify all possible cases (firms) into a potentially bankrupted or non-bankrupted firm. The most frequently used method for cut-off calculation is the half of the distance between averages of the values in two groups. If the value of this function for a particular enterprise is below this cut-off point than the enterprise is classified as potentially bankrupted and in the case where it is below this cut-off value the enterprise is classified as potentially non-bankrupted one. Discriminant model applies simultaneously many variables using weights. It transforms multivariable space into

one dimension. It is possible to estimate and interpret the impact of each variable on the dependent one. A very important advantage of this model is that it can be applied on a small sample and by this it is useful in bankruptcy prediction. This method is available in all popular statistical packages (Ptak-Chmielewska 2012). There are also some disadvantages of this method. The possibility of using qualitative variables is very limited. The assumption of normality distribution is very often violated. This assumption is not critical. More critical is the assumption of equality of variances between groups. Sometimes this assumption is also violated. If the sample is not balanced we must remember about the decision matrix to be specified (cost of misclassification). Sometimes the linear separation is not optimal, and non-linear solution is more effective. The probability of bankruptcy is not estimated directly like in logistic regression (Ptak-Chmielewska 2012). Despite such a limitation this method is still popular in prediction of bankruptcy. The final estimated models with linear discriminant functions are presented in Table 3.

Table 3. Discriminant analysis – results

Discriminant linear function	0	1
Intercept	-0.56368	-0.60428
X_1 - current liquidity	0.00674	0.00919
X_5 – gross margin	0.00457	0.05883
X_6 – operating profitability of the sales	0.31414	-1.97663
X_9 – assets turnover	0.01844	-0.02436
X_10 – current assets turnover	0.07558	0.16014
X_13 – capital ratio	0.19445	0.08945
X_14 – coverage of the short-term liabilities by the equity	-0.00100	0.00662
region_low_risk (binary)	1.24260	0.39534
legal_form2 (binary)	1.65670	0.95146

All included explanatory variables were significant at least at the significance level 0.1. A higher value of gross margin classifies to bankrupted enterprises. Higher values of operating profitability of sales ratio and assets turnover ratio classify to non-bankrupted enterprises. But higher values of current assets turnover ratio classify to the bankrupted enterprises group. High capital ratio classifies to the good enterprises group. Activity situated in low risk region classifies to the non-bankrupted enterprises group.

The accuracy of the model was presented in Table 4. Overall accuracy on the training sample amounted to 68.5%. Among bankrupted enterprises the accuracy was comparable to non-bankrupted and amounted to 68.2% and 68.7% respectively.

Table 4. Classification table for train sample – linear discrimination

	Model=0	Model=1	Total
Level=0	237	108	345
Level=1	69	148	217
Total	306	256	562

Support Vector Machines (SVM) is a model based on the decision planes that define decision boundaries. A decision plane separates a set of objects into different classes. Most classification tasks are not as simple as linear classification, and often more complex structures are necessary to get the optimal separation. Optimal separation is the correct classification of new objects (test cases) based on the available train cases. This classification tasks which are drawing separating lines to classify objects of different class memberships are called hyperplane classifiers. Support Vector Machines can handle such complex tasks. The objects are mapped using mathematical functions (kernels). This process of classifying the objects is known as mapping.

Support Vector Machine (SVM) is a method for classifying that performs tasks by constructing hyperplanes in a multidimensional space. This method separates cases from different classes. SVM can be both regression and classification. It utilises multiple continuous and categorical variables (categorical variables are transformed into dummies). To find an optimal hyperplane, SVM uses an iterative algorithm to minimize an error function. SVM models can be classified into four groups (error function) two for classification: C-SVM, nu-SVM and two for regression: epsilon-SVM, nu-SVM regression.

SVM for classification Type 1 (C-SVM), with the error function to be maximized:

$$\frac{1}{2}\boldsymbol{w^T w} + \boldsymbol{C}\sum_{i=1}^{N}\boldsymbol{\xi_i}$$

with the constraints:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0, i = 1, \ldots, N$$

where:

C - the capacity constant,

w - the vector of coefficients,

b - constant,

$\boldsymbol{\phi_i}$ - parameters to handle nonseparable data (inputs).

Additionally, y∈±1 represents class labels, x_i represents the independent variables. The kernel $\boldsymbol{\phi}$ is used for data transformation from the input to the feature space. Larger C, more penalize the error, and should be chosen carefully to avoid over-fitting.

SVM for classification Type 2 (nu-SVM), with the error function to be maximized:

$$\frac{1}{2}w^T w - vp + \frac{1}{N}\sum_{i=1}^{N}\xi_i$$

with the constraints:

$$y_i(w^T\phi(x_i)+b) \geq \rho - \xi_i \text{ and } \xi_i \geq 0, i = 1, \ldots, N \text{ and } \rho \geq 0.$$

The most popular kernels used in Support Vector Machines models are: linear, polynomial, radial function (RBF) or sigmoid:

Kernel Functions

$$K(X_i, X_j) = \begin{cases} X_i \cdot X_j & \text{Linear} \\ (\gamma X_i \cdot X_j + C)^d & \text{Polynomial} \\ \exp\left(-\gamma |X_i - X_j|^2\right) & \text{RBF} \\ \tanh(\gamma X_i \cdot X_j + C) & \text{Sigmoid} \end{cases}$$

where $K(X_i,X_j)=\phi(X_i)\cdot\phi(X_j)$ is the kernel function with Gamma as an adjustable parameter. The most popular choice is RBF to be used in Support Vector Machines.

As the final SVM – Support Vector Machines model was estimated with the method of optimization using the Interior Point with polynomial function (3 degree) for scaling. The C method for penalization was used with the parameter equal to 1. Maximum iterations set to 50 with tolerance 1e-08.

Table 5. Results of SVM train

Internal weights (product)	39.4414518
Burden	-22.224232
Violation of restrictions	337.700692
Longest vector	14.1651387
Number of support vectors	383
Number of support vectors on margin	373
Maximum value of F	3.36246618
Minimum value of F	-21.635810
Number of included effects	9
Data matrix (number of columns)	9
Kernel matrix (number of columns)	**220**

The first type error for the train sample (wrongly classified defaults) was equal to 0.49. The second type error (wrongly classified good firms) was to 0.07 (see Table 6).

Table 6. Classification table for train sample – SVM

	Model=0	Model=1	Total
Level=0	338	7	345
Level=1	145	72	217
Total	483	79	562

3. Models comparison

The comparison of models' accuracy was based on a test sample (see Table 7 and 8) and the overall accuracy as well as bankruptcy prediction accuracy was compared. Also bankrupted enterprises prediction accuracy (sensitivity) was compared. Despite higher overall accuracy for SVM the specificity was higher for the DISCRIM model.

Generally, in the literature only the overall prediction accuracy is counted. The input from this analysis comparing the DISCRIM and the SVM model shows that specificity is more important in bankruptcy prediction. The cost of wrong classification for enterprises in good condition is not as high as the cost of wrong classification for a bankrupted enterprise.

Applying machine learning techniques including SVM is not always the best solution in bankruptcy prediction. Simple models, understandable and clear interpretation bring more value in understanding bankruptcy risk drivers.

Table 7. Classification table for test sample – discriminant analysis and SVM

Discriminant analysis			
	Model=0	Model=1	Total
Level=0	92	57	149
Level=1	32	63	95
Total	124	120	244
SVM			
	Model=0	Model=1	Total
Level=0	142	7	149
Level=1	61	34	95
Total	203	41	244

Table 8. Overall accuracy and prediction of bankruptcy – comparison

Model	Overall accuracy	Bankruptcy prediction accuracy	Good SMEs prediction accuracy
DISCRIM	63.5%	66.3%	61.7%
SVM	72.1%	35.8%	95.3%

4. Discussion and Conclusions

In this paper the accuracy of two different methods was assessed. The first method was popular discriminant function, called the Fisher discrimination. The second was Support Vector Machines, recently developed and applied for classification. According to the results, the overall accuracy of SVM was higher compared to linear discrimination but the accuracy of bankrupted enterprises prediction (sensitivity) was higher in the case of simple linear discrimination (almost double).

False alarms for enterprises in good standing are much less costly compared to wrong classification of bankrupted ones. Such models should first of all classify bankrupted enterprises. Only with higher accuracy of bankrupted enterprises (sensitivity) make such models useful for an early warning process. Implication for policy is a possible application of such a model (based on the DISCRIM model) in an early warning signal process. It is quite an easy tool for business to assess the credibility of partners and customers (let us say suppliers).

Comparing the effectiveness of different models based only on AUC (Area Under the ROC Curve) or AR (Accuracy Ratio) does not give the full picture of the real accuracy (Zięba et al. 2016). We should always take a look on the classification of both sides: bankrupted and non-bankrupted enterprises. In this field this research brings added value to the bankruptcy prediction research area. Typically, the literature and applications focus on general accuracy measures, not being cautious about bankrupted enterprises misclassifications.

Results highlighted the importance of financial ratios like current liquidity, gross margin, operating profitability of the sales, assets turnover and the current assets turnover, capital ratio, coverage of short-term liabilities by the equity. Also non-financial information was important like the region of activity, legal form of the enterprise. This confirms the importance of different non-financial, macroeconomic aspects, etc.

Future research should focus on specificity of SME's financial analysis. In comparison with corporations, the information about SMEs is always limited. Not only financial ratios are important, non-financial information plays also an important role. Analysis cannot be based only on information on financial ratios.

Acknowledgements

This paper was presented at the MSA 2019 conference, which financed its publication. Organization of the international conference "Multivariate Statistical Analysis 2019" (MSA 2019) was supported from resources for popularization of scientific activities of the Minister of Science and Higher Education in the framework of agreement No. 712/P-DUN/202019.

References

ABDOU, H., POINTON, J., MASRY, A. E., (2008). Neural Nets Versus Conventional Techniques in Credit Scoring in Egyptian Banking. Expert Systems with Applications, 35(2), pp. 1275–1292.

ALTMAN, E. I., (1968). Financial ratios, Discriminant analysis and the prediction of corporate bankruptcy. Journal of Finance, 23(4), pp. 589–609.

ANDREEVA, G., CALABRESE, R., OSMETTI, S. A., (2014). A comparative analysis of the UK and Italian small businesses using Generalised Extreme Value models. https://arxiv.org/pdf/1412.5351.pdf.

APPENZELLER, D., SZARZEC, K., (2004). Forecasting the bankruptcy risk of Polish public companies. Rynek Terminowy, 1, pp. 120–128.

BACK, B., LAITINEN, T., SERE, K., VAN WEZEL, M., (1996). Choosing bankruptcy predictors using discriminant analysis, logit analysis, and genetic algorithms, Technical Report, Turku Centre for Computer Science.

BALINA, R., BĄK, M. J., (2016). Discriminant Analysis as a Prediction Method for Corporate Bankruptcy with the Industrial Aspects. Waleńczów: Wydawnictwo Naukowe Intellect.

BERG, D., (2007). Bankruptcy prediction by generalized additive models. Applied Stochastic Models in Business and Industry, 23(2), pp. 129–143.

BLANCO, A., PINO-MEJÍAS, R., LARA, J., (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from Peru, Expert Systems with Applications, 40(1), pp. 356–364.

BROŻYNA, J., MENTEL, G., PISULA, T., (2016). Statistical methods of the bankruptcy prediction in the logistics sector in Poland and Slovakia. Transformations in Business & Economics, 15, pp. 80–96.

BRYANT, S. M., (1997). A case-based reasoning approach to bankruptcy prediction modelling. Intelligent Systems in Accounting, Finance and Management, 6(3), pp. 195–214.

CALABRESE, R., MARRA, G., OSMETTI, S. A., (2015). Bankruptcy prediction of small and medium enterprises using a flexible binary generalized extreme value model. Journal of the Operational Research Society, 67(4).

CHAUDHURI, A., DE, K., (2011). Fuzzy support vector machine for bankruptcy prediction. Applied Soft Computing, 11(2), pp. 2472–2486.

DERELIOĞLU, G., GÜRGEN F., (2011). Knowledge discovery using neural approach for SME's credit risk analysis problem in Turkey. Expert Systems with Applications, 38(8), pp. 9313–9318.

DESAI, V. S., CROOK, J. N., OVERSTREET, G. A., (1996). A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment. European Journal of Operational Research, 95(1), pp. 24–47.

