STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 15, Number 4, Autumn 2014

Editorial Office **ISSN 1234-7655**
Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Beata Witek: Secretary
b.witek@stat.gov.pl. Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec: Technical Assistant
Address for correspondence
GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 491—498

EDITOR'S NOTE AND ACKNOWLEDGEMENTS

As the last issue of the 2014, this volume gives us an opportunity to thanks all the journal's collaborators and supporters. On behalf of the Editorial Board and the whole Editorial Office, I would like to express our gratefulness to the authors of articles published during the past year, and to acknowledge generous help which both the authors and editors obtained from the peer-reviewers. The names of the reviewers are listed in the Acknowledgements to Reviewers, below.

Two types of innovations mark this issue as an attempt to advance further our cooperation with authors and readers alike. The first one is addition of the *Early View* system for online printing of the articles which are still being under editorial processing though might give authors opportunity to improve them while viewing them in the printed format. The second is continuation of tentatively included into the previous issue of the section containing biographical notes about the authors, with information on their main fields of research interest and expertise. These both innovations are supposed to contribute to the journal's visibility and accessibility which are also enhanced by our continued efforts to have the journal included into the growing set of prestigious indexation bases − such as BazEkon (we still expect confirmation from SCOPUS, RePEc, and so on).

An overview of the contents of this issue embraces three groups of articles. In addition to traditional section devoted to *Sampling methods and estimation*, there is a section *Research articles and communicates* which contains papers of mixed status − next to papers with completed research results it presents 'communicates' with results obtained within an early stage of a larger or still ongoing research project (the latter can also be seen as an innovation toward encouraging submission of fresh-made papers). The third section contains papers based on presentations at the Multivariate Statistical Analysis 2014, conference held in Lodz (on November, 17-19, 2014).

The issue is opened with ***A General Class of Mean Estimators Using Mixture of Auxiliary Variables for Two-phase Sampling in the Presence of Non-response***, by **Zahoor Ahmad**, **Rahma Zubair** and **Ummara Shahid**. The authors suggest a general class of estimators for two-phase sampling to estimate the population mean in the case when non-responses occur at the first phase. Several continuous and categorical auxiliary variable(s) have been simultaneously used while constructing the class. Also, it is assumed that the information on all

auxiliary variables is not available for population, which is often the case. The expressions of the mean square error of the suggested class have been derived and several special cases of the proposed class have been identified. The empirical study has also been conducted. This paper is also aimed at filling the gap in the literature through suggesting estimation of the finite population mean using both qualitative and quantitative multi-auxiliary variables in the presence of non-response at the first phase under two-phase sampling.

Janusz L. Wywial's article ***On Conditional Simple Random Sample*** addresses the issue of estimation of the population average in a finite and fixed population on the basis of the conditional simple random sampling design dependent on order statistics of the auxiliary variable studied. The sampling scheme implementing the sampling design is proposed. The inclusion probabilities are derived. The well-known Horvitz-Thompson statistic under the conditional simple random sampling designs is considered as the estimator of population mean. It has been shown that the Horvitz-Thompson estimator under some particular cases of the conditional simple random sampling design is more accurate than the ordinary mean from the simple random sample.

Marta Zalewska and **Wojciech Zieliński** discuss the problem of ***Statistical Analysis of a Questionnaire: Voluntary Health Insurance Implementation Among Patients Suffering from Allergy and Asthma***. They propose to use a simultaneous confidence intervals in inference about true population proportions in situation when multiple answers in a questionnaire are allowed. A new method of calculating simultaneous confidence regions is provided. It is aimed at improving inference about the population based on such intervals. The inference about the respective population suffering from allergy and asthma proportions requires the construction of a two-dimensional confidence region. Much of authors' attention is paid to the case of three possible answers but the results may be generalized to any questionnaire with more than two excluding answers.

The next section (with research articles and communicates) starts with **Jacek Białek's** paper ***Proposition of Stochastic Postulates for Chain Indices***, which are based on the assumption that prices and quantities are stochastic processes, but the case when price processes are martingales is included too. General conditions which allow the chain indices to satisfy these postulates are discussed with intention to provide an alternative for the classic axiomatic price index theory. The novelty of the presented approach consists in treating the prices and quantities as stochastic processes, and the discussion is meant to introduce the author's future research agenda on chain index theory.

It is followed by a paper ***Lag Length Specification in Engle-Granger Cointegration Test: A Modified Koyck Mean Lag Approach Based on Partial Correlation*** by **Oluokun Kasali Agunloye**, **Dahud Kehinde Shangodoyin** and **Raghunath Arnab.** The authors discuss the problem of limitations of the Engle-Granger cointegration test due to its sensitivity to the choice of lag length and the poor performance of conventional lag selection criteria, such as standard information criteria. Testing for cointegration within the framework of the residual-based Engle-Granger cointegration methodology is the same as testing for the stationarity of the residual series via the augmented Dickey-Fuller test which is well known to be sensitive to the choice of lag length. The researchers are faced with the problem of deciding on the best optimal lag among the candidate optimal lag lengths. This paper introduces a new lag selection criterion called a modified Koyck mean lag approach based on partial correlation criterion for the selection of optimal lag length for the residual-based Engle-Granger cointegration test. Based on empirical findings, it has been observed that in some instances over-specification of lag length can bias the Engle-Granger cointegration test towards the rejection of a true cointegration relationship and non-rejection of a spurious cointegration relationship. Using real-life data, the authors present an empirical illustration which demonstrates that the proposed criterion outperformed the standard information criteria in selecting appropriate optimal truncation lag for the implementation of the Engle-Granger cointegration test using both augmented Dickey-Fuller and generalized least squares Dickey-Fuller tests.

Anna Turczak and **Patrycja Zwiech** discuss the issue of ***Variability of Household Disposable Income per capita by Types of Residence in Poland.*** They use micro-data for the years 1998-2012, though the analysis has been carried out separately for the subsequent years of this period. The study shows that households in Poland are differentiated with regard to income *per capita* by the classes of residence, with the differences within the groups being much bigger than the differences between the groups. What is particularly surprising is that the share of between-group variance in total variance in the population under study has been negligibly small (just a few percent) compared to the share of the mean within-group variance (more than 90 percent). In conclusion, the authors emphasize that the location of a household (city, small town or village) is also significant for the level of household disposable income *per capita*, but the differences are small in comparison to the differences between households of the same classes of residence. Consequently, the authors suggest that more appropriate way of dividing households would be the one explaining better the dispersion of household disposable income *per capita*. The authors are continuing

their analysis towards developing a new classification of households which will be adequate for the problem under study.

In the next article, ***Evaluation of Selected Approaches to Clustering Categorical Variables***, **Zdeněk Šulc** and **Hana Řezanková** consider a set of different similarity measures for defining their contribution to categorical variable clustering. They use three methods of hierarchical cluster analysis (complete, single and average linkage methods) and compare results of cluster analysis using three recent similarity measures (inverse occurrence frequency, occurrence frequency and Lin measures) with results obtained on a basis of two association measures for nominal variables (Cramér's V and the uncertainty coefficient) and the simple matching coefficient (the overlap measure). The quality of clustering is evaluated by the within-cluster variability of created clusters (the lower values the better). The normalized within-cluster mutability coefficient is applied for this purpose. The calculations are made on data from two real datasets (from a social survey).

Finally, two papers from the Multivariate Statistical Analysis 2014 conference constitute the last section of this issue. **Daniel Kosiorowski** compares four methods of forecasting functional time series in the article ***Functional Regression in Short-Term Prediction of Economic Time Series.*** Specifically, themes discussed are fully functional regression, functional autoregression FAR(1) model, and Hyndman and Shang principal component scores forecasting using one-dimensional time series method, and moving functional median. Both simulation studies and an analysis of empirical dataset concerning the Internet users' behaviours for two Internet services in 2013 are employed. In effect, Hyndman & Shao predicting method is shown to outperform other methods in the case of stationary functional time series without outliers. Similarly, the moving functional median induced by Frainman & Muniz depth for functional data outperforms other methods in the case of smooth departures from stationarity of the time series, as well as in the case of functional time series containing outliers.

Marta Małecka's paper ***Duration-Based Approach to VaR Independence Backtesting*** discusses the problem of low power of the VaR-based risk valuation models in investment companies. The problem becomes a particularly serious one in the case of finite-sample settings. A dynamic development in the area of VaR estimation and gradual implementation stimulate the need for statistical methods of VaR models evaluation. Following recent changes in Basel Accords, current UE banking supervisory regulations require internal VaR model backtesting, which gives another strong incentive for research on relevant statistical tests. An alternative to the popular Markov test is sought and the author presents an

overview of the group of duration-based VaR backtesting procedures along with their statistical properties, rejecting a non-realistic assumption of the infinite sample size. The Monte Carlo test technique has been adopted to provide exact tests, in which asymptotic distributions has been replaced with simulated finite sample distributions. A Monte Carlo study (based on the GARCH model) has been designed to investigate the size and the power of the tests. Through the comparative analysis it has been found that, in the light of observed statistical properties, the duration-based approach has been superior to the Markov test.

Włodzimierz Okrasa
Editor

ACKNOWLEDGEMENTS TO REVIEWERS

The Editor and Editorial Board wish to thank the following persons who served from 31 December 2013 to 31 December 2014 as peer-reviewers of manuscripts for the ***Statistics in Transition new series* – Volume 15, Numbers 1–4**; the authors' work has benefited from their feedback.

Abuzinadah Hanaa, King Abdulaziz University, Jeddah, Saudi Arabia

Al-Omari Amer Ibrahim, Al al-Bayt University, Mafraq, Jordan

Andersson Per Gösta, Örebro University, Sweden

Baszczyńska Aleksandra, University of Lodz, Poland

Berger Yves, University of Southampton, United Kingdom

Biecek Przemysław, University of Warsaw, Poland

Bodjanova Slavka, Texas A&M University-Kingsville, USA

Bourguignon Marcelo, Federal University of Pernambuco, Brazil

Brauneis Alexander, University of Klagenfurt, Austria

Bwanakare Second, University of Information Technology and Management in Rzeszow, Poland

Diana Giancarlo, University of Padova, Italy

Dihidar Kajal, Indian Statistical Institute Kolkata, India

Dittmann Paweł, Wroclaw University of Economics, Poland

Domański Czesław, University of Lodz, Poland

Domański Henryk, Polish Academy of Sciences, Warsaw, Poland

Dwivedi Alok, Texas Tech University Health Sciences Center, El Paso, USA

Dwivedi Laxmi Kant, International Institute for Population Sciences, Mumbai, India

Dziechciarz Józef, Wroclaw University of Economics, Poland

Eideh Abdulhakeem A. H., Al-Quds University, Jerusalem, Palestine

Garg Neha, Statistics School of Sciences, IGNOU, New Delhi, India

Getka-Wilczyńska Elżbieta, Warsaw School of Economics, Poland

Gurgul Henryk, AGH University of Science and Technology, Cracow, Poland

Hanagal David D., University of Pune, India and Royal Statistical Society, London, United Kingdom

Hidiroglou Mike, Statistics Canada, Canada

Jajuga Krzysztof, Wroclaw University of Economics, Poland

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 499

SUBMISSION INFORMATION FOR AUTHORS

Statistics in Transition new series (SiT) is an international journal published jointly by the Polish Statistical Association (PTS) and the Central Statistical Office of 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., followed by a hard copy addressed to
Prof. Wlodzimierz Okrasa,
GUS / Central Statistical Office
Al. Niepodległości 208, R. 287, 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, Autumn 2014
Vol. 15, No. 4, pp. 501–524

A GENERAL CLASS OF MEAN ESTIMATORS USING MIXTURE OF AUXILIARY VARIABLES FOR TWO-PHASE SAMPLING IN THE PRESENCE OF NON-RESPONSE

Zahoor Ahmad[1], **Rahma Zubair**[2], **Ummara Shahid**[3]

ABSTRACT

In this paper we have proposed a general class of estimators for two-phase sampling to estimate the population mean in the case when non-responses occur at the first phase. Furthermore, several continuous and categorical auxiliary variable(s) have been simultaneously used while constructing the class. Also, it is assumed that the information on all auxiliary variables is not available for population, which is often the case. The expressions of the mean square error of the suggested class have been derived and several special cases of the proposed class have been identified. The empirical study has also been conducted.