DU JARDIN, P., (2009). Bankruptcy prediction models: How to choose the most relevant variables?. Bankers, Markets & Investors, 98, pp. 39–46.

GAJDKA, J., STOS, D., (1996). Wykorzystanie analizy dyskryminacyjnej do badania podatności przedsiębiorstwa na bankructwo. In: J. Duraj ed. Przedsiębiorstwo na rynku kapitałowym, Wydawnictwo Uniwersytetu Łódzkiego, Łódź.

GĄSKA, D., (2016). Predicting Bankruptcy of Enterprises with the use of Learning Methods. Ph.D. dissertation, Wrocław University of Economics.

GORDINI, N., (2014). A genetic algorithm approach for SMEs bankruptcy prediction: Empirical evidence from Italy, Expert Systems with Applications, 41(14), pp. 6067–6536.

GRUSZCZYŃSKI, M., (2003). Models of microeconometrics in the analysis and forecasting of the financial risk of enterprises. Zeszyty Polskiej Akademii Nauk, 23.

GUPTA, J., GREGORIOU, A., HEALY, J., (2015). Forecasting bankruptcy for SMEs using hazard function. A review of quantitative finance and accounting, 45 (4), pp. 845–869.

HADASIK, D., (1998). The Bankruptcy of Enterprises in Poland and Methods of its Forecasting. Wydawnictwo Akademii Ekonomicznej w Poznaniu, 153.

HAMROL, M., CZAJKA, B., PIECHOCKI, M., (2004). Enterprise bankruptcy–discriminant analysis model. Przegląd Organizacji, 6, pp. 35–39.

HOŁDA, A., (2001). Forecasting the bankruptcy of an enterprise in the conditions of the Polish economy using the discriminatory function ZH. Rachunkowość, 5, pp. 306–10.

HUANG, Z., CHEN, H., HSU, C. J., CHEN, W. H., WU, S., (2004). Credit Rating Analysis with Support Vector Machines and Neural Networks: A Market Comparative Study. Decision Support System, 37(4), pp. 543–558.

HUANG, J. J., TZENG, J. H., ONG, C. S., (2006). Two-stage genetic programming (2sgp) for the credit scoring model. Applied Mathematics and Computation, 174, pp. 1039–1053.

JAGIEŁŁO, R., (2013). Discriminant and Logistic Analysis in the Process of Assessing the Creditworthiness of Enterprises. Materiały i Studia, 286. Warszawa: NBP.

KALAK I. E., HUDSON, R., (2016). The effect of size on the failure probabilities of SMEs: An empirical study on the US market using discrete hazard model. International Review of Financial Analysis, 43, pp. 135–145.

KARBOWNIK, L., (2017). Methods for Assessing the Financial Risk of Enterprises in the TSI Sector in Poland. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

KIM, H. S., SOHN, S. Y., (2010). Support Vector Machines for Default Prediction of SMEs Based on Technology Credit. European Journal of Operational Research, 201(3), pp. 838–846.

KOROL, T., PRUSAK, B., (2009). Upadłość przedsiębiorstwa a wykorzystanie sztucznej inteligencji, Warszawa: CeDeWu.

KOROL, T., (2004). Assessment of the Accuracy of the Application of Discriminatory Methods and Artificial Neural Networks for the Identification of Enterprises Threatened with Bankruptcy. Gdańsk: Doctoral dissertation.

KOROL, T., (2010a). Early Warning Systems of Enterprises to the Risk of Bankruptcy. Warszawa: Wolters Kluwer.

KOROL, T., (2010b). Forecasting bankruptcies of companies using soft computing techniques. Finansowy Kwartalnik Internetowy "e-Finanse", 6, pp. 1–14.

MĄCZYŃSKA, E., (1994). Assessment of the condition of the enterprise. Simplified methods. Życie Gospodarcze, 38, pp. 42–45.

MĄCZYŃSKA, E., (2004). Early warning systems. Nowe Życie Gospodarcze, 12, pp. 4–9.

MICHALUK, K., (2003). Effectiveness of corporate bankruptcy models in Polish economic conditions. In: L. Pawłowicz, R.Wierzba ed. Corporate Finance in the Face of Globalization Processes. Warszawa: Wydawnictwo Gdańskiej Akademii Bankowej.

MODINA, M., PIETROVITO, F., (2014). A default prediction model for Italian SMEs: the relevance of the capital structure. Applied Financial Economics, 24(23), pp. 1537–1554.

ORŁOWSKI, W., PASTERNAK, R., FLAHT, K., SZUBERT, D., (2010). Procesy inwestycyjne i strategie przedsiębiorstw w czasach kryzysu, Raport PARP Warszawa.

POCIECHA, J., PAWEŁEK, B., (2011). Bankruptcy Prediction and Business Cycle, Contemporary Problems of Transformation Process in the Central and East European Countries. Paper presented at 17th Ukrainian-Polish-Slovak Scientific

Seminar, Lviv, Ukraine, September 22–24; Lviv: The Lviv Academy of Commerce, pp. 9–24.

POCIECHA, J., PAWEŁEK, B., BARYŁA, M., AUGUSTYN, S., (2014). Statistical Methods of Forecasting Bankruptcy in the Changing Economic Situation. Kraków: Fundacja Uniwersytetu Ekonomicznego w Krakowie.

POGODZIŃSKA, M., SOJAK, S., (1995). The Use of Discriminant Analysis in Predicting Bankruptcy of Enterprises. Ekonomia XXV, 299.

PRUSAK, B., (2018). Review of Research into Enterprise Bankruptcy Prediction in Selected Central and Eastern European Countries. International Journal of Financial Studies, 6, 60.

PRUSAK, B., WIĘCKOWSKA, A., (2007). Multidimensional models of discriminant analysis in the study of the bankruptcy risk of Polish companies listed on the WSE. In: B. Prusak ed. Economic and Legal Aspects of Corporate Bankruptcy. Warszawa: Difin.

PRUSAK, B., (2005). Modern Methods of Forecasting Financial Risk of Enterprises. Warszawa: Difin.

PSILLAKI, M., TSOLAS, I. E., MARGARITIS, D., (2010). Evaluation of credit risk based on firm performance. European Journal of Operational Research, 201 (3), pp. 873–881.

PTAK-CHMIELEWSKA, A., MATUSZYK, A., (2017). The importance of financial and non-financial ratios in SMEs bankruptcy prediction. Bank i Kredyt, 49(1), pp. 45–62.

PTAK-CHMIELEWSKA, A., (2016). Statistical Models for Corporate Credit Risk Assessment–Rating Models. Acta Universitatis Lodziensis Folia Oeconomica, 3, pp. 98–111.

PTAK-CHMIELEWSKA, A., (2012). Wykorzystanie modeli przeżycia i analizy dyskryminacyjnej do oceny ryzyka upadłości przedsiębiorstw. Ekonometria, 4 (38), pp. 157–172.

POLSKA AGENCJA ROZWOJU PRZEDSIĘBIORCZOŚCI, (2012). Raport o stanie sektora małych i średnich przedsiębiorstw w Polsce w latach 2010–2011, Warsaw.

SARTORI, F., MAZZUCCHELLI, A., DI GREGORIO, A., (2016). Bankruptcy forecasting using case-based reasoning: the CRePERIE approach. Expert Systems with Applications, 64, pp. 400–411.

SOHN, S. Y., KIM, D. H., YOON, J. H., (2016). Technology credit scoring model with fuzzy logistic regression. Applied Soft Computing, 43, pp. 150–158.

SOJAK, S., STAWICKI, J., (2001). Wykorzystanie metod taksonomicznych do oceny kondycji ekonomicznej przedsiębiorstw. Zeszyty Teoretyczne Rachunkowości, 3(59), pp.45–52.

STĘPIEŃ, P., STRĄK, T., (2004). Multidimensional logit models for assessing the risk of bankruptcy of Polish enterprises. In: D.Zarzecki ed. Time for Money, t. I., Szczecin: Wydawnictwo Uniwersytetu Szczecińskiego.

WĘDZKI, D., (2000). The problem of using the ratio analysis to predict the bankruptcy of Polish enterprises–Case study. Bank i Kredyt, 5, pp. 54–61.

WĘDZKI, D., (2004). Logit model of bankruptcy for the Polish economy–Conclusions from the study. In: D.Zarzecki ed. Time for Money. Corporate finance. Financing enterprises in the EU. Szczecin: Wydawnictwo Uniwersytetu Szczecińskiego.

WIERZBA, D., (2000). Early Detection of Enterprises Threatened with Bankruptcy Based on the Analysis of Financial Ratios–Theory and Empirical Research. Zeszyty Naukowe nr 9. Warszawa: Wydawnictwo Wyższej Szkoły Ekonomiczno-Informatycznej w Warszawie.

YIP, A. Y. N., (2006). Business failure prediction: a case-based reasoning approach. Review of Pacific Basin Financial Markets and Policies, 09, pp. 491–508.

ZIĘBA, M., TOMCZAK, S. K., TOMCZAK, J. M., (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems With Applications, 58, pp. 93–101.

ZMIJEWSKI, M. E., (1984). Methodological issues related to the estimation of financial distress prediction models. Journal of Accounting Research, 22, pp. 59–82.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 197–206, DOI 10.21307/stattrans-2021-011
Received – 02.07.2020; accepted – 14.09.2020

Estimation of the number of residents included in a population frame

Danutė Krapavickaitė[1]

ABSTRACT

A high migration rate of a population causes the number of a country's residents to become extremely volatile, which negatively affects the quality of population frames. The aim of the paper is to present a method for estimating the number of residents included in a study population frame. The method involves the cross-classification of a population register with other databases which contain information relating to the activities of the population elements in a given country. The estimates from an ongoing sample survey are applied to some of the cells.

Key words: population register, sign of life, sample survey, logistic regression model.

1. Introduction

High migration rates have become a prominent characteristic of everyday life. People travel to get to know new countries, other styles of life, for studies, work, economic and other reasons. Sometimes, when departing from a country, they do not know whether their departure is temporary or permanent. Moreover, migration may be declared and non-declared, the latter making the number of country residents unknown.

Lithuania has been witnessing a rather high migration rate with a shift in direction, when immigration started outweighing emigration in 2019 (Statistics Lithuania, Official Statistics Portal). Table 1 shows that the intensity of migration in Lithuania is critical as compared to the moderate net migration rates in the European Union as a whole.

Due to migration, the number of country residents becomes volatile. As a consequence, the quality of the population frames worsens. Of course, there are many other reasons reducing frame quality. The quality requirements for frames in social statistics are very important, they are presented on the Eurostat's home page

[1] Vilnius Gediminas Technical University, Lithuania. E-mail: danute.krapavickaite@vilniustech.lt. ORCID: https://orcid.org/0000-0002-2159-1167.

(Eurostat, 2019). The number of residents included in the frame of the study population needs to be estimated. If one data source is insufficient, the integration of information from several data sources is needed. The construction of the residency index (Tiit, Maasing, 2016; Lehto, Söstra, Tiit, 2018) is one of the methods used to estimate the number of the study population residents or country residents included in the frame. Such methods use data from multiple sources, which makes quality evaluation in multisource statistics closely interlinked with and as important as the estimation of the parameters themselves.

Table 1. Crude rates of net migration plus statistical adjustment per 1,000 persons

Year	Crude rates of emigration from Lithuania	Crude rates of immigration to Lithuania	Crude rates of net migration, Lithuania	Crude rates of net migration, EU-28
2019	10.5	14.3	3.8	3.2(*)
2018	11.5	10.3	-1.2	2.8
2017	16.9	7.2	-9.7	2.3
2016	17.5	7.0	-10.5	2.3
2015	15.3	7.6	-7.7	3.5
2014	12.5	8.3	-4.2	2.1
2013	13.1	7.4	-5.7	3.5

(*) provisional data

Note. The crude rate of net migration is equal to the difference between the crude rate of increase and the crude rate of natural increase (that is, net migration is considered as the part of population change not attributable to births and deaths). The value is expressed per 1,000 inhabitants (European Commission, EMN Glossary Search).

The aim of the current paper is to present a way of estimating the number of residents included in the study population frame using the cross-classification of the population register with some other available registers and applying sample survey data. The idea is to divide the main study population frame into a union of nonintersecting cells and to estimate the number of residents included in each of those cells separately. For this purpose, various statistical methods may be used: estimates from the ongoing sample surveys, statistical models, etc.

The total number of country residents should be no less than the number of country residents included in the frame because the population may also include residents not included in any of the registers used. The estimation of the residing population size is more complicated. It is studied in the paper by Heijden, Cruyff, (2020). A dual estimation system based on the cross-classification of two or more databases and the application of the log-linear model to the cell sizes is used in this paper.

This study is partially based on the results included in the Komuso project initiated by the European Commission, 2016-2019. The results of the project on the quality for multisource statistics are shortly described in Ascari et al., 2020.

2. Presentation of a method for the estimation of the number of residents included in the study population frame

In order to demonstrate the method, data of Statistics Lithuania are used for the case study. Specifically, the problem is to classify the study population frame into country residents and non-residents and to simultaneously estimate the number of frame errors. The study population frame is based on the domain of the Population Register, which contains data on the individuals having a personal identification (ID) code of the Republic of Lithuania (LT) and older than 16 on October 1, 2016.

The sign of life is an indicator demonstrating the activity of the element in the database. In order to find signs of life of its elements, the frame has been merged with the database of the Lithuanian Labour Exchange (LE) and the database of the State Social Insurance Fund Board (SI) of Lithuania as of October 1, 2016.

2.1. Description of the data

The study population frame is further classified by the following variables:

- The *identification type* of an individual: a valid ID document, an invalid Lithuanian ID document or no document at all.
- The *address type*: a person declared his/her official address at the municipality of Lithuania; a person declared his/her home address in Lithuania; a person declared a foreign address; and other cases. "Other" means frame errors: a house at the address declared was demolished, is a non-residential building, or other illogical cases.
- *Signs of life.* The construction of the third classification variable is based on the cross-classified groups of individuals belonging to the LE and SI databases with the following levels: SI only; LE only; neither SI nor LE; SI and LE.
- Based on these three variables, the study population frame is divided into 32 non-intersecting groups, as shown in Table 2. The study goes further – to identify the elements of each cell: whether they live in Lithuania and are country residents or whether they are non-residents.

It is assumed that the fact of belonging to the LE or SI database means that an individual actually lives in Lithuania. There is possibility of a person living in one country while working in another; however, such individuals are excluded from the

current study due to a relatively insignificant number of such cases and for the sake of simplicity.

Table 2. Classification of the population frame by three variables

Document type		Address type			
	LE/SI	Municipality	Declared	Foreign	Other
Valid	SI only	A1	B1	C1	D1
Valid	LE only	A2	B2	C2	D2
Valid	Neither SI nor LE	A3	B3	C3	D3
Valid	SI and LE	A4	B4	C4	D4
Invalid	SI only	U1	V1	Z1	Y1
Invalid	LE only	U2	V2	Z2	Y2
Invalid	Neither SI nor LE	U3	V3	Z3	Y3
Invalid	SI and LE	U4	V4	Z4	Y4

2.2. Estimation of the unknown cell sizes

According to our assumption about the signs of life, the elements belonging to cells A1, B1, U1 and V1 should live in Lithuania because they have a sign of life there. The elements of cells A2, B2, A4, B4, U2 and V2 should also live in Lithuania. Cells D1,…, D4 and Y1,…, Y4 mean obscure addresses and indicate frame errors. For individuals from cells C1, C2, C4, Z1, Z2 and Z4 their signs of life are in Lithuania, however, according to the frame, their addresses are foreign. No signs of life in Lithuania indicates that individuals from the cell Z3 live abroad; individuals from the cell C3 may also live abroad. Persons from cells U3 and V3 are assumed to live abroad. Cells A3 and B3 consist of the individuals who are inactive in both LE and SI databases; they may live in Lithuania or abroad. It is also possible that some of them are already dead. The number of the three latter kinds of individuals has to be estimated. Let us denote by I2 the subset of individuals from the group B3 who live abroad, and by M2 the subset of individuals from the group B3 who passed away; let I1 and M1 be the corresponding subsets of individuals from the group A3.

Let us denote the number of individuals belonging to the cells of Table 2 by the corresponding lowercase letters: m_1, i_1, m_2, i_2, signifying the sizes of M1, I1, M2, I2; letters with hats mean their estimates: $\hat{m}_1, \hat{i}_1, \hat{m}_2, \hat{i}_2$.

Based on this investigation, the contents of the cells of Table 2 are resettled to the cells of Table 3.

Table 3. Grouped cells of Table 2

Population	According to the frame		
	Live in Lithuania	Foreign address	Errors
Live in Lithuania	A1+A2+B1+B2+U1+U2+V1+V2+A3+B3+ +A4+B4+U4+V4-I1-I2-M1-M2	C1+C2+C4+ +Z1+Z2+Z4	D1+D2+D3+D4 +Y1+Y2+Y4
Live abroad	I1+I2+U3+V3	C3+Z3	Y3
Deceased	M1+M2		

The union of cells B1, B2, B3 and B4 (Table 2) usually serves as a basis for the sampling frame in social surveys at Statistics Lithuania. Using information on the reasons for non-response from the Lithuanian Health Interview Survey, 2014, the number of individuals i_2 of I2$\subset$B1+B2+B3+B4 living abroad is estimated using sample design weights:

$$\hat{i}_2 = \hat{t}_y = \sum_{k\in(B1+B2+B3+B4)\cap s} w_k y_k \ ,$$

here *s* means a sample drawn from B1+B2+B3+B4, and w_k are the sample design weights. The values for a binary variable y are $y_k = 1$ if the *k*-th individual has not responded to the survey because he/she lives abroad, while $y_k = 0$ otherwise. Assuming that individuals from B1, B2 and B4 participate in the survey diligently, non-response is due to B3. It follows that the set of individuals living abroad I2$\subset$B3, and $y_k = 1$ for $k \in$ B3. The number m_2 of people in M2 who passed away from the cell B3 may be estimated based on the reasons of non-response using the same ongoing survey exactly in the same way.

For people belonging to A3 it is not possible to identify signs of life in Lithuania. Some additional databases are needed, to which no access was available in the current study. Assuming that the distribution of individuals living abroad in B3 coincides with the distribution of those in A3, it is possible to estimate the number of individuals included in the frame who live abroad and who passed away in A3 using the logistic regression model created in B3$\cap s$. Using data from the same ongoing survey and B3, the logistic regression model for the same variable *y* is constructed:

$$P\left(y_k = 1 \middle| x_k\right) = \frac{e^{a+bx_k}}{1+e^{a+bx_k}}, \quad k \in \mathrm{B3}\cap s,$$

and estimated. Using an assumption that individuals from B1, B2 and B4 live in Lithuania due to their signs of life, we restrict ourselves with $k \in \mathrm{B3}$. In general, *x* should be a vector of auxiliary variables, but the data study shows that there is only

one variable available in the sample data and in A3 correlated with *y*: *age*. Then, the probability of living abroad for units from A3 is estimated:

$$\hat{p}_k = \frac{e^{\hat{a}+\hat{b}x_k}}{1+e^{\hat{a}+\hat{b}x_k}}, \quad k \in \mathrm{A3}.$$

By summing the estimated probabilities over A3, the estimator for i_1 is obtained: $\hat{i}_1 = \sum_{k\in \mathrm{A3}} \hat{p}_k$. It estimates the number of individuals living abroad from A3. In a similar way, $\hat{m}_1$ (the number of individuals m_1 who passed away from A3) is estimated.

3. Numerical results

For this study, data of Statistics Lithuania for 2016 are used (European Commission, 2016–2019). The ID code is available in the frame and other databases. It allows us to merge the lists, fill in Table 2 and get the following Table 4.

Using sample data from the Lithuanian Health Interview Survey, 2014 (7000 individuals, of whom 25 passed away, and 71 went abroad), the estimates of coefficients for the logistic regression model constructed for individuals to live abroad are $\hat{a} = -2.9743$ (s.e.=0.2893), $\hat{b} = -0.0387$ (s.e.=0.0074). The estimated coefficients of the logistic regression model constructed for individuals to pass away are $\hat{a} = -9.9250$ (s.e.=0.9663), $\hat{b} = 0.0706$ (s.e.=0.0132). An assumption is made that the model did not change from 2014 to 2016, and these coefficients are applied to the data for 2016. In order to estimate the quality of the logistic regression model, the known number of individuals in the sample who went abroad t_{abroad} is compared to its estimate $\hat{t}_{abroad} = \sum_{k\in \mathrm{B3}\cap s} \hat{p}_k$, obtained using a logistic regression model, by the relative absolute ratio: $r_{abroad} = \left| t_{abroad} - \hat{t}_{abroad} \right| / t_{abroad}$.

The sample data are randomly divided into two parts: logistic regression coefficients are estimated from one of the parts and applied to estimate the number of individuals living abroad included in the second part. The relative absolute ratio is calculated in the second sample part. The procedure is repeated independently 200 times. The average for the relative absolute ratio obtained for the estimate of the number of individuals who went abroad is equal to $\bar{r}_{abroad} = 0.199$. The average relative absolute ratio $\bar{r}_{away} = 0.390$ for the estimate of the number of individuals who passed away is obtained in the same way. Estimates from the sample survey $\hat{i}_2$ and $\hat{m}_2$ have estimated coefficients of variation $c\hat{v}\left(\hat{i}_2\right) = 0.12$, $c\hat{v}\left(\hat{m}_2\right) = 0.2$. The accuracy of

the estimates is not high because the proportions of the sampled elements who went abroad and who passed away are low.

Table 4. Frame coverage and domain classification

Document type	LF/SI	Address type			
		Municipality	Declared	Foreign	Other
Valid	SI only	21 266	1 146 209	2 372	57
Valid	LF only	7 974	115 077	1 146	19
Valid	Neither SI nor LE	38 547	1267 490	257 544	133
Valid	SI &LE	1 265	22 323	33	1
Invalid	SI only	445	13 079	139	5
Invalid	LF only	189	1 568	10	0
Invalid	Neither SI nor LE	2 282	26 155	40 146	121
Invalid	SI &LE	17	313	1	0

Table 3 is presented in the same numerical way as Table 5. It provides the estimates for the number of residents included in the study population frame and the number of non-residents belonging to this frame. The number of individuals correctly classified in the frame as residents and non-residents and the number of mis-classified study population frame elements can be derived from Table 5.

Table 5. Basis for measuring the number of residents included in the study population frame, frame coverage and domain classification

Population domain	Frame domain		
	Lithuanian address	Foreign address	Frame errors
Living in Lithuania	2 600 857	3 701	215
Living abroad	54 252	297 690	121
Deceased	9 090		

Table 6 accumulates the main results of the study and provides the frame coverage. It is derived from Table 5.

Table 5 shows the cross-classification of the frame information on the country of residence and adjusted information on the residence place with the errors on the margins. Summing over its lines, both diagonals and margins, the contents of Table 6 are obtained.

Table 6. Number of residents included in the study population frame and frame coverage measures

Indicator	Value	Frame proportion
Number of country residents in the frame	2 604 558	0.8809
Number of country non-residents in the frame	351 942	0.1190
Correct domain classification	2 898 547	0.9803
Domain miss-classification	57 953	0.0196
Frame under-coverage	336	0.0001
Frame over-coverage	9 090	0.0031
In-scope frame size	2 956 500	0.9999
Overall population frame size	2 956 836	1.0000

One more aspect in the estimation of the number of residents included in the study population frame is the accuracy of the estimate. It depends on the quality of the frame and on all the statistical methods included in the estimation.

4. Discussion

In our investigation, *study population* refers as a set of individuals who are currently alive and show their activities (signs of life) in various databases. Based on these activities, one can decide if an element really exists at a certain moment and if his/her activities show that he/she should belong to a certain study population domain (with higher or lower probability).

Frame is a relatively frozen list of population elements, based on a certain register or several databases. The element is included in this list formally, and errors may occur.

We aim to merge what we know about the population elements from the sources different than the frame with what is available in the frame.

Access to additional databases, such records on health services, pensions and allowances, driving licence renewal, etc., could provide more information on signs of life for the elements of the frame; therefore cells C3, U3 and V3 could be classified or estimated more accurately. Unfortunately, although such data exist, they were not accessible for this study.

The number of signs of life used in this study is very small: only two. It is expected that if more signs of life were used, a more accurate estimate of the number of residents included in the study population frame could be obtained.