Key words: non-response, multi-auxiliary variables, regression-cum-ratio-exponential estimators, no information case.

1. Introduction

The most common method of data collection in survey research is sending the questionnaire through mail. The reason may be the minimum cost involved in this method. But this method has a major disadvantage that a large rate of non-response may occur, which may result in an unknown bias, while the estimate based only on responding units is representative of both responding and non-responding units.

A personal interview is another method of data collection which generally may result in a complete response, but the cost involved in personal interviews is much higher than the mail questionnaire method. We may conclude from the above discussion that the advantage of one method is the disadvantage of the other and vice versa. Hansen and Hurwitz (1946) combined the advantages of both procedures. They considered the issue of determining the number of mail

[1] Department of Statistics, University of Gujrat, Gujrat, Pakistan. E-mail: zahoor.ahmed@uog.edu.pk.
[2] Department of Statistics, University of Gujrat, Gujrat, Pakistan. E-mail: rahma_stat@ymail.com.
[3] Department of Statistics, University of Gujrat, Gujrat, Pakistan. E-mail: ummara.shahid@uog.edu.pk.

questionnaires along with the number of personal interviews to be carry out given non-response to the mail questionnaire in order to attain the required precision at minimum cost.

Hansen and Hurwitz (1946) discussed the sampling scheme considering non-response and constructed the following unbiased estimator for population mean $\overline{Y}$ of variable of interest y as

$$\overline{y}_1^* = w_1 \overline{y}_1 + w_1' \overline{y}_{1r}, \tag{1.1}$$

where $\overline{y}_1$ and $\overline{y}_{1r}$ denote the means for the respondent and re-contacted sample respectively, and further it is assumed that there is no non-response at re-contacted sample. The weights $w_1 = n_{11}/n_1$ and $w_1' = n_{12}/n_1$.

The variance of (1.1) is

$$Var(\overline{y}_1^*) = \left(N - n_1\right)\left(n_1 N\right)^{-1} S_y^2 + \frac{W_2(k_1 - 1)}{n_1} S_{y_2^*}^2, \tag{1.2}$$

where $S_y^2 = \left(N-1\right)^{-1} \sum_{j=1}^{N} (y_j - \overline{Y})^2$ and $S_{y_2^*}^2 = \left(N_2 - 1\right) \sum_{j=1}^{N_2} (y_j - \overline{Y}_2)^2$ are population variances for responding and non-responding portions with means $\overline{Y} = \left(N\right)^{-1} \sum_{i=1}^{N} y_i$ and $\overline{Y}_2 = \left(N_2\right)^{-1} \sum_{i=1}^{N_2} y_i$, $W_2 = N_2 \left(N\right)^{-1}$.

Singh et al. (2010) emphasized that precision of an estimator can be increased using auxiliary variable in estimation procedure when the study variable *y* is highly correlated with the auxiliary variable *x*. In the case of two phase sampling, Wu and Luan (2003) argue that when we take a large first phase sample from the population and a sub-sample from the first phase sample then there is an issue of small sample size and large non-response rate, and as a result the mean square error becomes larger. This effect can be compensated using auxiliary variables that are highly correlated with the study variable in the estimation procedure. The major advantage of using two-phase sampling is the gain in high precision without substantial increase in cost.

The availability of population auxiliary information plays an important role in efficiency of estimators in two-phase sampling. In the case of at least two auxiliary variables, Samiuddin and Hanif (2007) show that auxiliary information can be utilized in three ways depending on the availability of auxiliary information for population. Firstly, No Information Case (NIC): when population information on all auxiliary variables is not available. Secondly, Partial Information Case (PIC): when population information on some auxiliary variables is available. Thirdly, Full Information Case (FIC): when population information on all auxiliary variables is available. Ahmad and Hanif (2010) clarify that case for a specific estimation procedure, - the estimator for FIC will be more efficient

then the estimator for PIC and the estimator for PIC will me more efficient then the estimator for NIC.

Ahmad et al. (2009a, 2009b, 2010) and Ahmad and Hanif (2010) developed several univariate and multivariate classes of ratio and regression estimators using multi-auxiliary variables under these three cases of availability of auxiliary information for population.

Many survey statisticians have used the quantitative auxiliary variables for constructing their estimators in two-phase sampling. Furthermore, some authors have used qualitative auxiliary variables for estimating the unknown population parameters (see Jhajj et al. (2006), Shabbir and Gupta (2007), Samiuddin and Hanif (2007), Shahbaz and Hanif (2009), Haq et al. (2009), Hanif et al. (2010)).

As mentioned earlier, Hansen-Hurwitz (1946) dealt with non-response problem for simple random sampling and suggested an estimator without using auxiliary information. Many researchers such as Khair and Srivastava (1993,1995), Singh and Kumar (2008a, 2009a) developed different ratios, product and regression estimators to estimate population mean of study variables in two-phase sampling when non-response occurs at the second phase. Tabasum and Khan (2004) revisited the ratio-type estimator by Khair and Srivastava (1993) and found that the cost of this estimator is lower than the cost gained by Hansen-Hurwitz (1946) estimator. Singh et al. (2010) proposed two exponential-type estimators and/or auxiliary variables when non-response occurs during the study.

Ahmad et al. (2012, 2013a, 2013b) proposed the class of generalized estimators to estimate the population mean using multi-auxiliary quantitative variables in the presence of non-responses at the first phase, second phase and both phases.

After introducing the concept of estimating the mean of study variable using a mixture of auxiliary variables in the presence of non-responses, some important references regarding estimators of population mean in the presence of non-responses in single and two-phase sampling using quantitative and qualitative auxiliary variables have been discussed separately in Section 1. In Section 2 we have proposed a generalized class of regression-cum-ratio-exponential estimators for estimating the mean of study variable using a mixture of auxiliary variables in the presence of non-responses at the first phase and its special cases are also given in this section. A detailed empirical study have been conducted and discussed in Section 3. Some conclusions are provided in Section 4.

2. Generalized class of regression-cum-ratio-exponential estimators in two-phase sampling

Most of the literature is devoted to the case when non-responses occur at the second phase, but in two-phase sampling, when auxiliary information is obtained at the first phase sample that is relatively larger than the second phase sample, the non-response rate will be high as compared to the second phase. The two-phase

sampling scheme when non-responses occur at the first phase is discussed as follows.

Consider the total population (denoted by U) of N units is divided into two sections: one is the section (denoted by U_1) of N_1 units, which would be available at the first attempt at the first phase, and the other section (denoted by U_2) of N_2 units, which are not available at the first attempt at the first phase but will be available at the second attempt. From N units, a first phase sample (denoted by u_1) of n_1 units is drawn by simple random sampling without replacement (SRSWOR). At the first phase let m_1' units supply information which is denoted by v_1' and m_2' units refuse to respond, which is denoted by v_2', where $v_1' = u_1 \cap U_1$ and $v_2' = u_1 \cap U_2$. A subsample (denoted by v_{2m}') of r_1 units is randomly taken from the m_2' non-respondents by applying the strategy defined by Hansen and Hurwitz (1946) and this subsample is specified by $r_1 = m_2'/k_1$, $k_1 > 1$. It is assumed that no non-response is observed in this subsample. A second phase sample (denoted by u_2) of n_2 units (i.e. $n_2 < n_1$) is drawn from n_1 by SRSWOR and the variable of interest y is measured at the second phase. The above sampling scheme can be easily understandable from Figure 1.

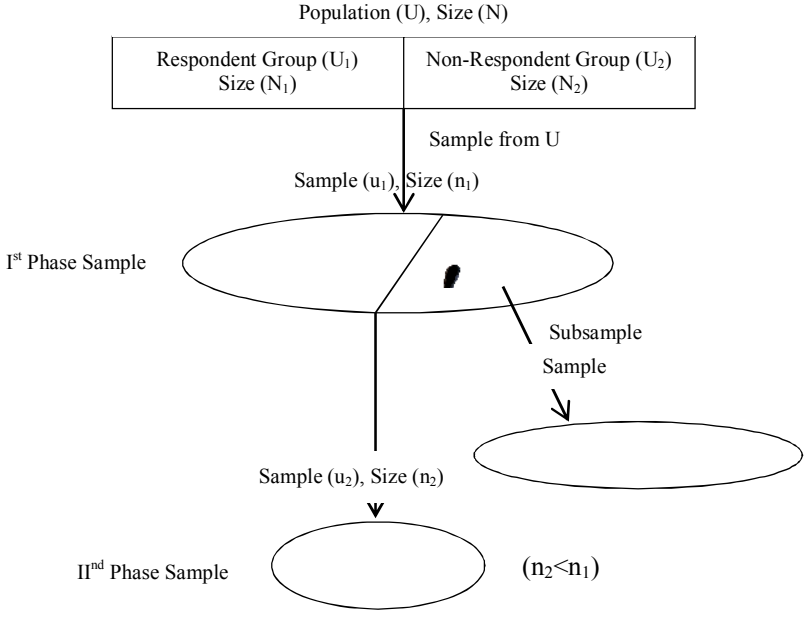

Figure 1. Two-phase sampling scheme when non-responses occur at the first phase

The literature is evident that there is no estimator that can utilize auxiliary information on both quantitative and qualitative variables. But, in sample surveys,

the information on both quantitative and qualitative variables is collected either at the first phase and/or the second phase. For example, we want to estimate the average CGPA of a student in BS (Honor). The information on variables like previous degree marks, attendance, number of hours spent in library, if a student is a member of rural or urban area, father's profession, having a laptop or not, having internet facility or not, etc., can be used as auxiliary information to estimate average CGPA with more efficiency. Hence, there is a need to develop an estimator that can utilize auxiliary information on both quantitative and qualitative variables.

For the first time a combination of regression and ratio technique for simple random sampling called regression-cum-ratio estimator was used by Mohanty (1967) to estimate the population mean of study variable. Similarly, the sum of the ratio and exponential components with some suitable weights can be combined with regression component to develop a general class of regression-cum-ratio-exponential estimators. Furthermore, the objective of suggesting such a class is to search for the best member from all members of the class.

We have suggested the general class of estimators for two-phase sampling to estimate the population mean of the study variable in the case when non-response occurs at the first phase. Moreover, several quantitative and qualitative auxiliary variables have been used simultaneously while constructing the class. Also, it is assumed that population information is not available for all auxiliary variables that is the natural case.

The proposed class is

$$t_{mix} = t_1(t_2 + t_3), \tag{2.1}$$

where

$$t_1 = \eta \bar{y}_2 + a\sum_{i=1}^{q_1} \alpha_{1i}\left(\bar{x}^*_{(1)i} - \bar{x}_{(2)i}\right) + b\sum_{i'=1}^{q_2} \alpha_{2i'}\left(\tau^*_{(1)i'} - \tau_{(2)i'}\right),$$

$$t_2 = c\left[\prod_{j=1}^{q_3}\left(\frac{\bar{z}^*_{(1)j}}{\bar{z}_{(2)j}}\right)^{d\alpha_{3j}} \prod_{j'=1}^{q_4}\left(\frac{\omega^*_{(1)j'}}{\omega_{(2)j'}}\right)^{h\alpha_{4j'}}\right]$$

and $$t_3 = e\exp\left[\sum_{k=1}^{q_5} f\alpha_{5k}\left(\frac{\bar{w}_{(2)k} - \bar{w}^*_{(1)k}}{\bar{w}_{(2)k} + \bar{w}^*_{(1)k}}\right) + \sum_{k'=1}^{q_6} l\alpha_{6k'}\left(\frac{\phi_{(2)k} - \phi^*_{(1)k}}{\phi_{(2)k} + \phi^*_{(1)k}}\right)\right] \text{ for } c + e = 1,$$

where *a, b, c, d, e, f, h, l* are constants to be chosen for generating members of this class and η & α_i's for all $i = 1,2,...,6$ are unknown constants to be determined by minimizing the mean square error of t_{mix} given in (2.1) and $\sum_{i=1}^{6} q_i = m.$ Where

y : denotes the study variable

x_i : denotes the i^{th} auxiliary quantitative variables for $i = 1, 2, 3, ..., q_1$

$\tau_{i'}$: denotes the i'^{th} auxiliary qualitative variables for $i' = 1, 2, 3, ..., q_2$

z_j : denotes the j^{th} auxiliary quantitative variables for $j = 1, 2, 3, ..., q_3$

$\omega_{j'}$: denotes the j'^{th} auxiliary qualitative variables for $j' = 1, 2, 3, ..., q_4$

w_k : denotes the k^{th} auxiliary quantitative variables for $k = 1, 2, 3, ..., q_5$

$\phi_{k'}$: denotes the k'^{th} auxiliary qualitative variables for $k' = 1, 2, 3, ..., q_6$

$\bar{X}_i = N^{-1}\sum_{t=1}^{N} x_{ti}$: denotes the population mean of i^{th} auxiliary variable

$\Phi_{i'} = N^{-1}\sum_{t=1}^{N} \tau_{ti'}$: denotes the population proportion of i'^{th} auxiliary attribute

$\bar{Z}_j = N^{-1}\sum_{t=1}^{N} z_{tj}$: denotes the population mean of j^{th} auxiliary variable

$\Psi_{j'} = N^{-1}\sum_{t=1}^{N} \omega_{tj'}$: denotes the population proportion of j'^{th} auxiliary attribute

$\bar{W}_k = N^{-1}\sum_{t=1}^{N} w_{tk}$: denotes the population mean of k^{th} auxiliary variable

$\mathrm{E}_{k'} = N^{-1}\sum_{t=1}^{N} \phi_t$: denotes the population proportion of k'^{th} auxiliary attribute

$\bar{x}_{(1)i} = n_1^{-1}\sum_{t=1}^{n_1} x_{ti}$: denotes the first phase sample mean of i^{th} auxiliary variable

$\bar{z}_{(1)j} = n_1^{-1}\sum_{t'=1}^{n_1} z_{t'j}$: denotes the first phase sample mean of j^{th} auxiliary variable

$\bar{w}_{(1)k} = n_1^{-1}\sum_{t''=1}^{n_1} w_{t''k}$: denotes the first phase sample mean of k^{th} auxiliary variable

$\tau_{(1)i'} = n_1^{-1}\sum_{t=1}^{n_1} \tau_{ti'}$: denotes the first phase sample proportion of i'^{th} auxiliary attribute

$\omega_{(1)j'} = n_1^{-1}\sum_{t'=1}^{n_1} \omega_{t'j'}$: denotes the first phase sample proportion of j'^{th} auxiliary attribute

$\phi_{(1)k'} = n_1^{-1}\sum_{t^*=1}^{n_1} \phi_{t^*k'}$: denotes the first phase sample proportion of k'^{th} auxiliary attribute

$\bar{y}_2$: denotes the mean of the study variable for the second phase sample

$\overline{x}_{(2)i} = n_2^{-1}\sum_{t=1}^{n_2} x_{ti}$: denotes the second phase sample mean of i^{th} auxiliary variable

$\overline{z}_{(2)j} = n_2^{-1}\sum_{t'=1}^{n_2} z_{t'j}$: denotes the second phase sample mean of j^{th} auxiliary variable

$\overline{w}_{(2)k} = n_2^{-1}\sum_{t''=1}^{n_2} w_{t''k}$: denotes the second phase sample mean of k^{th} auxiliary variable

$\tau_{(2)i'} = n_2^{-1}\sum_{t=1}^{n_2} \tau_{ti'}$: denotes the second phase sample proportion of i'^{th} auxiliary attribute

$\omega_{(2)j'} = n_2^{-1}\sum_{t'=1}^{n_2} \omega_{t'j'}$: denotes the second phase sample proportion of j'^{th} auxiliary attribute

$\phi_{(2)k'} = n_2^{-1}\sum_{t''=1}^{n_2} \phi_{t''k'}$: denotes the second phase sample proportion of k'^{th} auxiliary attribute

$\overline{x}_{(r_1)i} = \frac{1}{r_1}\sum_{t=1}^{r_1} x_t$: denotes the first phase subsample mean of i^{th} auxiliary variable

$\overline{z}_{(r_1)j} = \frac{1}{r_1}\sum_{t=1}^{r_1} z_t$: denotes the first phase subsample mean of j^{th} auxiliary variable

$\overline{w}_{(r_1)k} = \frac{1}{r_1}\sum_{t=1}^{r_1} w_t$: denotes the subsample mean of k^{th} auxiliary variable

$\tau_{(r_1)i'} = \frac{1}{r_1}\sum_{t=1}^{r_1} \tau_{ti'}$: denotes the first phase subsample proportion of i'^{th} auxiliary attribute

$\omega_{(r_1)j'} = \frac{1}{r_1}\sum_{t'=1}^{r_1} \omega_{t'j'}$: denotes the first phase subsample proportion of j'^{th} auxiliary attribute

$\phi_{(r_1)k'} = \frac{1}{r_1}\sum_{t''=1}^{r_1} \phi_{t''k'}$: denotes the first phase subsample proportion of k'^{th} auxiliary attribute

$\overline{x}^{*}_{(1)i} = w_1\overline{x}_{(1)i} + w_2\overline{x}_{(r_1)i}$: denotes the first phase sample mean of i^{th} auxiliary variable considering non-response

$\overline{z}^{*}_{(1)j} = w_1\overline{z}_{(1)j} + w_2\overline{z}_{(r_1)j}$: denotes the first phase sample mean of j^{th} auxiliary variable considering non-response

$\overline{w}^{*}_{(1)k} = w_1\overline{w}_{(1)k} + w_2\overline{w}_{(r_1)k}$: denotes the first phase sample mean of k^{th} auxiliary variable considering non-response

$\tau^{*}_{(1)i'} = w_1\tau_{(1)i'} + w_2\tau_{(r_1)i'}$: denotes the first phase sample proportion of i'^{th} auxiliary variable considering non-response

$\omega^{*}_{(1)j'} = w_1\omega_{(1)j'} + w_2\omega_{(r_1)j'}$: denotes the first phase sample proportion of j'^{th} auxiliary variable considering non-response

$\phi^{*}_{(1)k'} = w_1\phi_{(1)k'} + w_2\phi_{(r_1)k'}$: denotes the first phase sample proportion of k'^{th} auxiliary variable considering non-response

First, considering the regression component t_1 of (2.1), let $\overline{e}_{y(2)} = \overline{y}_{(2)} - \overline{Y}$ be the sampling errors of y, $\overline{e}^{*}_{x(1)i} = \overline{x}^{*}_{(1)i} - \overline{X}_i$ and $\overline{e}^{*}_{\tau(1)i'} = \tau^{*}_{(1)i'} - \Phi_{i'}$ be the sampling errors of i^{th} auxiliary variable and i'^{th} auxiliary attributes respectively in the presence of non-responses at the first phase, and let $\overline{e}_{x(1)i} = \overline{x}_{(1)i} - \overline{X}_i$ and $\overline{e}_{\tau(1)i'} = \tau_{(1)i'} - \Phi_{i'}$ be the sampling errors of i^{th} auxiliary variable and i'^{th} auxiliary attribute respectively at the first phase. Using sampling errors form, and after simplification, t_1 becomes

$$t_1 = \eta\left(\overline{e}_{y(2)} + \overline{Y}\right) + a\sum_{i=1}^{q_1}\alpha_{1i}\left(\overline{e}^{*}_{x(1)i} - \overline{e}_{x(2)i}\right) + b\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\overline{e}^{*}_{\tau(1)i'} - \overline{e}_{\tau(2)i'}\right),$$

In matrix notation we can write

$$t_1 = \eta\left(\overline{e}_{y(2)} + \overline{Y}\right) - a\boldsymbol{\alpha}_1^{\mathbf{t}}\mathbf{d}_{\mathbf{x}} - b\boldsymbol{\alpha}_2^{\mathbf{t}}\mathbf{d}_{\boldsymbol{\tau}}, \tag{2.2}$$

where $\boldsymbol{\alpha}_1^{\mathbf{t}}$ and $\boldsymbol{\alpha}_2^{\mathbf{t}}$ are vectors of an unknown coefficients, and vectors $\mathbf{d}_{\mathbf{x}}^{\mathbf{t}} = \left[d_i\right]_{1\times q_1}$, with $d_i = \left(\overline{e}_{x(2)i} - \overline{e}^{*}_{x(1)i}\right)$ and $\mathbf{d}_{\boldsymbol{\tau}}^{\mathbf{t}} = \left[d_{i'}\right]_{1\times q_2}$, with $d_i = \left(\overline{e}_{\tau(2)i'} - \overline{e}^{*}_{\tau(1)i'}\right)$.