The problem studied in the paper by Tiit and Maasing (2016) is the construction of the residency index for the Estonian population. It is a complex index based on data

from 27 databases and the same index for the previous year with the values between 0 and 1 calculated for every element of the "extended total population" of the country. It is calculated for every year and used to estimate the population size. With a proper threshold, this index can be used to classify the "extended total population" into the groups of elements belonging/not belonging to the study population. It may also show the over-coverage of any frame domain. In Estonian case, the residency index allows estimating the size of the resident population and population register over-coverage due to persons who have left the country without declaration. The problem solved and the method used in the said paper by Tiit and Maasing are quite close to our solution, although more complex and using more signs of life.

Based on the idea of signs of life it seems that for any frame for which over-coverage is expected, a certain classification method may be applied based on the data showing activities of the population elements. If it is successful, the estimation of the frame over-coverage will not cause any problems. Any methods for the estimation of the active population frame size can be used for calibration in sample surveys, register-based statistics and administrative data-based population censuses.

Acknowledgments

The author is thankful to the anonymous reviewers for their comments and suggestions regarding improvements of the current paper.

References

ASCARI, G., BLIX, K., BRANCATO, G., BURG, T., McCOURT, A., VAN DELDEN, A., KRAPAVICKAITĖ, D., PLOUG, N., SCHOLTUS, S., STOLZE, P., DE WAAL, T., ZHANG, L.-C., (2020). Quality of Multisource Statistics – the KOMUSO Project. *The Survey Statistician*, January No 81, pp. 35–49, http://isi-iass.org/home/services/the-survey-statistician/ (accessed January 2020).

EUROPEAN COMMISSION, (2016-2019). *ESSnet on Quality of Multisource Statistics – Komuso*, https://ec.europa.eu/eurostat/cros/content/essnet-quality-multisource-statistics-komuso_en/ (accessed January 2020).

EUROPEAN COMMISSION, (2020). *EMN Glossary Search*, https://ec.europa.eu/home-affairs/what-we-do/networks/european_migration_network/glossary_search/ (accessed January 2020).

EUROSTAT, (2019). *Quality Guidelines for Frames in Social Statistics – QGFSS*, https://ec.europa.eu/eurostat/cros/system/files/qgfss-v1.51.pdf/ (accessed August 2020).

LEHTO, K., SŐSTRA, K., TIIT, E.-M., (2018). Index-based Methodology in Demographic Statistics. *The Survey Statistician*, January No 77, p. 45. https://isi-iass.org/home/services/the-survey-statistician/ (accessed January 2020).

STATISTICS LITHUANIA, (2020). Official Statistics Portal, https://www.stat.gov.lt/home/ (accessed August 2020).

TIIT, E.-M., MAASING, E., (2016). Residency index and its applications in censuses and population statistics. *Eesti statistika kvartalikri.* (*Quarterly Bulletin of Statistics Estonia*). No 3, pp. 41–60, http://www.stat.ee/publication-2016_quarterly-bulletin-of-statistics-estonia-3-16 (accessed January 2020).

VAN DER HEIJDEN, P. G. M., CRUYFF, M., (2020). Wider applications for dual and multiple system estimation. *The Survey Statistician*, January No 81, pp. 16–20, http://isi-iass.org/home/services/the-survey-statistician/ (accessed January 2020).

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 207–216, DOI 10.21307/stattrans-2021-012
Received – 05.05.2020; accepted – 27.11.2020

A new family of robust regression estimators utilizing robust regression tools and supplementary attributes

Irsa Sajjad[1], Muhammad Hanif[2], Nursel Koyuncu[3], Usman Shahzad[2,4], Nadia H. Al-Noor[5]

ABSTRACT

Zaman and Bulut (2018a) developed a class of estimators for a population mean utilising LMS robust regression and supplementary attributes. In this paper, a family of estimators is proposed, based on the adaptation of the estimators presented by Zaman (2019), followed by the introduction of a new family of regression-type estimators utilising robust regression tools (LAD, H-M, LMS, H-MM, Hampel-M, Tukey-M, LTS) and supplementary attributes. The mean square error expressions of the adapted and proposed families are determined through a general formula. The study demonstrates that the adapted class of the Zaman (2019) estimators is in every case more proficient than that of Zaman and Bulut (2018a). In addition, the proposed robust regression estimators based on robust regression tools and supplementary attributes are more efficient than those of Zaman and Bulut (2018a) and Zaman (2019).The theoretical findings are supported by real-life examples.

Key words: supplementary attributes, ratio-type estimators, SRS, robust regression tools, percentage relative efficiency..

1. Introduction

The estimation theory is important in different interdisciplinary territories of research including financial matters, clinical preliminaries, population studies, agriculture, engineering, and so on. Additionally, the issue of estimation of mean is critical in research, for example, the estimation of average crop yield, normal life

[1] Department of Lahore Business School - University of Lahore, Islamabad, Pakistan. E-mail: irsasajjad@yahoo.com. ORCID: https://orcid.org/0000-0002-7246-7043.

[2] Department of Mathematics and Statistics - PMAS-Arid Agriculture University, Rawalpindi, Pakistan. E-mail: mhpuno@hotmail.com. ORCID: https://orcid.org/0000-0002-1976-4452.

[3] Hacettepe University, Department of Statistics, Beytepe, Ankara, Turkey. E-mail: nkoyuncu@hacettepe.edu.tr. ORCID: https://orcid.org/0000-0003-1065-3411.

[4] Department of Mathematics and Statistics, International Islamic University, Islamabd, Pakistan. Corresponding E-mail: usman.stat@yahoo.com. ORCID: https://orcid.org/0000-0002-0178-5298.

[5] Department of Mathematics, College of Science, Mustansiriyah University, Baghdad, Iraq. E-mail: nadialnoor@uomustansiriyah.edu.iq. ORCID: https://orcid.org/0000-0002-4433-9044.

expectancy of people in an area and many others. To improve the efficiency of estimation of parameters, an auxiliary variable is widely used in the literature. An alternate way to enhance the efficiency of an estimator is to utilize supplementary attributes (for more details, interested readers may refer to Naik and Gupta (1996), Shahzad (2016), and Shahzad et al. (2018)). Ratio and product estimation techniques are widely used under SRS (simple random sampling) scheme. Both of these schemes have their own advantages and disadvantages. For instance, a ratio estimator is suitable for a positive linear relationship between the study and supplementary attribute, a product estimator is suitable for a negative linear relationship between the study and supplementary attribute. The usual regression estimator solves this issue and provides much better results for both positive and negative correlations. Note that the usual regression estimator based on ordinary least square (OLS) regression coefficient. However, when data are contaminated with outliers, OLS will not perform well, and as a result we get poor results. For solving this issue, Zaman and Bulut (2018a) utilized one of the robust regression technique namely LMS (Least Median of Squares), and provide a set of estimators utilizing auxiliary attribute under SRS scheme. In the current article, a class of robust ratio estimators is constructed by adapting the estimators of Zaman (2019), and a new class of robust regression estimators is introduced utilizing the supplementary attributes and robust-regression tools (least absolute deviations (LAD), Huber's M-estimator (H-M), least median of squares (LMS), least trimmed squares (LTS), Huber's MM-estimator (H-MM), Hampel-M estimator, Tukey-M estimator) under simple random sampling.

Zaman and Bulut (2018a) developed a class of estimators utilizing the known parameters of a supplementary attribute and LMS regression coefficient, as given below:

$$\bar{y}_{zb_1} = \frac{\bar{y}+b_{\varphi(lms)}(P_1-p_1)}{p_1} P_1 , \tag{1}$$

$$\bar{y}_{zb_2} = \frac{\bar{y}+b_{\varphi(lms)}(P_1-p_1)}{\left(P_1+C_\varphi(\varphi)\right)} \left(P_1 + C_\varphi\right), \tag{2}$$

$$\bar{y}_{zb_3} = \frac{\bar{y}+b_{\varphi(lms)}(P_1-p_1)}{\left(P_1+\beta_2(\varphi)\right)} \left(P_1 + \beta_2(\varphi)\right), \tag{3}$$

$$\bar{y}_{zb_4} = \frac{\bar{y}+b_{\varphi(lms)}(P_1-p_1)}{\left(P_1C_\varphi+\beta_2(\varphi)\right)} \left(P_1C_\varphi + \beta_2(\varphi)\right), \tag{4}$$

$$\bar{y}_{zb_5} = \frac{\bar{y}+b_{\varphi(lms)}(P_1-p_1)}{\left(P_1\beta_2(\varphi)+C_\varphi\right)} \left(P_1\beta_2(\varphi) + C_\varphi\right), \tag{5}$$

where C_φ, the coefficient of variation, $\bar{y}$, the sample mean, $\beta_2(\varphi)$, the coefficient of kurtosis, and $b_{\varphi(lms)}$ is the LMS robust regression coefficient. Further, P_1 and p_1 represent population and sample proportions, respectively. For more details about proportions, interested readers may refer to Zaman and Bulut (2018a).

The MSE of Zaman and Bulut (2018a) family of estimators is given below:

$$MSE\left(\bar{y}_{zb_i}\right) = \gamma\left[S_y^2 + g_i^2 S_\varphi^2 + 2B_i g_i S_\varphi^2 + B_i^2 S_\varphi^2 - 2g_i S_{y\varphi} - 2B_i S_{y\varphi}\right]; i = 1, \ldots, 5 \tag{6}$$

where

$g_1 = 1,\ g_2 = \frac{\bar{Y}}{P_1 + C_\varphi}, g_3 = \frac{\bar{Y}}{P_1 + \beta_2(\varphi)}, g_4 = \frac{C_\varphi \bar{Y}}{C_\varphi\ P_1 + \beta_2(\varphi)}, g_5 = \frac{\beta_2(\varphi)\bar{Y}}{C_\varphi\ P_1 + C_\varphi}, S_{y\varphi} = \rho S_y S_\varphi$ and $= \left(\frac{1}{n} - \frac{1}{N}\right)$. Further, S_y^2 and S_φ^2 are the unbiased variances of Y and P_1 respectively, ρ is the coefficient of correlation.

2. Adapted Family of Estimators

Zaman (2019) developed a class of estimators utilizing known characteristics of supplementary information. By analogy to the approach of Zaman (2019), a supplementary attribute is utilized here, as given below:

$$\bar{y}_{z_1} = k\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{p_1}P_1 + (1-k)\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{\left(P_1 + C_\varphi(\varphi)\right)}\left(P_1 + C_\varphi\right), \tag{7}$$

$$\bar{y}_{z_2} = k\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{p_1}P_1 + (1-k)\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{\left(P_1 + \beta_2(\varphi)\right)}\left(P_1 + \beta_2(\varphi)\right), \tag{8}$$

$$\bar{y}_{z_3} = k\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{p_1}P_1 + (1-k)\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{\left(P_1 C_\varphi + \beta_2(\varphi)\right)}\left(P_1 C_\varphi + \beta_2(\varphi)\right), \tag{9}$$

$$\bar{y}_{z_4} = k\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{p_1}P_1 + (1-k)\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{\left(P_1\beta_2(\varphi) + C_\varphi\right)}\left(P_1\beta_2(\varphi) + C_\varphi\right), \tag{10}$$

In general form, we can write the adapted class of estimators as given below:

$$\bar{y}_{z_i} = k\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{p_1}P_1 + (1-k)\frac{\bar{y} + b_{\varphi(i)}(P_1 - p_1)}{\left(P_1 U_\varphi + V_\varphi\right)}\left(P_1 U_\varphi + V_\varphi\right), \tag{11}$$

where U_φ and V_φ are the known characteristics of an auxiliary attribute. The MSE of $\mathrm{MSE}\left(\bar{y}_{z_i}\right)$ is as follows:

$$MSE\left(\bar{y}_{z_i}\right) = \gamma\left[S_y^2 - 2\delta S_{y\varphi} + \delta^2 S_\varphi^2\right], \tag{12}$$

where $\delta = \left[k\left(B_{\varphi(i)} + g_1\right) + (1-k)\left(B_{\varphi(i)} + g_i\right)\right]$.