Now, considering the ratio component t_2 of (2.1) let $\overline{e}^{*}_{z(1)j} = \overline{z}^{*}_{(1)j} - \overline{Z}_j$, $\overline{e}^{*}_{\omega(1)j'} = \omega^{*}_{(1)j'} - \Psi_{j'}$ be the sampling errors of j^{th} auxiliary quantitative variable and j'^{th} auxiliary qualitative variable at the first phase in the presence of non-responses and $\overline{e}_{z(1)j} = \overline{z}_{(1)j} - \overline{Z}_j$, $\overline{e}_{\omega(1)j'} = \omega_{(1)j'} - \Psi_{j'}$ be the sampling errors of j^{th} auxiliary quantitative variable and j'^{th} auxiliary qualitative variable at the first phase. Then, simplifying and using binomial expansion up to the second order terms, t_2 becomes

$$t_2 = c\prod_{j=1}^{q_3}\left[1 + \frac{d\alpha_{3j}}{\overline{Z}_j}\overline{e}^{*}_{z(1)j} + \frac{d\alpha_{3j}(d\alpha_{3j}-1)}{2\overline{Z}_j^2}\overline{e}^{*2}_{z(1)j}\right]\left[1 - \frac{d\alpha_{3j}}{\overline{Z}_j}\overline{e}_{z(2)j} + \frac{d\alpha_{3j}(d\alpha_{3j}+1)}{2\overline{Z}_j^2}\overline{e}^{2}_{z(2)j}\right]$$

$$\prod_{j'=1}^{q_4}\left[1+\frac{h\alpha_{4j'}}{\Psi_{j'}}\overline{e}^{*}_{\omega(1)j'}+\frac{h\alpha_{4j'}(h\alpha_{4j'}-1)}{2\Psi^2_{j'}}\overline{e}^{*^2}_{\omega(1)j'}\right]\left[1-\frac{h\alpha_{4j'}}{\Psi_{j'}}\overline{e}_{\omega(2)j'}+\frac{h\alpha_{4j'}(h\alpha_{4j'}+1)}{2\Psi^2_{j'}}\overline{e}^{2}_{\omega(2)j'}\right],$$

or

$$t_2 = c\left[1-\sum_{j=1}^{q_3}\left\{\frac{d\alpha_{3j}}{\overline{Z}_j}\left(\overline{e}_{z(2)j}-\overline{e}^{*}_{z(1)j}\right)+\frac{d^2\alpha^2_{3j}}{2\overline{Z}^2_j}\left(\overline{e}^{2}_{z(2)j}+\overline{e}^{2^*}_{z(1)j}\right)+\frac{d\alpha_{3j}}{2\overline{Z}^2_j}\left(\overline{e}^{2}_{z(2)j}-\overline{e}^{2^*}_{z(1)j}\right)\right\}\right]$$

$$\left[1-\sum_{j'=1}^{q_4}\left\{\frac{h\alpha_{4j}}{\Psi_{j'}}\left(\overline{e}_{\omega(2)j'}-\overline{e}^{*}_{\omega(1)j'}\right)+\frac{h^2\alpha^2_{4j'}}{2\Psi^2_{j'}}\left(\overline{e}^{2}_{\omega(2)j'}+\overline{e}^{2^*}_{\omega(1)j'}\right)+\frac{h\alpha_{4j'}}{2\Psi^2_{j'}}\left(\overline{e}^{2}_{\omega(2)j'}-\overline{e}^{2^*}_{\omega(1)j'}\right)\right\}\right].$$

Ignoring the third and higher order terms and writing in matrix notations, we have

$$t_2 = c\Big[1-h\boldsymbol{\alpha}_4^{\mathbf{t}}\boldsymbol{\Psi}\mathbf{d}_{\boldsymbol{\omega}}-2^{-1}h^2\boldsymbol{\alpha}_4^{2^{\mathbf{t}}}\boldsymbol{\Psi}^2\mathbf{u}_{\boldsymbol{\omega}}+2^{-1}h\boldsymbol{\alpha}_4^{\mathbf{t}}\boldsymbol{\Psi}^2\mathbf{v}_{\boldsymbol{\omega}}-d\boldsymbol{\alpha}_3^{\mathbf{t}}\mathbf{Z}\mathbf{d}_{\mathbf{z}}$$
$$+hd\boldsymbol{\alpha}_3^{\mathbf{t}}\mathbf{Z}\mathbf{d}_{\mathbf{z}}\boldsymbol{\alpha}_4^{\mathbf{t}}\boldsymbol{\Psi}\mathbf{d}_{\boldsymbol{\omega}}+2^{-1}d^2\boldsymbol{\alpha}_3^{2^{\mathbf{t}}}\mathbf{Z}^2\mathbf{u}_{\mathbf{z}}+2^{-1}d\boldsymbol{\alpha}_3^{\mathbf{t}}\mathbf{Z}^2\mathbf{v}_{\mathbf{z}}\Big], \qquad (2.3)$$

where $\boldsymbol{\alpha}_3^{\mathbf{t}}$ and $\boldsymbol{\alpha}_4^{\mathbf{t}}$ are vectors of unknown coefficient, $\boldsymbol{\Psi}=\left[\Psi^{-1}_{j'}\right]_{q_4\times q_4}$; $j'=1,2,3...q_4$, and $\mathbf{Z}=\left[\overline{Z}^{-1}_{j}\right]_{q_3\times q_3}$; $j=1,2,3...q_3$ are diagonal matrices and vectors $\mathbf{d}_{\mathbf{z}}^{\mathbf{t}}=\left[d_j\right]_{1\times q_3}$ with $d_j=\left(\overline{e}_{z(2)j}-\overline{e}^{*}_{z(1)j}\right)$, $\mathbf{d}_{\boldsymbol{\omega}}^{\mathbf{t}}=\left[d_{j'}\right]_{1\times q_4}$ with $d_{j'}=\left(\overline{e}_{\omega(2)j'}-\overline{e}^{*}_{\omega(1)j'}\right)$, $\mathbf{u}_{\boldsymbol{\omega}}^{\mathbf{t}}=\left[u_{j'}\right]_{1\times q_4}$ with $u_{j'}=\left(e^{2}_{\omega_{2j'}}+e^{2^*}_{\omega_{1j'}}\right)$, $\mathbf{u}_{\mathbf{z}}^{\mathbf{t}}=\left[u_j\right]_{1\times q_3}$ with $u_j=\left(e^{2}_{z_{2j}}+e^{2^*}_{z_{1i}}\right)$, $\mathbf{v}_{\boldsymbol{\omega}}^{\mathbf{t}}=\left[v_{j'}\right]_{1\times q_4}$ with $v_{j'}=\left(e^{2}_{\omega_{2j'}}-e^{2^*}_{\omega_{1j'}}\right)$ and $\mathbf{v}_{\mathbf{z}}^{\mathbf{t}}=\left[v_j\right]_{1\times q_3}$ with $v_j=\left(e^{2}_{z_{2i}}-e^{2^*}_{z_{1j}}\right)$.

Now, considering the exponential component (t_3) of (2.1), let $\overline{e}^{*}_{w(1)k}=\overline{w}^{*}_{(1)k}-\overline{W}_k$, $\overline{e}^{*}_{\phi(1)k'}=\phi^{*}_{(1)k'}-\mathrm{E}_{k'}$ $\overline{e}^{*}_{\phi(2)k'}=\phi^{*}_{(2)k'}-\mathrm{E}_{k'}$ be the sampling errors of k^{th} quantitative auxiliary variable and k'^{th} qualitative auxiliary variable at the first phase with non-response and $\overline{e}_{w(1)k}=\overline{w}_{(1)k}-\overline{W}_k$, $\overline{e}_{\phi(1)k'}=\phi_{(1)k'}-\mathrm{E}_{k'}$ be the sampling errors of k^{th} quantitative auxiliary variable and k'^{th} qualitative auxiliary variable at the first phase. Then, simplifying and using binomial expansion up to the second order term, t_3 becomes

$$t_3 = e\exp\left[\sum_{k=1}^{q_5} f\alpha_{5k}\left(\frac{\overline{e}_{\overline{w}_{(2)k}}-\overline{e}^{*}_{\overline{w}_{(1)k}}}{\overline{e}_{\overline{w}_{(2)k}}+\overline{e}^{*}_{\overline{w}_{(1)k}}+2\overline{W}_k}\right)+\sum_{k'=1}^{q_6} l\alpha_{6k'}\left(\frac{\overline{e}_{\phi_{(2)k'}}-\overline{e}^{*}_{\phi_{(1)k'}}}{\overline{e}_{\phi_{(2)k'}}+\overline{e}^{*}_{\phi_{(1)k'}}+2\mathrm{E}_{k'}}\right)\right],$$

or

$$t_3 = e\exp\left[\sum_{k=1}^{q_5}\frac{f\alpha_{5k}}{2\overline{W}_k}\left(\overline{e}_{\overline{w}_{(2)k}}-\overline{e}^{*}_{\overline{w}_{(1)k}}\right)\left(1+\frac{\overline{e}_{\overline{w}_{(2)k}}+\overline{e}^{*}_{\overline{w}_{(1)k}}}{2\overline{W}_k}\right)^{-1}+\sum_{k'=1}^{q_6}\frac{l\alpha_{6k'}}{2\mathrm{E}_{k'}}\left(\overline{e}_{\phi_{(2)k'}}-\overline{e}^{*}_{\phi_{(1)k'}}\right)\left(1+\frac{\overline{e}_{\phi_{(2)k'}}+\overline{e}^{*}_{\phi_{(1)k'}}}{2\mathrm{E}_{k'}}\right)^{-1}\right]$$

or
$$t_3 = e\exp\left[\sum_{k=1}^{q_5}\frac{f\alpha_{5k}}{2\bar{W}_k}\left(\bar{e}_{\bar{w}_{(2)k}}-\bar{e}^{*}_{\bar{w}_{(1)k}}\right)\left(1-\frac{\bar{e}_{\bar{w}_{(2)k}}+\bar{e}^{*}_{\bar{w}_{(1)k}}}{2\bar{W}_k}+\frac{\left(\bar{e}_{\bar{w}_{(2)k}}+\bar{e}^{*}_{\bar{w}_{(1)k}}\right)^2}{4\bar{W}_k^2}\right)\right.$$
$$\left.+\sum_{k'=1}^{q_6}\frac{l\alpha_{6k'}}{2\mathrm{E}_{k'}}\left(\bar{e}_{\phi_{(2)k'}}-\bar{e}^{*}_{\phi_{(1)k'}}\right)\left(1-\frac{\bar{e}_{\phi_{(2)k'}}+\bar{e}^{*}_{\phi_{(1)k'}}}{2\mathrm{E}_{k'}}+\frac{\left(\bar{e}_{\phi_{(2)k'}}+\bar{e}^{*}_{\phi_{k'}}\right)^2}{4\mathrm{E}^2{}_{k'}}\right)\right]$$

Using exponential series and writing in matrix notation, after ignoring the third and higher order terms
$$t_3 = e + 2^{-1}ef\boldsymbol{\alpha}_5^{\mathrm{t}}\mathbf{W}\mathbf{d}_{\mathbf{w}} - 4^{-1}ef\boldsymbol{\alpha}_5^{\mathrm{t}}\mathbf{W}^2\mathbf{v}_{\mathbf{w}} + 2^{-1}el\boldsymbol{\alpha}_6^{\mathrm{t}}\mathbf{E}\mathbf{d}_{\mathbf{f}} - 4^{-1}el\boldsymbol{\alpha}_6^{\mathrm{t}}\mathbf{E}^2\mathbf{v}_{\mathbf{f}}, \tag{2.4}$$

where $\boldsymbol{\alpha}_5^{\mathrm{t}}$ and $\boldsymbol{\alpha}_6^{\mathrm{t}}$ are vectors of unknown coefficient, $\mathbf{W} = \left[\bar{W}_k^{-1}\right]_{q_{5\times}q_5}$; $k = 1,2,3...q_5$ and $\mathbf{E} = \left[\mathrm{E}_{k'}^{-1}\right]_{q_{6\times}q_6}$; $k' = 1,2,3...q_6$ are diagonal matrices and vectors $\mathbf{d}_{\mathbf{w}}^{\mathrm{t}} = \left[d_k\right]_{1\times q_5}$ with $d_k = \left(\bar{e}_{w(2)k} - \bar{e}^{*}_{w(1)k}\right)$, $\mathbf{d}_{\phi}^{\mathrm{t}} = \left[d_{k'}\right]_{1\times q_6}$ with $d_{k'} = \left(\bar{e}_{\phi(2)k'} - \bar{e}^{*}_{\phi(1)k'}\right)$, $\mathbf{v}_{\mathbf{w}}^{\mathrm{t}} = \left[v_k\right]_{1\times q_5}$ with $v_k = \left(e^2_{w_{2k}} - e^{2^*}_{w_{1k}}\right)$ and $\mathbf{v}_{\phi}^{\mathrm{t}} = \left[v_{k'}\right]_{1\times q_6}$ with $v_{k'} = \left(e^2_{\phi_{2k'}} - e^{2^*}_{\phi_{1k'}}\right)$.