By replacing $(\delta = B)$ in the above MSE expression, we get the minimum MSE of $\bar{y}_{z_i}$ as follows:

$$MSE\left(\bar{y}_{z_i}\right) = \gamma S_y^2(1-\rho^2), \tag{13}$$

which is the MSE of traditional regression estimator, i.e. $\bar{y}_{reg} = \bar{y} + b_{\varphi(i)}(P_1 - p_1)$. Note that Abd-Elfattah et al. (2010) consider same class in the same context utilizing OLS regression coefficient. However, Zaman (2019) introduced robust regression techniques instead of OLS regression in the presence of outliers.

3. Proposed Estimators

Taking motivation from ratio type estimators of Zaman (2019), we propose the following family of robust regression estimators as given by

$$\hat{\bar{y}}_{p_1} = w_1\left[\bar{y} + b_{\varphi(mm)}(P_1 - p_1)\right] + w_2\left[\bar{y} + b_{\varphi(lms)}(P_1 - p_1)\right], \tag{14}$$

$$\hat{\bar{y}}_{p_2} = w_1\left[\bar{y} + b_{\varphi(lad)}(P_1 - p_1)\right] + w_2\left[\bar{y} + b_{\varphi(tuket)}(P_1 - p_1)\right], \tag{15}$$

$$\hat{\bar{y}}_{p_3} = w_1\left[\bar{y} + b_{\varphi(lts)}(P_1 - p_1)\right] + w_2\left[\bar{y} + b_{\varphi(huber)}(P_1 - p_1)\right], \tag{16}$$

$$\hat{\bar{y}}_{p_4} = w_1\left[\bar{y} + b_{\varphi(mm)}(P_1 - p_1)\right] + w_2\left[\bar{y} + b_{\varphi(hample)}(P_1 - p_1)\right], \tag{17}$$

$$\hat{\bar{y}}_{p_5} = w_1\left[\bar{y} + b_{\varphi(huber)}(P_1 - p_1)\right] + w_2\left[\bar{y} + b_{\varphi(mm)}(P_1 - p_1)\right], \tag{18}$$

$$\hat{\bar{y}}_{p_6} = w_1\left[\bar{y} + b_{\varphi(huber)}(P_1 - p_1)\right] + w_2\left[\bar{y} + b_{\varphi(lms)}(P_1 - p_1)\right], \tag{19}$$

In general form, we can write the proposed family of estimators as

$$\hat{\bar{y}}_{p_i} = w_1\left[\bar{y} + b_{\varphi(i)}(P_1 - p_1)\right] + w_2\left[\bar{y} + b_{\varphi(j\neq i)}(P_1 - p_1)\right]; (i, j \neq i) = 1, \ldots, 6 \tag{20}$$

However, it is interesting to note that if we put $(w_1, w_2) = (0, 1)$, $\hat{\bar{y}}_{p_i}$ will be converted into a traditional robust regression type estimator introduced by Nasir et al. (2018) under SRS for quantitative sensitive study variable. Hence, these estimator is a special case of the proposed class. The proposed family relies on the robust-regression tools, i.e. $b_{\varphi(i)}$ (LAD, H-M, LMS, LTS, H-MM, Hampel-M, Tukey-M) for $(i = 1, \ldots, 6)$ respectively. For deep knowledge of $b_{\varphi(i)}$, interested readers may refer to Zaman and Bulut (2018b).

To obtain MSE, let us define $\bar{y} = \left(1 + \eta_y\right)\overline{Y}$, $p_1 = \left(1 + \eta_\varphi\right)P_1$. Utilizing these notations $\eta_i (i = y, \varphi)$, we can write
$E\left(\eta_y\right) = E\left(\eta_\varphi\right) = 0, E\left(\eta_y^2\right) = \gamma C_y^2, E\left(\eta_\varphi^2\right) = \gamma C_\varphi^2$ and $\left(\eta_y \eta_\varphi\right) = \gamma C_{y\varphi}$.

Now, expanding $\hat{\bar{y}}_{p_i}$ in terms of η_y and η_φ as

$$\hat{\bar{y}}_{p_i} = \left[w_1 \bar{Y}\left(1 + \eta_y\right) - b_{\varphi(i)} P_1 \eta_\varphi\right] + \left[w_2 \bar{Y}\left(1 + \eta_y\right) - b_{\varphi(j \neq i)} P_1 \eta_\varphi\right]. \quad (21)$$

Squaring (21), applying expectation, we get theoretical MSE of the estimator $\hat{\bar{y}}_{p_i}$ up to the order n^{-1}, as

$$MSE\left(\hat{\bar{y}}_{p_i}\right) = \bar{Y}^2 + w_1^2 \delta_A + w_2^2 \delta_B + 2 w_1 w_2 \delta_C - 2 w_1 \delta_D - 2 w_2 \delta_E , \quad (22)$$

where

$$\delta_A = \left[\bar{Y}^2 + \gamma\left\{S_y^2 + B_{\varphi(i)}\left(B_{\varphi(i)} S_\varphi - 2\rho S_y\right) S_\varphi\right\}\right],$$

$$\delta_B = \left[\bar{Y}^2 + \gamma\left\{S_y^2 + B_{\varphi(j \neq i)}\left(B_{\varphi(j \neq i)} S_\varphi - 2\rho S_y\right) S_\varphi\right\}\right],$$

$$\delta_C = \left[\bar{Y}^2 + \gamma\left\{S_y^2 - \left(B_{\varphi(i)} + B_{\varphi(j \neq i)}\right) S_{y\varphi} + B_{\varphi(i)} B_{\varphi(j \neq i)} S_\varphi^2\right\}\right],$$

$$\delta_D = \delta_E = \bar{Y}^2.$$

By partially differentiating (22) w.r.t. w_1 and w_2, we obtained the optimum values as given by

$$w_1^{opt} = \left[\frac{\delta_B \delta_D - \delta_C \delta_E}{\delta_A \delta_B - \delta_C^2}\right],$$

and

$$w_2^{opt} = \left[\frac{\delta_A \delta_E - \delta_C \delta_D}{\delta_A \delta_B - \delta_C^2}\right],$$

Substitution of w_1^{opt} and w_2^{opt} in (22) provides the minimum MSE of $\hat{\bar{y}}_{p_i}$ as

$$MES_{min}\left(\hat{\bar{y}}_{p_i}\right) = \left[\bar{Y}^2 - \frac{\delta_B \delta_D^2 - 2\delta_C \delta_D \delta_E + \delta_A \delta_E^2}{\delta_A \delta_B - \delta_C^2}\right]. \quad (23)$$

The general theoretical condition of proposed vs. existing estimators as given below:

$$MSE(\bar{y}_{zi}) - MSE\left(\hat{\bar{y}}_{pi}\right) > 0$$

4. Numerical Illustration

A numerical illustration is performed utilizing the previous studies of Koyuncu (2012).

Data 1 [Source: Sukhatme and Sukhatme (1970)]

$\emptyset$ = A circle consisting of more than five villages.

Y = Number of villages in the circles.

$B_{lad} = 4, B_{lms} = 5, B_{huber} = 4.660824, B_{hample} = 4.672494,$
$B_{lts} = 5, B_{tukey} = 4.655754, B_{mm} = 4.647839, \text{var}(\bar{y}) = 4.0738$.

Data 2 [Source: Sukhatme and Sukhatme (1970)]

$\emptyset$ = A circle consisting of more than five villages.

Y = Area under the wheat crop within the circles.

$B_{lad} = 1678.281, B_{lms} = 1896, B_{huber} = 1462.839, B_{hample} = 1438.403,$
$B_{lts} = 1896, B_{tukey} = 1574.684, B_{mm} = 1573.993, \text{var}(\bar{y}) = 513592$.

For remaining characteristics of the data sets, interested readers may refer to Koyuncu (2012). The data sets are also available in the appendix.

Figures 1- 4 clearly show that our considered data sets suffer from non-normality and the presence of outliers. Hence, suitable for robust regression tools. In Table 1, results of PRE, which are figured utilizing PRE equations displayed in Sections 1, 2, and 3, are provided. Note that by ignoring fractional values in the proposed class, all members of the proposed class are providing the same results. So, we ignore the fractional part and provide a single value of PRE in Table 1. When we look at Table 1, we see that the proposed class has the maximum PRE among all estimators given in Sections 1 and 2. From the consequence of this numerical delineation, it is unmistakably concluded that all new estimators are more effective than existing and adapting estimators.

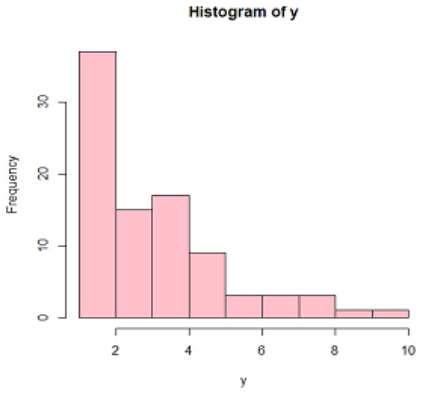


Figure 1. Population-1

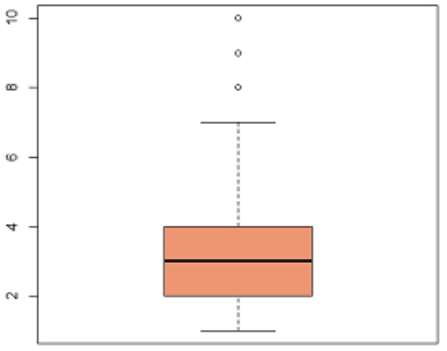


Figure 2. Boxplot Population-1

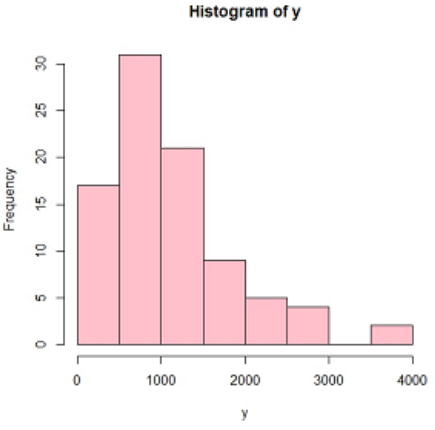


Figure 3. Population-2

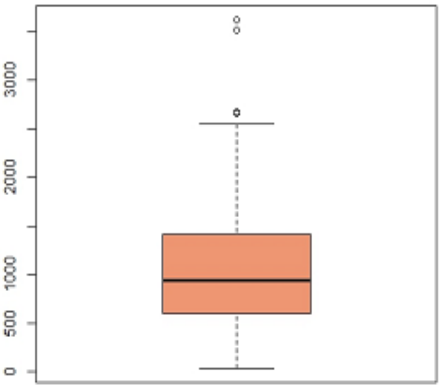


Figure 4. Boxplot Population-2

Table 1. The MSE and PRE of Proposed and Existing Estimators w.r.t. var($\bar{y}$)

Estimator	Pop-1		Pop-2	
	MSE	PRE	MSE	PRE
$\bar{y}_{zb_1}$	2.72724	149.37	326355.7	157.37
$\bar{y}_{zb_2}$	0.06246	6521.38	13236.41	3880.14
$\bar{y}_{zb_3}$	0.05682	7169.33	11947.89	4298.59
$\bar{y}_{zb_4}$	0.06021	6764.89	12766.2	4023.06
$\bar{y}_{zb_5}$	0.06442	6323.61	13621.23	3770.52
$\bar{y}_{reg}$	0.05423	7511.55	10120.65	5074.69
$\bar{y}_p$	0.05397	7547	10037	5116

5. Conclusion

This paper proposes two classes of estimators. It was discovered that the proposed robust regression estimators were more efficient than the estimators of Zaman and Bulut (2018a) and Zaman (2019). The outcomes displayed here support this conclusion by hypothetical improvement and numerical examination.

References

ABD-ELFATTA,H 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, Appl. Math. Comput., Vol. 215, pp. 4198–4202.

KOYUNCU, N., (2012). Efficient estimators of population mean using auxiliary attributes. Applied Mathematics and Computation, Vol. 218, pp. 10900–10905.

NAIK, V. D., GUPTA, P. C., (1996). A note on estimation of mean with known population of an auxiliary character, J. Indian Soc. Agr. Stat., Vol. 48, pp. 151–158.

NASIR, A., AHMAD, I., HANIF, M., SHAHZAD, U., (2018). Robust-regression-type estimators for improving mean estimation of sensitive variables by using auxiliary information, Commun. Stat. Theory Methods, doi:10.1080/03610926.2019.1645857.