Substituting the expressions of t_1, t_2 and t_3 from (2.2), (2.3) and (2.4) in (2.1), we get
$$\begin{aligned}
t_{mix} = {} & \left[\eta\left(\bar{e}_{y(2)} + \bar{Y}\right) - a\boldsymbol{\alpha}_1^{\mathrm{t}}\mathbf{d}_{\mathbf{x}} - b\boldsymbol{\alpha}_2^{\mathrm{t}}\mathbf{d}_{\tau}\right]\left[c - ch\boldsymbol{\alpha}_4^{\mathrm{t}}\boldsymbol{\Psi}\mathbf{d}_{\boldsymbol{\omega}} - 2^{-1}ch^2\boldsymbol{\alpha}_4^{2^{\mathrm{t}}}\boldsymbol{\Psi}^2\mathbf{u}_{\boldsymbol{\omega}}\right.\\
& + 2^{-1}ch\boldsymbol{\alpha}_4^{\mathrm{t}}\boldsymbol{\Psi}^2\mathbf{v}_{\boldsymbol{\omega}} - cd\boldsymbol{\alpha}_3^{\mathrm{t}}\mathbf{Z}\mathbf{d}_{\mathbf{x}} + chd\boldsymbol{\alpha}_3^{\mathrm{t}}\mathbf{Z}\mathbf{d}_{\mathbf{x}}\boldsymbol{\alpha}_4^{\mathrm{t}}\boldsymbol{\Psi}\mathbf{d}_{\boldsymbol{\omega}}\\
& + 2^{-1}cd^2\boldsymbol{\alpha}_3^{2^{\mathrm{t}}}\mathbf{Z}^2\mathbf{u}_{\mathbf{z}} + 2^{-1}cd\boldsymbol{\alpha}_3^{\mathrm{t}}\mathbf{Z}^2\mathbf{v}_{\mathbf{z}} + e + ef\,2^{-1}\boldsymbol{\alpha}_5^{\mathrm{t}}\mathbf{W}\mathbf{d}_{\mathbf{w}}\\
& \left. - 4^{-1}ef\boldsymbol{\alpha}_5^{\mathrm{t}}\mathbf{W}^2\mathbf{v}_{\mathbf{w}} + 2^{-1}el\boldsymbol{\alpha}_6^{\mathrm{t}}\mathbf{E}\mathbf{d}_{\phi} - 4^{-1}el\boldsymbol{\alpha}_6^{\mathrm{t}}\mathbf{E}^2\mathbf{v}_{\phi}\right]; c + e = 1
\end{aligned} \tag{2.5}$$

Ignoring the third and higher order terms of the expression given by (2.5) and applying the expectation, we get
$$\begin{aligned}
E\left(t_{mix} - \bar{Y}\right) = {} & (\eta - 1)\bar{Y} - \eta ch\boldsymbol{\alpha}_4^{\mathrm{t}}\boldsymbol{\Psi}\boldsymbol{\delta}_4 - \eta cd\boldsymbol{\alpha}_3^{\mathrm{t}}\mathbf{Z}\boldsymbol{\delta}_3 + 2^{-1}\eta ef\boldsymbol{\alpha}_5^{\mathrm{t}}\mathbf{W}\boldsymbol{\delta}_5\\
& + 2^{-1}\eta el\boldsymbol{\alpha}_6^{\mathrm{t}}\mathbf{E}\boldsymbol{\delta}_6 + 2^{-1}\bar{Y}\eta ch^2\boldsymbol{\alpha}_4^{2^{\mathrm{t}}}\boldsymbol{\Psi}^2\left(\lambda_4\mathbf{S}_{\boldsymbol{\omega}}^2 + \theta\mathbf{S}_{\boldsymbol{\omega}_{(2)}}^2\right)\\
& + \bar{Y}\eta chd\boldsymbol{\alpha}_3^{\mathrm{t}}\mathbf{Z}\boldsymbol{\Delta}_{34}\boldsymbol{\Psi}\boldsymbol{\alpha}_4 + 2^{-1}\bar{Y}\eta cd^2\boldsymbol{\alpha}_3^{2^{\mathrm{t}}}\mathbf{Z}^2\left(\lambda_4\mathbf{S}_{\mathbf{z}}^2 + \theta\mathbf{S}_{\mathbf{z}_{(2)}}^2\right)\\
& + 2^{-1}\eta\bar{Y}ch\boldsymbol{\alpha}_4^{\mathrm{t}}\boldsymbol{\Psi}^2\left(\lambda_3\mathbf{S}_{\boldsymbol{\omega}}^2 + \theta\mathbf{S}_{\boldsymbol{\omega}_{(2)}}^2\right) + 2^{-1}\eta\bar{Y}cd\boldsymbol{\alpha}_3^{\mathrm{t}}\mathbf{Z}^2\left(\lambda_3\mathbf{S}_{\mathbf{z}}^2 + \theta\mathbf{S}_{\mathbf{z}_{(2)}}^2\right)\\
& - 4^{-1}\bar{Y}\eta ef\boldsymbol{\alpha}_5^{\mathrm{t}}\mathbf{W}^2\left(\lambda_3\mathbf{S}_{\mathbf{w}}^2 + \theta\mathbf{S}_{\mathbf{w}_{(2)}}^2\right) - 4^{-1}\bar{Y}\eta el\boldsymbol{\alpha}_6^{\mathrm{t}}\mathbf{E}\left(\lambda_3\mathbf{S}_{\phi}^2 + \theta\mathbf{S}_{\phi_{(2)}}^2\right)\\
& + ach\boldsymbol{\alpha}_1^{\mathrm{t}}\boldsymbol{\Delta}_{44}\boldsymbol{\Psi}\boldsymbol{\alpha}_4 + acd\boldsymbol{\alpha}_1^{\mathrm{t}}\boldsymbol{\Delta}_{13}\mathbf{Z}\boldsymbol{\alpha}_3 - 2^{-1}aef\boldsymbol{\alpha}_1^{\mathrm{t}}\boldsymbol{\Delta}_{15}\mathbf{W}\boldsymbol{\alpha}_5\\
& - 2^{-1}ael\boldsymbol{\alpha}_1^{\mathrm{t}}\boldsymbol{\Delta}_{16}\mathbf{E}\boldsymbol{\alpha}_6 + bch\boldsymbol{\alpha}_2^{\mathrm{t}}\boldsymbol{\Delta}_{44}\boldsymbol{\Psi}\boldsymbol{\alpha}_4 + bcd\boldsymbol{\alpha}_2^{\mathrm{t}}\boldsymbol{\Delta}_{23}\mathbf{Z}\boldsymbol{\alpha}_3\\
& - 2^{-1}bef\boldsymbol{\alpha}_2^{\mathrm{t}}\boldsymbol{\Delta}_{25}\mathbf{W}\boldsymbol{\alpha}_5 - 2^{-1}bel\boldsymbol{\alpha}_2^{\mathrm{t}}\boldsymbol{\Delta}_{26}\mathbf{E}\boldsymbol{\alpha}_6,
\end{aligned} \tag{2.6}$$

where

$$\boldsymbol{\delta_1} = E\left(\mathbf{d_x} e_{\bar{y}_2}\right) = \left[\lambda_3 S_{x_i y}\right]_{q_1 \times 1},\ \boldsymbol{\delta_2} = E\left(\mathbf{d_\tau} e_{\bar{y}_2}\right) = \left[\lambda_3 S_{\tau_{i'} y}\right]_{q_2 \times 1}, \boldsymbol{\delta_3} = E\left(\mathbf{d_z} e_{\bar{y}_2}\right) = \left[\lambda_3 S_{z_j y}\right]_{q_3 \times 1}$$

$$\boldsymbol{\delta_4} = E\left(\mathbf{d_\omega} e_{\bar{y}_2}\right) = \left[\lambda_3 S_{\omega_{j'} y}\right]_{q_4 \times 1}, \boldsymbol{\delta_5} = E\left(\mathbf{d_w} e_{\bar{y}_2}\right) = \left[\lambda_3 S_{w_k y}\right]_{q_5 \times 1},\ \boldsymbol{\delta_6} = E\left(\mathbf{d}_\phi e_{\bar{y}_2}\right) = \left[\lambda_3 S_{\phi_{k'} y}\right]_{q_6 \times 1}$$

$$E\left(\mathbf{u}_\omega\right) = \left(\lambda_4 \mathbf{S}_\omega^2 + \theta \mathbf{S}_{\omega_{(2)}}^2\right), E\left(\mathbf{v}_\omega\right) = \left(\lambda_3 \mathbf{S}_\omega^2 + \theta \mathbf{S}_{\omega_{(2)}}^2\right) E\left(\mathbf{u}_\mathbf{z}\right) = \left(\lambda_4 \mathbf{S}_\mathbf{z}^2 + \theta \mathbf{S}_{\mathbf{z}_{(2)}}^2\right); \lambda_4 = \left(\lambda_2 + \lambda_1\right),$$

$$E\left(\mathbf{v}_\mathbf{z}\right) = \left(\lambda_3 \mathbf{S}_\mathbf{z}^2 + \theta \mathbf{S}_{\mathbf{z}_{(2)}}^2\right), E\left(\mathbf{v}_\mathbf{w}\right) = \left(\lambda_3 \mathbf{S}_\mathbf{w}^2 + \theta \mathbf{S}_{\mathbf{w}_{(2)}}^2\right), E\left(\mathbf{v}_\phi\right) = \left(\lambda_3 \mathbf{S}_\phi^2 + \theta \mathbf{S}_{\phi_{(2)}}^2\right); \lambda_3 = \left(\lambda_2 - \lambda_1\right),$$

$$\boldsymbol{\Delta_{13}} = E(\mathbf{d_x} \mathbf{d_z^t}) = \left[\lambda_3 S_{x_i z_j} + \theta S_{x_i z_{j(2)}}\right]_{q_1 \times q_3},\ \boldsymbol{\Delta_{15}} = E(\mathbf{d_x} \mathbf{d_w^t}) = \left[\lambda_3 S_{x_i w_k} + \theta S_{x_i w_{k(2)}}\right]_{q_1 \times q_5},$$

$$\boldsymbol{\Delta_{16}} = E(\mathbf{d_x} \mathbf{d_\phi^t}) = \left[\lambda_3 S_{x_i \phi_{k'}} + \theta S_{x_i \phi_{k'(2)}}\right]_{q_1 \times q_6},\ \boldsymbol{\Delta_{23}} = E(\mathbf{d_\tau} \mathbf{d_z^t}) = \left[\lambda_3 S_{\tau_{i'} z_j} + \theta S_{\tau_{i'} z_{j(2)}}\right]_{q_2 \times q_3},$$

$$\boldsymbol{\Delta_{25}} = E(\mathbf{d_\tau} \mathbf{d_w^t}) = \left[\lambda_3 S_{\tau_{i'} w_k} + \theta S_{\tau_{i'} w_{k(2)}}\right]_{q_2 \times q_5},\ \boldsymbol{\Delta_{26}} = E(\mathbf{d_\tau} \mathbf{d_\phi^t}) = \left[\lambda_3 S_{\tau_{i'} \phi_{k'}} + \theta S_{\tau_{i'} \phi_{k'(2)}}\right]_{q_2 \times q_6},$$

$$\boldsymbol{\Delta_{34}} = E(\mathbf{d_z} \mathbf{d_\omega^t}) = \left[\lambda_3 S_{z_j \omega_{j'}} + \theta S_{z_j \omega_{j'(2)}}\right]_{q_3 \times q_4},\ \boldsymbol{\Delta_{44}} = E(\mathbf{d_\omega} \mathbf{d_\omega^t}) = \left[\lambda_3 S_{\omega_{j'} \omega_{j'}} + \theta S_{\omega_{j'} \omega_{j'(2)}}\right]_{q_4 \times q_4}.$$

Expression given in (2.6) can be written as

$$\begin{aligned}
Bias(t_m) &= (\eta - 1)\bar{Y} + cd\Big[a\boldsymbol{\alpha}_1^t \boldsymbol{\Delta}_{13} \mathbf{Z} \boldsymbol{\alpha}_3 + b\boldsymbol{\alpha}_2^t \boldsymbol{\Delta}_{23} \mathbf{Z} \boldsymbol{\alpha}_3 + 2^{-1} \eta \bar{Y} d \boldsymbol{\alpha}_3^{2t} \mathbf{Z}^2 \left(\lambda_4 \mathbf{S}_\mathbf{z}^2\right. \\
&+ \theta \mathbf{S}_{(2)\mathbf{z}}^2\left.\right) + 2^{-1} \bar{Y} \eta \boldsymbol{\alpha}_3^t \mathbf{Z}^2 \left(\lambda_3 \mathbf{S}_\mathbf{z}^2 + \theta \mathbf{S}_{(2)\mathbf{z}}^2\right) - \eta \boldsymbol{\alpha}_3^t \mathbf{Z} \boldsymbol{\delta}_3 \Big] \\
&+ ch\Big[a\boldsymbol{\alpha}_1^t \boldsymbol{\Delta}_{14} \boldsymbol{\Psi} \boldsymbol{\alpha}_4 + b\boldsymbol{\alpha}_2^t \boldsymbol{\Delta}_{24} \boldsymbol{\Psi} \boldsymbol{\alpha}_4 + 2^{-1} \bar{Y} \eta d \boldsymbol{\alpha}_4^{2t} \boldsymbol{\Psi}^2 (\lambda_4 \mathbf{S}_\omega^2 \\
&+ \theta \mathbf{S}_{(2)\omega}^2 + 2^{-1} \bar{Y} \eta \boldsymbol{\alpha}_4^t \boldsymbol{\Psi}^2 \left(\lambda_3 \mathbf{S}_\omega^2 + \theta \mathbf{S}_{(2)\omega}^2\right) - \eta \boldsymbol{\alpha}_4^t \boldsymbol{\Psi} \boldsymbol{\delta}_4 \\
&+ \bar{Y} \eta d \boldsymbol{\alpha}_3^t \mathbf{Z} \boldsymbol{\Delta}_{34} \boldsymbol{\Psi}\Big] + ef\left[-2^{-1} a \boldsymbol{\alpha}_1^t \boldsymbol{\Delta}_{15} \mathbf{W} \boldsymbol{\alpha}_5 - 2^{-1} b \boldsymbol{\alpha}_2^t \boldsymbol{\Delta}_{25} \mathbf{W} \boldsymbol{\alpha}_5\right. \\
&- 4^{-1} \bar{Y} \eta \boldsymbol{\alpha}_5^t \mathbf{W}^2 (\lambda_3 \mathbf{S}_\mathbf{w}^2 + \theta \mathbf{S}_{(2)\mathbf{w}}^2) + 2^{-1} \eta \boldsymbol{\alpha}_5^t \mathbf{W} \boldsymbol{\delta}_5\Big] + el\Big[-2^{-1} a \boldsymbol{\alpha}_1^t \boldsymbol{\Delta}_{16} \mathbf{E} \boldsymbol{\alpha}_6 \\
&- 2^{-1} b \boldsymbol{\alpha}_2^t \boldsymbol{\Delta}_{26} \mathbf{E} \boldsymbol{\alpha}_6 - 4^{-1} \bar{Y} \eta \boldsymbol{\alpha}_6^t \mathbf{E}^2 \left(\lambda_3 \mathbf{S}_\phi^2 + \theta \mathbf{S}_{(2)\phi}^2\right) + 2^{-1} \eta \boldsymbol{\alpha}_6^t \mathbf{E} \boldsymbol{\delta}_6\Big]. \qquad (2.7)
\end{aligned}$$

$Bias(t_{mix})$ = Bias due to regression – cum – ratio (quantitative) + Bias due to regression – cum – ratio (qualitative) + Bias due to regression – cum – exponential (quantitative) + Bias due to regression – cum – exponential (qualitative)

For obtaining the mean square error and optimum value of generalized class, ignoring the second and higher order terms after multiplication from (2.5), we have

$$\begin{aligned}
t_{mix} - \bar{Y} &= \eta \bar{e}_{y_{(2)}} + (\eta - 1)\bar{Y} - a\boldsymbol{\alpha}_1^t \mathbf{d_x} - b\boldsymbol{\alpha}_2^t \mathbf{d}_\tau - \bar{Y} \eta cd \boldsymbol{\alpha}_3^t \mathbf{Z} \mathbf{d_z} \\
&\quad - \bar{Y} \eta ch \boldsymbol{\alpha}_4^t \boldsymbol{\Psi} \mathbf{d}_\omega + 2^{-1} \bar{Y} \eta ef \boldsymbol{\alpha}_5^t \mathbf{W} \mathbf{d_w} + 2^{-1} \bar{Y} \eta el \boldsymbol{\alpha}_6^t \mathbf{E} \mathbf{d}_\phi,
\end{aligned}$$

or

$$t_{mix} - \bar{Y} = \eta \bar{e}_{y_{(2)}} + (\eta - 1)\bar{Y} - \mathbf{h}^t \mathbf{H},$$

where $\mathbf{h}^t = \left[\boldsymbol{\alpha}_1^t \quad \boldsymbol{\alpha}_2^t \quad \boldsymbol{\alpha}_3^t \quad \boldsymbol{\alpha}_4^t \quad \boldsymbol{\alpha}_5^t \quad \boldsymbol{\alpha}_6^t\right]_{1 \times m}$

and $\mathbf{H}^{t} = \left[a\mathbf{d}_{x} \;\; b\mathbf{d}_{\tau} \;\; \bar{Y}\eta cd\mathbf{Z}\mathbf{d}_{z} \;\; \bar{Y}\eta ch\mathbf{\Psi}\mathbf{d}_{\omega} \;\; -2^{-1}\bar{Y}\eta ef\mathbf{W}\mathbf{d}_{w} \;\; -2^{-1}\bar{Y}\eta el\mathbf{E}\mathbf{d}_{\phi} \right]_{1\times m}$.

Squaring and taking the expectation, we have

$$E(t_{mix} - \bar{Y})^{2} = E(\eta\bar{e}_{y_{(2)}} + (\eta - 1)\bar{Y} - \mathbf{h}^{t}\mathbf{H})^{2}. \tag{2.8}$$

To find the optimum value of the unknown vector of row vectors *h* for which mean square error will be the minimum, differentiating (4.8) with respect to *h* and equating to zero, we get

$$\eta E(\mathbf{H}\bar{e}_{y_{(2)}}) + (\eta - 1)\bar{Y}E(\mathbf{H}) - E(\mathbf{H}\mathbf{H}^{t})\mathbf{h} = \mathbf{0}$$

or

$$\eta\mathbf{\Omega} - \mathbf{\Lambda}\mathbf{h} = \mathbf{0}, \tag{2.9}$$

where

$E(\mathbf{H}\bar{e}_{y_2}) = \mathbf{\Omega}$, with, $\mathbf{\Omega}^{t} = \left[a\boldsymbol{\delta}_{1} \;\; b\boldsymbol{\delta}_{2} \;\; \bar{Y}cd\mathbf{Z}\boldsymbol{\delta}_{3} \;\; \bar{Y}ch\mathbf{\Psi}\boldsymbol{\delta}_{4} \;\; -2^{-1}\bar{Y}ef\mathbf{W}\boldsymbol{\delta}_{5} \;\; -2^{-1}\bar{Y}el\mathbf{E}\boldsymbol{\delta}_{6} \right]$,

$E(\mathbf{H}) = \mathbf{0}$ and $E(\mathbf{H}\mathbf{H}^{t}) = \mathbf{\Lambda} = \left[\Lambda_{ij} \right]_{m\times m}$. (2.10)

The elements in $\left[\Lambda_{ij} \right]$ are

$\mathbf{\Lambda}_{11} = a^{2}E(\mathbf{d}_{x}\mathbf{d}_{x}^{t}) = a^{2}\mathbf{\Delta}_{11}$, $\mathbf{\Lambda}_{12} = aE(\mathbf{d}_{x}\mathbf{d}_{\tau}^{t})b = ab\mathbf{\Delta}_{12}$,

$\mathbf{\Lambda}_{13} = \eta acd\bar{Y}E(\mathbf{d}_{x}\mathbf{d}_{z}^{t})\mathbf{Z} = \eta acd\bar{Y}\mathbf{\Delta}_{13}\mathbf{Z}$, $\mathbf{\Lambda}_{14} = \bar{Y}\eta achE(\mathbf{d}_{x}\mathbf{d}_{\omega}^{t})\mathbf{\Psi} = \bar{Y}\eta ach\mathbf{\Delta}_{14}\mathbf{\Psi}$,

$\mathbf{\Lambda}_{15} = -2^{-1}\bar{Y}\eta aefE(\mathbf{d}_{x}\mathbf{d}_{w}^{t})\mathbf{W} = -2^{-1}\bar{Y}\eta aef\mathbf{\Delta}_{15}\mathbf{W}$,

$\mathbf{\Lambda}_{16} = -2^{-1}\bar{Y}\eta ael(\mathbf{d}_{x}\mathbf{d}_{\phi}^{t})\mathbf{E} = -2^{-1}\bar{Y}\eta ael\mathbf{\Delta}_{16}\mathbf{E}$, $\mathbf{\Lambda}_{22} = bE(\mathbf{d}_{\tau}\mathbf{d}_{\tau}^{t})b^{t} = b^{2}\mathbf{\Delta}_{22}$,

$\mathbf{\Lambda}_{23} = \eta bcd\bar{Y}E(\mathbf{d}_{\tau}\mathbf{d}_{z}^{t})\mathbf{Z} = \eta bcd\bar{Y}\mathbf{\Delta}_{23}\mathbf{Z}$, $\mathbf{\Lambda}_{24} = \bar{Y}\eta bchE(\mathbf{d}_{\tau}\mathbf{d}_{\omega}^{t})\mathbf{\Psi} = \bar{Y}\eta bch\mathbf{\Delta}_{24}\mathbf{\Psi}$,

$\mathbf{\Lambda}_{25} = -2^{-1}\bar{Y}\eta befE(\mathbf{d}_{\tau}\mathbf{d}_{w}^{t})\mathbf{W} = -2^{-1}\bar{Y}\eta bef\mathbf{\Delta}_{25}\mathbf{W}$,

$\mathbf{\Lambda}_{26} = -2^{-1}\bar{Y}\eta belE(\mathbf{d}_{\tau}\mathbf{d}_{\phi}^{t})\mathbf{E} = -2^{-1}\bar{Y}\eta bel\mathbf{\Delta}_{26}\mathbf{E}$,

$\mathbf{\Lambda}_{33} = (\eta cd\bar{Y})^{2}E(\mathbf{d}_{z}\mathbf{d}_{z}^{t})\mathbf{Z}^{2} = (\eta cd\bar{Y})^{2}\mathbf{\Delta}_{33}\mathbf{Z}^{2}$,

$\mathbf{\Lambda}_{34} = (\bar{Y}\eta c)^{2}dh\mathbf{Z}E(\mathbf{d}_{z}\mathbf{d}_{\omega}^{t})\mathbf{\Psi} = (\bar{Y}\eta c)^{2}dh\mathbf{Z}\mathbf{\Delta}_{34}\mathbf{\Psi}$,