SHAHZAD, U., (2016). On the Estimation Of Population Mean Under Systematic Sampling Using Auxiliary Attributes, Oriental Journal Physical Sciences, Vol. 1 (1&2), pp. 17–22.

SHAHZAD, U., HANIF, M., KOYUNCU, N., SANAULLAH, A., (2018). On the estimation of population variance using auxiliary attribute in absence and presence of non-response, Electronic Journal of Applied Statistical Analysis, Vol. 11, pp. 608–621.

SUKHATME, P. V., SUKHATME, B. V., (1970). Sampling Theory of Surveys with Applications, Iowa State University Press, Ames, USA,.

ZAMAN, T., BULUT, H., (2018a). Modified ratio estimators for population mean using robust regression based on auxiliary attribute, IJMR, Vol. 4, pp. 1–6.

ZAMAN, T., BULUT, H., (2018b). Modified ratio estimators using robust regression methods, Commun. Stat. Theory Methods, doi:10.1080/03610926.2018.1441419.

ZAMAN, T., (2019). Improvement of modified ratio estimators using robust regression methods. Applied Mathematics and Computation, Vol. 348, pp. 627–631.

APPENDIX

Data 1

$\emptyset$ = 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0.

$\boldsymbol{Y}$ = 6, 5, 4, 5, 4, 2, 4, 2, 5, 1, 3, 4, 3, 1, 1, 3, 4, 8, 2, 4, 3, 4, 4, 3, 5, 2, 3, 1, 2, 4, 3, 2, 4, 4, 1, 7, 3, 2, 2, 5, 2, 4, 6, 1, 2, 2, 3, 1, 1, 10, 5, 9, 2, 2, 3, 8, 3, 2, 5, 2, 2, 2, 2, 3, 4, 1, 5, 4, 3, 5, 7, 3, 2, 7, 4, 1, 2, 1, 1, 8, 2, 1, 3, 2, 6, 1, 1, 4, 4.

Data 2

$\emptyset$ = 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0.

$\boldsymbol{Y}$ = 1562, 1003, 1691, 271, 458, 736, 1224, 996, 475, 34, 1027, 1393, 692, 524, 602, 1522, 2087, 2474, 461, 846, 1036, 948, 1412, 438, 2111, 977, 814, 319, 583, 1150, 670, 499, 714, 1081, 389, 2675, 868, 1412, 445, 706, 642, 2050, 2530, 247, 421, 687, 941, 710, 387, 3516, 2002, 3622, 1400, 1584, 830, 167, 622, 591, 273, 781, 1101, 799, 601, 928, 1141, 1208, 1633, 902, 1286, 1299, 1947, 741, 574, 2554, 669, 1187, 852, 51, 1265, 1423, 794, 1604, 1621, 1764, 2668, 1076, 348, 1224, 1490.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1, pp. 217–226, DOI 10.21307/stattrans-2021-013
Received – 17.07.2020; accepted – 06.11.2020

Analysing the impact of dependency on conditional survival functions using copulas

Hadi Safari-Katesari[1], **Samira Zaroudi**[2]

ABSTRACT

Nowadays, insurance contract reserves for coupled lives are considered jointly, which has a significant influence on the process of determining actuarial reserves. In this paper, conditional survival distributions of life insurance reserves are computed using copulas. Subsequently, the results are compared with an independence case. These calculations are based on selected Archimedean copulas and apply when the 'death of one individual' condition exists. The estimation outcome indicates that the insurer reserves calculated by means of Archimedean copulas are far more effective than those resulting from an independence assumption. The study demonstrates that copula-based dependency modelling improves the calculations of reserves made for actuarial purposes.

Key words: conditional survival distribution, copula, Kendall's tau, reserves, life table.

1. Introduction

There exist extensive literature on modelling insurance contracts of two lives using the independence assumption between two lives with focusing on the starting time of the contracts. The main problem for calculating reserves (provisions) between two lifetimes is finding a way to reduce costs of insurance companies. In this paper, we relax the independence assumption and show that the dependency between two lives has a great effect on the solvency of financial businesses such as insurance companies.

In the past few years, many works have been done on application of copula for modelling the dependency between two lifetimes. See, for example Hougaard (2000), Spreeuw (2006), Spreeuw and Owadally (2013), and Zaroudi et al. (2018a,b). Lee et al. (2014) studied the dependence between the policyholders in multiple-life of insurance contracts. Also, many works considered the relationship between two lifetimes of coupled lives, such as Carriere (2000) and Ji et al. (2011). This dependency has a significant impact on the determination of pricing of the insurance contracts. If one considers remaining lifetimes of coupled lives dependently at the beginning period of issuing the insurance policy, then the death chance of two lives might depend on the life status of each of them. A copula is a useful tool for building multivariate distributions when the marginal distributions are available. The copula captures the linear and nonlinear dependencies available in dataset simultaneously and is an applicable tool for modelling the dependency structures between different random

[1]Corresponding Author: School of Mathematical and Statistical Sciences, Southern Illinois University, Carbondale, IL 62901-4408, USA. E-mail: hadi.safari@siu.edu. ORCID: https://orcid.org/0000-0003-2630-3133.
[2]School of Mathematical and Statistical Sciences, Southern Illinois University, Carbondale, IL 62901-4408, USA. ORCID: https://orcid.org/0000-0001-8290-6137.

variables (r.v.'s). Moreover, it opens a new vision for considering the properties of joint distributions of marginal r.v.'s such that it separates the study of dependency effects from the effects of the margins (Mari and Kotz, 2004). Following the studies by Spreeuw and Owadally (2013), and Spreeuw (2006), in this paper, we discuss methods of calculating the insurer's reserves based on copula models using the life table of France. Moreover, we apply Kendall's tau for measuring the dependency of variables using copula models. Our method is useful by selecting an optimal copula and bringing an approach for reserves calculation of contracts. Statistical analyses were performed with R version 3.6.2: The R Project for Statistical Computing.

The framework of this paper goes as follows: In Section 2, we review copula models, family of Archimedean copulas, Kendall's tau correlation coefficient, survival functions, and the relationship between copula and survival functions based on 'death of one individual'. In Section 3, we calculate reserves for life insurance policy and three Archimedean copulas for 'death of one individual' using life table of France. In Section 4, we apply a numerical example for reserves of two lifetimes in insurance contracts. Section 5 provides conclusion remarks.

2. Copula

The term copula is a Latin word 'copuler' which means 'tie and link' and was introduced at first by Sklar Theorem (Sklar, 1959) describing the relationship between r.v.'s to build multivariate distribution functions (Nelsen, 2006). In this paper, only bivariate copulas are used for considering the dependency of two lives when one of them dies during the time period of the insurance contract. In the next subsection, the bivariate copulas are introduced and their properties are investigated. For more information on properties of copula in insurance and financial concept, see Frees and Valdez (1998), Katesari and Vajargah (2015), Katesari and Zarodi (2016), and Safari-Katesari and Zaroudi (2020).

2.1. Random Variables and Copula

Copula is a statistical tool for modelling the dependence between r.v.'s. Let us start with the definition of copula from Denuit et al. (2005) as follows:

Definition 2.1 *A bivariate copula* $C:[0,1]^2 \to [0,1]$ *is a non-decreasing and right-continuous function which preserves the following:*

1. $\lim_{u_i \to 0} C(u_1,u_2) = 0$, *for* $i = 1,2$,
2. $\lim_{u_2 \to 1} C(u_1,u_2) = u_1$ *and* $\lim_{u_1 \to 1} C(u_1,u_2) = u_2$,
3. $C(v_1,v_2) - C(u_1,v_2) - C(v_1,u_2) + C(u_1,u_2) \geq 0$,

for any $u_1,u_2,v_1,v_2 \in [0,1]$, $u_1 \leq v_1$ *and* $u_2 \leq v_2$.

By Sklar's Theorem (Sklar, 1959), one can easily consider the dependence between multivariate distribution functions and their corresponding margins with copulas. Now, we consider Sklar's Theorem, which is as follows:

Theorem 2.1 *Consider the joint distribution $H_X(x_1,x_2)$ with continuous marginal cumulative distribution functions (CDFs) F_1 and F_2. Then, a copula C is available for all $x_1,x_2 \in [-\infty,+\infty]$ as follows:*

$$H_X(x_1,x_2) = C(F_1(x_1),F_2(x_2)). \tag{1}$$

Conversely, the function H_X denoted in Eq. (1) is a bivariate distribution function when C is a copula function with marginal CDFs F_1 and F_2.

Moreover, we use the following corollary and definition from Nelsen (2006) and Denuit et al. (2005), respectively.

Corollary 2.1 *Following Theorem 2.1, for any $(u_1,u_2) \in [0,1]^2$, we have:*

$$C(u_1,u_2) = H_X(F_1^{-1}(u_1),F_2^{-1}(u_2)).$$

Definition 2.2 *For any $(u_1,u_2) \in [0,1]^2$, $C(u_1,u_2) = u_1u_2$ defined as the independent copula.*

In what follows, we will introduce one of the most famous family of copulas named Archimedean copulas.

2.2. Archimedean Copulas

The copula described by Genest and MacKay (1986a) and Genest and MacKay (1986b) called the Archimedean copulas and there exist rich applications of this copula family in the mathematical and statistical literature. Many actuarial and financial dataset have been analyzed using Archimedean copulas; for more information, see Frees and Valdez (1998), Denuit et al. (2005) and Klugman and Parsa (1999). One of the important features of this family is that they are parametric. A generating function is the general structure of Archimedean copula, which is introduced in the following definition from Denuit et al. (2005).

Definition 2.3 *Assume $\phi : [0,1] \to [0,+\infty]$ is a continuous, strictly decreasing function and has following characteristics:*

$$\phi(1) = 0, \quad \phi^{'}(\tau) < 0, \quad \phi^{''}(\tau) > 0, \quad \forall \tau \in (0,1). \tag{2}$$

Then, the bivariate copula

$$C_\phi(u_1,u_2) = \begin{cases} \phi^{-1}(\phi(u_1)+\phi(u_2)), & if \quad \phi(u_1)+\phi(u_2) \leq \phi(0), \\ 0, & otherwise. \end{cases} \tag{3}$$

is generated by $\phi(\cdot)$ defined in Eq. (2) where the copula $C_\phi(u_1,u_2)$ is the Archimedean copula for $0 \leq u_1,u_2 \leq 1$. The generator function of the copula is ϕ.

One of the most important criteria for explaining the relationship between the dependency and the Archimedean copula is expressed in the following Theorem from Nelsen (2006).

Theorem 2.2 *Consider X and Y as r.v.'s with Archimedean copula C generated by ϕ. Then, Kendall's tau (τ) correlation for X and Y is as follows:*

$$\tau = 1 + 4\int_0^1 \frac{\phi(t)}{\phi'(t)}\, dt. \tag{4}$$

In this paper, we apply three copula functions in the Archimedean family, which are defined in Table 1. In this table, the second column displays the values of dependence parameter (θ) indicating the case of independence, the third column shows the relationships between θ and τ, and the fourth and fifth columns display generation functions and domains of θ for these copulas, respectively (Nelsen, 2006).