$\mathbf{\Lambda}_{35} = -2^{-1}\eta^{2}\bar{Y}^{2}cdef\mathbf{Z}E(\mathbf{d}_{z}\mathbf{d}_{w}^{t})\mathbf{W} = -2^{-1}\eta^{2}\bar{Y}^{2}cdef\mathbf{Z}\mathbf{\Delta}_{35}\mathbf{W}$,

$\mathbf{\Lambda}_{36} = -2^{-1}\eta^{2}\bar{Y}^{2}cdel\,\mathbf{Z}E(\mathbf{d}_{z}\mathbf{d}_{\phi}^{t})\mathbf{E} = -2^{-1}\eta^{2}\bar{Y}^{2}cdel\,\mathbf{Z}\mathbf{\Delta}_{36}\mathbf{E}$,

$\mathbf{\Lambda}_{44} = (\bar{Y}\eta ch)^{2}E(\mathbf{d}_{\omega}\mathbf{d}_{\omega}^{t})\mathbf{\Psi}^{2} = (\bar{Y}\eta ch)^{2}\mathbf{\Delta}_{44}\mathbf{\Psi}^{2}$,

$\mathbf{\Lambda}_{45} = -2^{-1}\eta^{2}\bar{Y}^{2}chef\;\mathbf{\Psi}E(\mathbf{d}_{\omega}\mathbf{d}_{w}^{t})\mathbf{W} = -2^{-1}\eta^{2}\bar{Y}^{2}chef\mathbf{\Psi}\mathbf{\Delta}_{45}\mathbf{W}$,

$\mathbf{\Lambda}_{46} = -2^{-1}\bar{Y}^{2}\eta^{2}ehcl\mathbf{\Psi}E(\mathbf{d}_{\omega}\mathbf{d}_{\phi}^{t})\mathrm{E} = -2^{-1}\bar{Y}^{2}\eta^{2}ehcl\mathbf{\Psi}\mathbf{\Delta}_{46}\mathbf{E}$,

$\mathbf{\Lambda}_{55} = \left(2^{-1}\bar{Y}\eta ef\right)^{2}E(\mathbf{d}_{w}\mathbf{d}_{w}^{t})\mathbf{W}^{2} = \left(2^{-1}\bar{Y}ef\right)^{2}\mathbf{\Delta}_{55}\mathbf{W}^{2}$,

$\mathbf{\Lambda}_{56} = 4^{-1}\bar{Y}^{2}\eta^{2}e^{2}\,fl\mathbf{W}E(\mathbf{d}_{w}\mathbf{d}_{\phi}^{t})\mathbf{E} = 4^{-1}\bar{Y}^{2}e^{2}\,fl\mathbf{W}\mathbf{\Delta}_{56}\mathbf{E}$,

$\mathbf{\Lambda}_{66} = \left(2^{-1}\bar{Y}\eta efl\right)^{2}E(\mathbf{d}_{\phi}\mathbf{d}_{\phi}^{t})\mathbf{E}^{2} = \left(2^{-1}\bar{Y}\eta efl\right)^{2}\mathbf{\Delta}_{66}\mathbf{E}^{2}$,

Now, (2.9) can be written as

$$\mathbf{h} = \eta \boldsymbol{\Lambda}^{-1} \boldsymbol{\Omega} . \tag{2.11}$$

From (2.8)

$$MSE(t_{mix}) = E\left\{\eta \bar{e}_{y_{(2)}} + (\eta - 1)\bar{Y} - \mathbf{h}^{t}\mathbf{H}\right\}\left\{\eta \bar{e}_{y_{(2)}} + (\eta - 1)\bar{Y} - \mathbf{h}^{t}\mathbf{H}\right\}$$

or

$$MSE(t_{mix}) = E\left\{\eta \bar{e}_{y(2)} + (\eta - 1)\bar{Y}\right\}\left\{\eta \bar{e}_{y(2)} + (\eta - 1)\bar{Y} - \mathbf{h}^{t}\mathbf{H}\right\}$$

or

$$MSE(t_{mix}) = \eta^2 E\left(\bar{e}_{y(2)}\right)^2 + \left\{(\eta - 1)\bar{Y}\right\}^2 - \eta \mathbf{h}^{t} E\left(\mathbf{H}\bar{e}_{y(2)}\right)$$

or

$$MSE(t_{mix}) = \eta^2 \lambda_2 S_y^2 + (\eta - 1)^2 \bar{Y}^2 - \eta \mathbf{h}^{t} \boldsymbol{\Omega} .$$

By using (2.11), we have

$$MSE(t_{mix}) = \bar{Y}^2\left(\eta^2 - 2\eta + 1\right) + \eta^2\left\{\lambda_2 S_y^2 - \boldsymbol{\Omega}^{t}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Omega}\right\} \tag{2.12}$$

or

$$MSE(t_{mix}) = \left(\eta^2 - 2\eta + 1\right)\bar{Y}^2 + \eta^2 \Gamma ,$$

where $\Gamma = \lambda_2 S_y^2 - \boldsymbol{\Omega}^{t}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Omega}$.

Differentiating MSE w.r.t η and equating to zero

$$2\eta\bar{Y}^2 - 2\bar{Y}^2 + 2\eta\Gamma = 0 ,$$

where $\eta\left(\bar{Y}^2 + \Gamma\right) = \bar{Y}^2$.

$$\eta_{opt} = \bar{Y}^2\left(\bar{Y}^2 + \Gamma\right)^{-1} = \left(1 + \Gamma\bar{Y}^{-2}\right)^{-1} .$$

Then, the minimum *MSE* of the general class is

$$MSE(t_{mix}) = \left[\left\{\left(1 + \bar{Y}^{-2}\Gamma\right)^{-1} - 1\right\}\bar{Y}\right]^2 + \left(1 + \bar{Y}^{-2}\Gamma\right)^{-2}\Gamma . \tag{2.13}$$

Remark 1. The general class of Ahmad et al. (2012) is a member of our proposed class after substituting $b = h = l = 0$ and $\eta = 1$ in (2.1).

As the proposed class is general in nature, special cases of the proposed class (2.1) may be deduced under the assumption $c + e = 1$ using different values of generalizing constants. The special cases with their expressions of bias and MSE's are given in the Remarks 2 and 3. Further, special cases for two and three components are given in Tables 1 and 2 respectively.

Remark 2. *(using two components of generalized class)*

We can obtain a regression (qualitative)-cum-ratio (quantitative) estimator by substituting $\eta = b = c = d = 1, \quad a = e = f = h = l = 0$ in (2.1), i.e.

$$t_{mix(23)} = \left[\bar{y}_2 + \sum_{i'=1}^{q_2} \alpha_{2i'} \left(\tau^*_{(1)i'} - \tau_{(2)i'} \right) \right] \left[\prod_{j=1}^{q_3} \left(\frac{\bar{Z}^*_{(1)j}}{\bar{Z}_{(2)j}} \right)^{\alpha_{3j}} \right] \tag{2.14}$$

The bias of (2.14) can be obtained by substituting $\eta = b = c = d = 1$, $a = e = f = h = l = 0$ in (2.7) as

$$Bias(t_{mix(23)}) = \boldsymbol{\alpha}_2^{t} \boldsymbol{\Delta}_{23} \mathbf{Z} \boldsymbol{\alpha}_3 + 2^{-1} \bar{Y} \boldsymbol{\alpha}_3^{2^t} \mathbf{Z}^2 \left(\lambda_4 \mathbf{S}_z^2 + \theta \mathbf{S}_{(2)z}^2 \right)$$
$$+2^{-1} \bar{Y} \boldsymbol{\alpha}_3^{t} \mathbf{Z}^2 \left(\lambda_3 \mathbf{S}_z^2 + \theta \mathbf{S}_{(2)z}^2 \right) - \boldsymbol{\alpha}_3^{t} \mathbf{Z} \boldsymbol{\delta}_3 .$$

The optimum values are

$\boldsymbol{\alpha}_2 = \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\delta}_2 + \bar{Y}^2 Z^2 \left(\boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\Delta}_{23} \mathbf{M}^{-1} \boldsymbol{\Delta}_{32} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\delta}_2 - \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\Delta}_{23} \mathbf{M}^{-1} \boldsymbol{\delta}_3 \right)$ and

$\boldsymbol{\alpha}_3 = -\bar{Y} \left(\mathbf{M}^{-1} \mathbf{Z} \boldsymbol{\Delta}_{32} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\delta}_2 - \mathbf{M}^{-1} \mathbf{Z} \boldsymbol{\delta}_3 \right)$.

Substituting $\eta = b = c = d = 1$, $a = e = f = h = l = 0$ in (2.12) and using (2.10) and Result-1 of the Appendix, the mean square error of (2.14) can be obtained as

$$MSE(t_{mix(23)}) = \lambda_2 S_y^2 - \left\{ \boldsymbol{\delta}_2^{t} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\delta}_2 + \bar{Y}^2 \left(\boldsymbol{\delta}_2^{t} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\Delta}_{23} \mathbf{Z}^2 \mathbf{M}^{-1} \boldsymbol{\Delta}_{32} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\delta}_2 - \boldsymbol{\delta}_3^{t} \mathbf{Z}^2 \mathbf{M}^{-1} \boldsymbol{\Delta}_{32} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\delta}_2 \right) \right\}$$
$$+\bar{Y}^2 \left(\boldsymbol{\delta}_2^{t} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\Delta}_{23} \mathbf{M}^{-1} \mathbf{Z} - \boldsymbol{\delta}_3^{t} \mathbf{M}^{-1} \mathbf{Z} \right) \mathbf{Z} \boldsymbol{\delta}_3 ,$$

where $\mathbf{M}^{-1} = \bar{Y}^{-1} \mathbf{R}^{-1}$ and $\mathbf{R}^{-1} = \bar{Y}^{-2} \left(\boldsymbol{\Delta}_{33} \mathbf{Z}^2 - \boldsymbol{\Delta}_{32} \boldsymbol{\Delta}_{22}^{-1} \boldsymbol{\Delta}_{23} \mathbf{Z}^2 \right)^{-1}$.

Remark 3. *(using three components of generalized class)*

A regression-cum-ratio estimator using a mixture of auxiliary variables can be obtained by substituting $\eta = a = b = c = d = 1$, $h = e = f = l = 0$ in generalized class (2.1) and we get

$$t_{mix(123)} = \left[\bar{y}_2 + \sum_{i=1}^{q_1} \alpha_{1i} \left(\bar{x}^*_{(1)i} - \bar{x}_{(2)i} \right) + \sum_{i'=1}^{q_2} \alpha_{2i'} \left(\tau^*_{(1)i'} - \tau_{(2)i'} \right) \right] \left[\prod_{j=1}^{q_3} \left(\frac{\bar{Z}^*_{(1)j}}{\bar{Z}_{(2)j}} \right)^{\alpha_{3j}} \right] . \tag{2.15}$$

The bias of (2.15) can be obtained by substituting $\eta = a = b = c = d = 1$ and $h = e = f = l = 0$ in (2.7) as

$$Bias(t_{mix(123)}) = \boldsymbol{\alpha}_1^{t} \boldsymbol{\Delta}_{13} \mathbf{Z} \boldsymbol{\alpha}_3 + \boldsymbol{\alpha}_2^{t} \boldsymbol{\Delta}_{23} \mathbf{Z} \boldsymbol{\alpha}_3 + 2^{-1} \bar{Y} \boldsymbol{\alpha}_3^{2^t} \mathbf{Z}^2 \left(\lambda_4 \mathbf{S}_z^2 + \theta \mathbf{S}_{(2)z}^2 \right)$$
$$+2^{-1} \bar{Y} \boldsymbol{\alpha}_3^{t} \mathbf{Z}^2 \left(\lambda_3 \mathbf{S}_z^2 + \theta \mathbf{S}_{(2)z}^2 \right) - \boldsymbol{\alpha}_3^{t} \mathbf{Z} \boldsymbol{\delta}_3 .$$

where $\boldsymbol{\alpha}_1^{t}$, $\boldsymbol{\alpha}_2^{t}$ and $\boldsymbol{\alpha}_3^{t}$ are vectors of unknown constants of the vector $\mathbf{h}^{t}$.