Table 1: Three copulas in Archimedean copula family

Family	Independence	τ	$\phi(\tau)$	$\theta \in$
Family 3	$\theta = 0$	$(1/\theta) - e$	$\exp[\tau^{-\theta}] - e$	$(0,\infty)$
G-H	$\theta = 1$	$\theta - 1/\theta$	$(-\ln\tau)^{\theta}$	$[-1,\infty)$
Clayton	$\theta = 0$	$\theta/(\theta+2)$	$\tau^{-\theta} - 1$	$[-1,\infty)\backslash\{0\}$

2.3. Survival Function and Copula

A contract of life insurance defines on couple (two persons) lives with x and y ages at time 0, respectively. The remaining lifetimes of persons x and y are defined by T_x and T_y, where the marginal survival functions and the copula survival function are denoted by $S_{T_x}(s_1)$, $S_{T_y}(s_2)$ and $S_{T_x,T_y}(s_1,s_2)$, respectively. The survival function of the remaining lifetime at time t for the person aged x, given the death of person aged y at time t_y is defined as (Spreeuw, 2006):

$$\begin{aligned} S_{T_x;t}(s|T_y = t_y) &= P(T_x > t+s | T_x > t, T_y = t_y) = \frac{-\frac{\mathrm{d}}{\mathrm{d}t_y} P(T_x > t+s, T_y > t_y)}{-\frac{\mathrm{d}}{\mathrm{d}t_y} P(T_x > t, T_y > t_y)} \\ &= \frac{-\frac{\mathrm{d}}{\mathrm{d}t_y} S_{T_x,T_y}(t+s,t_y)}{-\frac{\mathrm{d}}{\mathrm{d}t_y} S_{T_x,T_y}(t,t_y)} = \frac{\left(C_2\left(S_{T_x}(t+s),\xi\right)\right)_{\xi=S_{T_y}(t_y)}}{\left(C_2\left(S_{T_x}(t),\xi\right)\right)_{\xi=S_{T_y}(t_y)}}. \end{aligned} \tag{5}$$

Notice that the function defined in Eq. (5) is a survival copula function where the partial derivative of copula C for the second argument denoted with C_2 such that $C_2\left(S_{T_x}(t),\xi\right) \neq 0$ is defined for $\xi = S_{T_y}(t_y)$. Then, following Spreeuw (2006) and using Eq. (3), we have:

$$C(S_{T_x}(t+s), S_{T_y}(t_y)) = \phi^{-1}(\phi(S_{T_x}(t+s)) + \phi(S_{T_y}(t_y))) = S_{T_x,T_y}(t+s,t_y),$$

and

$$\begin{aligned}(C_2(S_{T_x}(t+s),\xi))_{\xi=S_{T_y}(t_y)} &= \phi^{'}(S_{T_y}(t_y))(\phi^{-1})^{'}(\phi(S_{T_x,T_y}(t+s,t_y))) \\ &= \phi^{'}(S_{T_y}(t_y))(\phi^{-1})^{'}(\phi(S_{T_x}(t+s))+\phi(S_{T_y}(t_y))).\end{aligned} \tag{6}$$

3. Reserves

For calculating the insurer's reserves, we use the expected *present* value for the payment of insurance contract at the beginning of each year until the death of one individual. For a whole life annuity-due, the net single premium is defined as (Gerber, 1997):

$$\ddot{a}_x = \sum_{k=0}^{\infty} v^k {}_kp_x, \tag{7}$$

where $v = (1+i)^{-1}$ is a discount factor and i is an annual interest rate. Notice that we consider i the same for the entire years.

Definition 3.1 *(Gerber, 1997). The survival function for remaining lifetime T_x is denoted by $S_{T_x}(t)$ and defined as:*

$${}_tp_x = S_{T_x}(t) = P(T_x > t).$$

Using definition 3.1, we can rewrite Eq. (7) for 'death of one individual' status where person aged x is still alive until t and person aged y will die between t_y and $t_y + \mathrm{d}t$ as follows:

$$\ddot{a}_{x;t|t_y} = \sum_{s=0}^{\infty} v^s \, S_{T_x;t}(s|T_y = t_y). \tag{8}$$

By combining Eq. (5) and Eq. (6), we obtain:

$$S_{T_x;t}(s|T_y = t_y) = \frac{(\phi^{-1})^{'}(\phi(S_{T_x}(t+s))+\phi(S_{T_y}(t_y)))}{(\phi^{-1})^{'}(\phi(S_{T_x}(t))+\phi(S_{T_y}(t_y)))}. \tag{9}$$

Notice that Eq. (8) is the annual reserves of the whole-life annuity (defined at Gerber, 1997) for the 'death of one individual' status, which by Eq. (9) can be written as follows:

$$\ddot{a}_{x;t|t_y} = \sum_{s=0}^{\infty} v^s \, \frac{(\phi^{-1})^{'}(\phi(S_{T_x}(t+s))+\phi(S_{T_y}(t_y)))}{(\phi^{-1})^{'}(\phi(S_{T_x}(t))+\phi(S_{T_y}(t_y)))}. \tag{10}$$

Table 2 provided the reserve for some Archimedean copulas at 'death of one individual' status using Eq. (10).

Table 2: Reserve for Archimedean copulas at 'death of one individual' status

Copula	Reserves ($\ddot{a}_{x;t\|t}$)
Clayton	$\sum_{s=0}^{\infty} v^s \left(\frac{S_{T_x}(t+s)^{-\theta}+S_{T_y}(t_y)^{-\theta}-1}{S_{T_x}(s)^{-\theta}+S_{T_y}(t_y)^{-\theta}-1}\right)^{-\frac{1}{\theta}-1}$
G-H	$\sum_{s=0}^{\infty} v^s \frac{\exp\{-[(-\ln S_{T_x}(t+s))^{\theta}+(-\ln S_{T_y}(t))^{\theta}]^{\frac{1}{\theta}}\}}{\exp\{-[(-\ln S_{T_x}(t))^{\theta}+(-\ln S_{T_y}(t))^{\theta}]^{\frac{1}{\theta}}\}} \times \left(\frac{(-\ln S_{T_x}(t+s))^{\theta}+(-\ln S_{T_y}(t_y))^{\theta}}{(-\ln S_{T_x}(t))^{\theta}+(-\ln S_{T_y}(t_y))^{\theta}}\right)^{\frac{1}{\theta}-1}$
Family 3	$\sum_{s=0}^{\infty} v^s \left(\frac{\ln(\exp(S_{T_x}(t+s)^{-\theta})+\exp(S_{T_y}(t_y)^{-\theta})-e)}{\ln(\exp(S_{T_x}(t)^{-\theta})+\exp(S_{T_y}(t_y)^{-\theta})-e)}\right)^{-\frac{1}{\theta}-1} \times \frac{\exp(S_{T_x}(t)^{-\theta})+\exp(S_{T_y}(t_y)^{-\theta})-e}{\exp(S_{T_x}(t+s)^{-\theta})+\exp(S_{T_y}(t_y)^{-\theta})-e}$

4. Application

In this section, we calculate reserves using copula models for life insurance companies. The results are compared with the independence case, which demonstrates the advantage of the copula model in decreasing the amount of reserves and impact on the solvency of the insurance companies. Moreover, statistical computations were conducted with R version 3.6.2: The R Project for Statistical Computing. In order to calculate the insurer's reserves based on the copulas defined in Table 2 for the 'death of one individual' status, one needs the marginal distributions of male and female's survival. For this goal, we use the life table of France in the year 2008. The complete life table can be found at https://www.ined.fr/fr/tout-savoir-population/chiffres/france/mortalite-cause-deces/table-mortalite/.

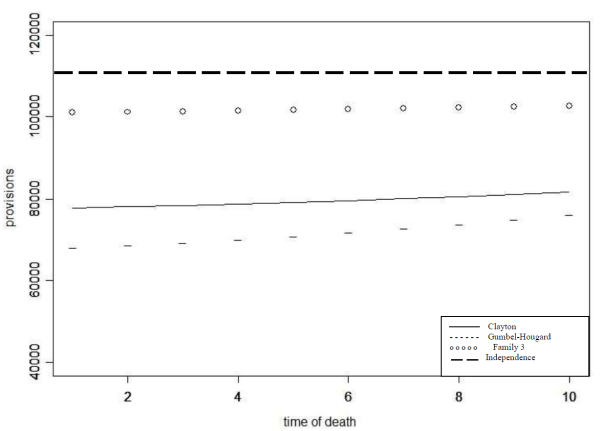


Figure 1: Reserves for time of death of person aged y at period 10.

We assume an insurance contract for a couple of 60 years old. As long as a person (male regardless of the state of the woman) is alive, the insurer liabilities are 10,000 units at the beginning of each year. Table 3 provided the percentage of difference in calculated reserves average using some copulas in Archimedean family (Clayton, Gumbel-Hougaard and Family 3) relative to the independence copula for the time period of 10, 20, and 30

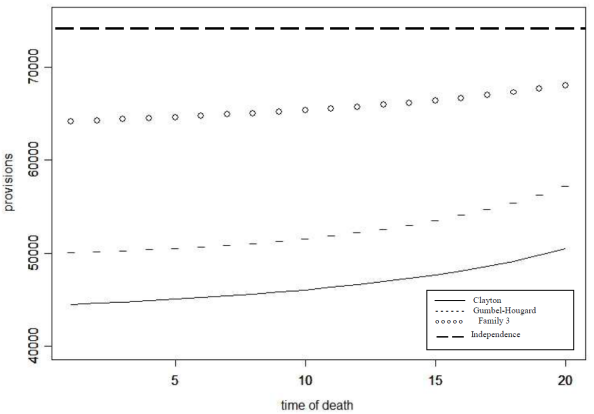


Figure 2: Reserves for time of death of person aged y at period 20.

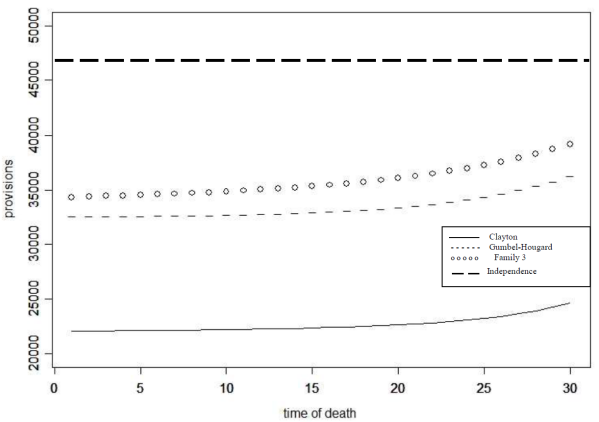


Figure 3: Reserves for time of death of person aged y at period 30.

years. Moreover, the Kendall's tau coefficient defined in Eq. (4) is 0.5 at the start point of the contract with an annual interest rate of 4%. Table 3 demonstrated that applying copula in

Table 3: The percentage of difference in calculated reserves average using Archimedean copula relative to the independence copula

Copula	period 10	period 20	period 30
Clayton	-29.2	-36	-44
Gumbel-Hougaard	-38.1	-28.7	-15.5
Family 3	-11.5	-11.3	-14

Family 3 reduced the reserve from 11% through 14% compared to the independence copula in the case that if one of the persons died at the age of y for 10, 20, and 30 years. As can be

seen from Figures 1, 2, and 3, the reserves for all three copulas are less than independence case during the time. Figure 1 shows the reserve at the time period of 10 years which the copula of Family type 3 has the highest and the Gumbel-Hougaard copula has the lowest amount of reserve. Figure 2 and Figure 3 displayed the reserves at the time period of 20 years and 30 years, respectively, which the Family 3 has the highest, and the Clayton has the lowest amount of reserves. To this end, our estimation results indicate that the insurer's reserves with using Archimedean copula families are less than the independence case, which increases the solvency of insurance companies.

5. Conclusion

Since life insurance reserves are calculated by the death of one individual, dependency of the two lifetimes plays an important role for actuarial computations of insurance companies. In this paper, the insurer's reserves are calculated using some Archimedean copulas for different time periods. The results showed that fitting the appropriate copula optimizes the amount of insurers' funds, which can be spent for reserves of future liabilities. Thus, considering dependency between two lifetimes for calculating the optimal reserves by the insurer is highly recommended.

References

CARRIERE, J. F., (2000). Bivariate survival models for coupled lives. Scandinavian Actuarial Journal, **(1)**, pp. 17–32.

DENUIT, M., DHAENE, J., GOOVAERTS, M., and KAAS, R., (2005). *Actuarial theory for dependent risks: measures, orders and models*. John Wiley and Sons.

FREES, E. W., VALDEZ, E. A., (1998). Understanding relationships using copulas. *North American actuarial journal*, 2**(1)**, pp. 1–25.

GENEST, C., MACKAY, J., (1986a). The joy of copulas: Bivariate distributions with uniform marginals. *The American Statistician*, 40**(4)**, pp. 280–283.

GENEST, C., MACKAY, R. J., (1986b). Copules archimédiennes et families de lois bidimensionnelles dont les marges sont données. *Canadian Journal of Statistics*, 14**(2)**, pp. 145–159.

GERBER, H. U., (1997). *Life insurance mathematics*. Springer Science and Business Media.

HOUGAARD, P., (2000). *Analysis of multivariate survival data*. Springer Science and Business Media.

JI, M., HARDY, M., and LI, J. S. H., (2011). Markovian approaches to joint-life mortality. *North American Actuarial Journal*, 15**(3)**, pp. 357–376.

KATESARI, H. S., VAJARGAH, B. F., (2015). Testing Adverse Selection Using Frank Copula Approach in Iran Insurance Markets. *Journal of mathematics and computer Science*. 15**(2)**, pp. 154–158.

KATESARI,H. S., ZARODI, S., (2016). Effects of Coverage Choice by Predictive Modeling on Frequency of Accidents. *Caspian Journal of Applied Sciences Research*, 5**(3)**, pp. 28–33.

KLUGMAN, S. A., PARSA, R., (1999). Fitting bivariate loss distributions with copulas. *Insurance: mathematics and economics*, 24**(2)**, pp. 139–148.

LEE, I., LEE, H. and KIM, H.T., (2014). Analysis of reserves in multiple life insurance using copula. *Communications for Statistical Applications and Methods*, 21**(1)**, pp. 23–43.

MARI, D., KOTZ, S., (2004). *Correlation and Dependent*. Imperial College Press.

NELSEN, R. B., (2006). *An Introduction to Copulas*. Springer Science and Business Media.

SAFARI-KATESARI, H., ZAROUDI, S., (2020). Count copula regression model using generalized beta distribution of the second kind. *Statistics in Transition New Series*, 21(2), pp. 1–12.

SKLAR, M., (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8, pp. 229–231.

SPREEUW, J., (2006). Types of dependence and time-dependent association between two lifetimes in single parameter copula models. *Scandinavian Actuarial Journal*, **5**, pp. 286–309.

SPREEUW, J., OWADALLY, I., (2013). Investigating the broken-heart effect: a model for short-term dependence between the remaining lifetimes of joint lives. *Annals of Actuarial Science*, 7**(2)**, pp. 236–257.

ZAROUDI, S., BEHZADI, M. H., and FARIDROHANI, M. R., (2018a). Application of Copula in Life Insurance. *International Journal of Applied Mathematics and Statistics*^{TM}, 57**(3)**, pp. 162–168.

ZAROUDI, S., BEHZADI, M. H., and FARIDROHANI, M. R., (2018b). A Copula Approach for Finding the Type of Dependency with Mortality Force Function in Insurance Market. *Journal of Advances and Applications in Statistics*, 53**(2)**, pp. 103–121.

STATISTICS IN TRANSITION new series, March 2021
Vol. 22, No. 1 pp. 227–231,

About the Authors

Al-Noor Nadia H. has completed her PhD in Mathematical Statistics. Currently, she is serving as a Professor of Mathematical Statistics in the Department of Mathematics, College of Science, Mustansiriyah University, Baghdad, Iraq. She has published over 40 research papers in research journals. She has supervised over 10 postgraduate students' dissertations and thesis. Her main research interests are in probability and mathematical statistics, survey sampling, statistical inference, reliability theory, robust regression and non-parametric estimation methods.

Dabgotra Apurba Vishal is a research scholar at the Department of Statistics, University of Jammu, Jammu & Kashmir, India. She has completed her MSc in Statistics from the same institute. Her main areas of interest include: sampling theory, multivariate analysis and categorical data analysis in particular. Currently, she is a full time PhD Scholar of Sampling Survey in the University of Jammu.

Ferretti Camilla graduated summa cum laude in 2005 in Mathematics and received a PhD also in Mathematics in 2010 at the University of Florence (Italy). From 2009 to 2019 she worked as a post-doc research fellow in Economic Statistics at the Department of Economics and Social Sciences of the University of Sacred Heart in Piacenza (Italy). Since July 2020 she has been working as an Expert in Statistics at the Payment System Directorate, Bank of Italy. Her main research interests concern statistical modelling and stochastic processes, with applications to economic problems. Her research results have been published in international journals and presented at prestigious workshops and conferences.

Filip Dariusz is an assistant professor of finance at Cardinal Stefan Wyszynski University in Warsaw (UKSW - Faculty of Social and Economic Sciences). His research fields include the development of financial markets and financial institutions in CEE countries, effectiveness of mutual funds and determinants of their performance. He has published over 30 research papers in international/national journals, such as Baltic Journal of Economics, Finance a úvěr-Czech Journal of Economics and Finance, Prague Economic Papers and Financial Assets and Investing.

Hanif Muhammad has obtained his MSc degree in Statistics from the University of Agriculture, Faisalabad, Pakistan. He has completed his MPhil in Statistics from Government College University, Lahore, Pakistan. He has completed his PhD degree in Statistics from Zhejiang University, Hangzhou, China. Currently, he is an Associate

Professor and serving as a chairman of the Department of Mathematics and Statistics in Pir Mehr Ali Shah Arid Agriculture University Rawalpindi, Pakistan. He has published over 50 research papers in research journals. His main research interest is in stochastic process, probability and mathematical statistics, survey sampling and non parametric estimation.

Ismail Mohd Tahir is an Associate Professor at the School of Mathematical Sciences, Universiti Sains Malaysia. His research interests are financial time series, econometrics, categorical data analysis, and applied statistics. He has published over 150 publications in reviewed journals and proceedings (some of them are listed in ISI, Scopus, Zentralblatt, MathSciNet, and other indices). Currently, he is an Exco Member of the Malaysian Mathematical Sciences Society and an active member of other scientific and professional bodies.

Koyuncu Nursel is a Full Professor at the Department of Statistics, Faculty of Science, Hacettepe University. Her main areas of interest include: sampling, control charts, calibration, simulation techniques.

Krapavickaitė Danutė works as a Professor at the Department of Mathematical Statistics, Vilnius Gediminas Technical University, Lithuania. Main research interests are survey statistics and statistical analysis of data. She has wise experience in the methodological work of official statistics; she is a member of the steering committee of the Baltic-Nordic-Ukrainian network on survey statistics; an editor of the Lithuanian Statistical Journal (2011-2016), an editor of the International Association of Survey Statisticians (IASS) Newsletter The Survey Statistician since July 2018, a member of the professional associations of Lithuania and IASS.

Kumar Sunil is an Assistant Professor at the Department of Statistics, University of Jammu, Jammu and Kashmir (India). Also, he is Visiting Faculty at Indian Statistical Institute, Kolkata (India) and IIM Indore (India). He has received his PhD degree in Statistics at Vikram University (India). He is a referee and an editorial member of several international journals in the frame of Statistics and Management studies. His main research interests are: sample survey, non-response, estimation theory, consumer behaviour, latent class analysis and applications.

Małecka Marta has graduated from University of Łódź, Poland, where she obtained degrees in three study programmes run at two faculties: Faculty of Economics and Sociology and Faculty of Mathematics and Informatics. In 2014 she obtained her PhD degree in Economics at the University of Łódź. She was granted the Award of the Polish Financial Supervision Authority for the doctoral thesis in finance in 2015. In years 2014-2018 she was a coordinator of the National Science Centre project „Hypothesis Testing in Market Risk Evaluation". For over 10 years her academic writing has

explored various aspects of statistical testing in finance. Current areas of focus include market risk management, extreme risk measures and testing risk models.

Mir Shakeel Ahmad is a Full Professor of Agriculture Statistics at Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir. His research interests are field designs/survey strategy development and statistical analysis, optimisation techniques analysis, statistical inference and data analysis in particular. Professor Mir has published over 180 research papers in international/national journals and conferences. He has also published seven books/monographs. Professor Mir is an active member of many scientific professional societies.

Młodak Andrzej is a consultant in the Centre for Small Area Estimation of the Statistical Office in Poznań and an Associate Professor at the Inter-faculty Department of Mathematics and Statistics, Calisia University – Kalisz, Poland. His main areas of interest include: multivariate data analysis, regional statistics, mathematical statistics, mathematical economics and statistical education. Currently, he is a member of two editorial boards: "Wiadomości Statystyczne. The Polish Statistician" (Deputy Editor-in-Chief) and "Statistics in Transition – new series" (Associate Editor).

Ogbonna Ahamuefula E. is a Research Fellow at the Centre for Econometric and Allied Research (CEAR), University of Ibadan, Nigeria and a PhD student in the Department of Statistics, University of Ibadan, Nigeria. His research interests revolve round theoretical and applied econometrics, with recent focus on Bayesian econometrics, financial time series and macroeconomic modelling. He has several research papers in reputable ISI and Scopus ranked journals, applying econometric methodologies like Bayesian Model Averaging, Autoregressive Distributed Lag – MIDAS, Volatility Modelling, Fourier Unit Root testing framework, among other things. He is also competent in programming with relevant analytical software like SPSS, Eviews, STATA, R, RATS and MATLAB, among other things.

Otekunrin Oluwaseun A. is a Senior lecturer in Statistics at the University of Ibadan, Ibadan, Nigeria. She has research publications in reputable local and international journals and is a reviewer of manuscripts for reputable journals. The scope of her research covers constructions of experimental designs with applications in decision sciences and health policy.

Ptak-Chmielewska Aneta is an Associate Professor at the Institute of Statistics and Demography, Collegium of Economic Analysis, Warsaw School of Economics. Simultaneously she holds the position of Credit Risk Model Validation Chapter Lead in ING Tech Poland in Warsaw. Her main areas of interest include: business demography, enterprises survival and bankruptcy measurement, data mining and

machine learning techniques. Currently, she is a member of the Scientific Council of Economy and Finance Discipline in Warsaw School of Economics.

Rogala Tomasz is an Assistant Professor of Mathematics in Institute of Mathematics in the Faculty of Mathematics and Natural Sciences – School of Exact Sciences at Cardinal Stefan Wyszyński University. His research fields include mathematics of finance and probability with special regard to the theory of martingales, optimal stopping, dynamic programming and markets with transaction costs.

Safari-Katesari Hadi is a PhD candidate at the Department of Mathematics, Southern Illinois University, Carbondale, IL, USA. His research interest is multivariate statistical analysis, copula and dependency models and biostatistical data. He has published two research papers in international journals.

Sajjad Irsa has completed her MSc and MPhil in Statistics from Pir Mehr Ali Shah Arid Agriculture University Rawalpindi, Pakistan. She is serving as a lecturer in the Department of Lahore Business School University of Lahore, Islamabad Campus, Pakistan. Her research interests include sampling, stochastic process, methods and probability.

Shah Immad A. is a PhD in Agricultural Statistics and is working as a Research Fellow in a NMHS Fellowship at the Sher-e-Kashmir University of Agricultural Sciences and Technology of Kashmir. His research interests are design of experiments, multivariate techniques, statistical inference and modelling. Dr. Immad has published over 18 research papers in international/national journals and conferences. He has also published a statistical manual on "Application of statistical software packages in Agriculture and allied sciences". Additionally, he has won many awards from various professional organisations and is a merit holder/topper in his doctoral degree programme. He is an active member of many scientific professional societies.

Shahzad Usman has obtained his MSc degree in Statistics from the International Islamic University, Islamabad, Pakistan. He has completed his MPhil in Statistics from Pir Mehr Ali Shah Arid Agriculture University Rawalpindi, Pakistan. Currently, he is studying for his PhD in Statistics from International Islamic University, Islamabad, Pakistan. He has served as a lecturer at the Pir Mehr Ali Shah Arid Agriculture University Rawalpindi, Pakistan. He has published over 40 research papers in research journals. His research interests include survey sampling, extreme value theory, stochastic process, probability and non-parametric statistics.

Yaya OlaOluwa S. is an Assistant Professor at the Department of Statistics, University of Ibadan, Nigeria. His research interests are economic, financial and computational time series with papers written on fractional integration, unit roots, volatility modelling

and regime switching. He has a significant number of published articles in reputable ISI and Scopus journals. He is an active member of professional bodies such as the Royal Statistical Society, International Statistical Institute, Nigerian Statistical Society.

Zaroudi Samira received her PhD in 2018 at the Department of Statistics, Azad University, Tehran, Iran. Currently, she is pursuing her master's degree in Mathematics at Southern Illinois University, Carbondale, IL, USA. Her research interest is multivariate statistical analysis, copula and dependency models and actuarial data. She has published three research papers in international journals.

Żądło Tomasz is an Associate Professor at the Department of Statistics, Econometrics and Mathematics at the University of Economics in Katowice. His research interests focus on small area estimation, survey sampling, mixed models and bootstrap methods. He is an elected member of the International Statistical Institute and a country representative of the International Association of Survey Statisticians.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page*: http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).