Let $\mathbf{h}^{t} = \left[\boldsymbol{\alpha}_1^{t} \quad \boldsymbol{\alpha}_2^{t} \quad \boldsymbol{\alpha}_3^{t} \right]_{1 \times m_1}$, where $m_1 = q_1 + q_2 + q_3$.

The optimum value of $\mathbf{h}^{t}$ for which MSE of $t_{mix(123)}$ will be the minimum can be written directly from (2.11) as:

$$\mathbf{h}_{\mathbf{m_1} \times \mathbf{1}} = \eta \boldsymbol{\Lambda}^{-1}_{\mathbf{m_1} \times \mathbf{m_1}} \boldsymbol{\Omega}_{\mathbf{m_1} \times \mathbf{1}} .$$

The mean square error of (2.15) can be obtained by substituting $\eta = a = b = c = d = 1$ and $h = e = f = l = 0$ in (2.13) as:

$$MSE(t_{mix(m_1)}) = \left[\left\{ \left(1+\overline{Y}^{-2}\Gamma_{m_1}\right)^{-1} - 1 \right\}^2 \overline{Y}^2 \right] + \left(1+\overline{Y}^{-2}\Gamma_{m_1}\right)^{-2} \Gamma_{m_1},$$

where $\Gamma_{m_1} = \lambda_2 S_y^2 - \mathbf{\Omega_{1\times m_1}}\, \mathbf{\Lambda^{-1}}_{\mathbf{m_1\times m_1}}\, \mathbf{\Omega_{m_1\times 1}}$;

$$\mathbf{\Lambda_{m_1\times m_1}} = \begin{bmatrix} \mathbf{\Delta}_{11} & \mathbf{\Delta}_{12} & \overline{Y}\mathbf{\Delta}_{13}\mathbf{Z} \\ \mathbf{\Delta}_{21} & \mathbf{\Delta}_{22} & \overline{Y}\mathbf{\Delta}_{23}\mathbf{Z} \\ \overline{Y}\mathbf{Z}\mathbf{\Delta}_{31} & \overline{Y}\mathbf{Z}\mathbf{\Delta}_{32} & \overline{Y}^2\mathbf{\Delta}_{33}\mathbf{Z}^2 \end{bmatrix} \text{ and } \mathbf{\Omega_{m_1\times 1}} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \boldsymbol{\delta}_2 \\ \overline{Y}\mathbf{Z}\boldsymbol{\delta}_3 \end{bmatrix}.$$

The inverse of $\mathbf{\Lambda_{m_1\times m_1}}$ i.e. $\mathbf{\Lambda^{-1}}_{\mathbf{m_1\times m_1}}$ can be obtained using the Result-1 given in the Appendix.

The proposed general class comprises six components, three pairs are based on regression, ratio and exponential forms and each form utilizes categorical and continuous auxiliary variables separately. Moreover, it is assumed that $c + e = 1$. Following the Remarks 2 and 3, special cases consist of four and five components and even the single case based on all the components can be deduced using suitable values of generalizing constants. The special cases in which either c or e are involved, need no additional work, but the cases that involve both c and e need one extra step in finding the optimum value of either c or e. After finding this additional optimum value, the bias, existing optimum values and means square errors will be changed accordingly for these particular special cases. We have not included these cases in the article due to the limitation of length of the article. The special cases for two and three are given in the following tables.

Table1. Special cases of generalized class using two components

Estimator	(a,b,c,d,e,f,h,l,η)
$t_{mix(23)} = \left[\overline{y}_2 + \sum_{i'=1}^{q_2} \alpha_{2i'} \left(\tau^*_{(1)i'} - \tau_{(2)i'} \right) \right] \left[\prod_{j=1}^{q_3} \left(\frac{\overline{z}^*_{(1)j}}{\overline{z}_{(2)j}} \right)^{\alpha_{3j}} \right]$	(0,1,1,1,0,0,0,0,1)
$t_{mix(16)} = \left[\overline{y}_2 + \sum_{i=1}^{q_1} \alpha_{1i} \left(\overline{x}^*_{(1)i} - \overline{x}_{(2)i} \right) \right] \left[\exp\left\{ \sum_{k'=1}^{q_6} \alpha_{6k'} \left(\frac{\phi_{(2)k'} - \phi^*_{(1)k'}}{\phi_{(2)k'} + \phi^*_{(1)k'}} \right) \right\} \right]$	(1,0,0,0,1,0,0,1,1)
$t_{mix(25)} = \left[\overline{y}_2 + \sum_{i'=1}^{q_2} \alpha_{2i'} \left(\tau^*_{(1)i'} - \tau_{(2)i'} \right) \right] \left[\exp\left\{ \sum_{k=1}^{q_5} \alpha_{5k} \left(\frac{\overline{w}_{(2)k} - \overline{w}^*_{(1)k}}{\overline{w}_{(2)k} + \overline{w}^*_{(1)k}} \right) \right\} \right]$	(0,1,0,0,1,1,0,0,1)
$t_{mix(15)} = \left[\overline{y}_2 + \sum_{i=1}^{q_1} \alpha_{1i} \left(\overline{x}^*_{(1)i} - \overline{x}_{(2)i} \right) \right] \left[\exp\left\{ \sum_{k=1}^{q_5} \alpha_{5k} \left(\frac{\overline{w}_{(2)k} - \overline{w}^*_{(1)k}}{\overline{w}_{(2)k} + \overline{w}^*_{(1)k}} \right) \right\} \right]$	(1,0,0,0,1,1,0,0,1)

Table1. Special cases of generalized class using two components (cont.)

Estimator	(a,b,c,d,e,f,h,l,η)
$t_{mix(26)}=\left[\bar{y}_2+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\exp\left\{\sum_{k'=1}^{q_6}\alpha_{6k'}\left(\frac{\phi_{(2)k'}-\phi^*_{(1)k'}}{\phi_{(2)k'}+\phi^*_{(1)k'}}\right)\right\}\right]$	(0,1,0,0,1,0,0,1,1)
$t_{mix(14)}=\left[\bar{y}_2+\sum_{i=1}^{q_1}\alpha_{1i}\left(\bar{x}^*_{(1)i}-\bar{x}_{(2)i}\right)\right]\left[\prod_{j'=1}^{q_4}\left(\frac{\omega^*_{(1)j'}}{\omega_{(2)j'}}\right)^{\alpha_{4j'}}\right]$	(1,0,1,0,0,0,1,0,1)
$t_{mix(12)}=\left[\bar{y}_2+\sum_{i=1}^{q_1}\alpha_{1i}\left(\bar{x}^*_{(1)i}-\bar{x}_{(2)i}\right)+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]$	(1,1,1,0,0,0,0,0,1)
$t_{mix(34)}=\bar{y}_2\left[\prod_{j=1}^{q_3}\left(\frac{\bar{z}^*_{(1)j}}{\bar{z}_{(2)j}}\right)^{\alpha_{3j}}\prod_{j'=1}^{q_4}\left(\frac{\omega^*_{(1)j'}}{\omega_{(2)j'}}\right)^{\alpha_{4j'}}\right]$	(0,0,1,1,0,0,1,0,1)
$t_{mix(24)}=\left[\bar{y}_2+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\prod_{j'=1}^{q_4}\left(\frac{\omega^*_{(1)j'}}{\omega_{(2)j'}}\right)^{\alpha_{4j'}}\right]$	(0,1,1,0,0,0,1,0,1)
$t_{mix(13)}=\left[\bar{y}_2+\sum_{i=1}^{q_1}\alpha_{1i}\left(\bar{x}^*_{(1)i}-\bar{x}_{(2)i}\right)\right]\left[\prod_{j=1}^{q_3}\left(\frac{\bar{z}^*_{(1)j}}{\bar{z}_{(2)j}}\right)^{\alpha_{3j}}\right]$	(1,0,1,1,0,0,0,0,1)
$t_{mix(56)}=\bar{y}_2\exp\left\{\sum_{k=1}^{q_5}\alpha_{5k}\left(\frac{\bar{w}_{(2)k}-\bar{w}^*_{(1)k}}{\bar{w}_{(2)k}+\bar{w}^*_{(1)k}}\right)+\sum_{k'=1}^{q_6}\alpha_{6k'}\left(\frac{\phi_{(2)k'}-\phi^*_{(1)k'}}{\phi_{(2)k'}+\phi^*_{(1)k'}}\right)\right\}$	(0,0,1,0,1,1,0,1,1)

Table 2. Special cases of generalized class using three components

Estimator	(a,b,c,d,e,f,h,l,η)
$t_{mix(123)}=\left[\bar{y}_2+\sum_{i=1}^{q_1}\alpha_{1i}\left(\bar{x}^*_{(1)i}-\bar{x}_{(2)i}\right)+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\prod_{j=1}^{q_3}\left(\frac{\bar{z}^*_{(1)j}}{\bar{z}_{(2)j}}\right)^{\alpha_{3j}}\right]$	(1,1,1,1,0,0,0,0,1)
$t_{mix(134)}=\left[\bar{y}_2+\sum_{i=1}^{q_1}\alpha_{1i}\left(\bar{x}^*_{(1)i}-\bar{x}_{(2)i}\right)\right]\left[\prod_{j=1}^{q_3}\left(\frac{\bar{z}^*_{(1)j}}{\bar{z}_{(2)j}}\right)^{\alpha_{3j}}\prod_{j'=1}^{q_4}\left(\frac{\omega^*_{(1)j'}}{\omega_{(2)j'}}\right)^{\alpha_{4j'}}\right]$	(1,0,1,1,0,0,1,0,1)

Table 2. Special cases of generalized class using three components (cont.)

Estimator	(a,b,c,d,e,f,h,l,η)
$t_{mix(234)}=\left[\bar{y}_2+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\prod_{j=1}^{q_3}\left(\frac{\bar{z}^*_{(1)j}}{\bar{z}_{(2)j}}\right)^{\alpha_{3j}}\prod_{j'=1}^{q_4}\left(\frac{\omega^*_{(1)j'}}{\omega_{(2)j'}}\right)^{\alpha_{4j'}}\right]$	(0,1,1,1,0,0,1,0,1)
$t_{mix(125)}=\left[\bar{y}_2+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\exp\left\{\sum_{k=1}^{q_5}\alpha_{5k}\left(\frac{\bar{w}_{(2)k}-\bar{w}^*_{(1)k}}{\bar{w}_{(2)k}+\bar{w}^*_{(1)k}}\right)\right\}\right]$	(1,1,0,0,1,1,0,0,1)
$t_{mix(126)}=\left[\bar{y}_2+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\exp\left\{\sum_{k'=1}^{q_6}l\alpha_{6k'}\left(\frac{\phi_{(2)k'}-\phi^*_{(1)k'}}{\phi_{(2)k'}+\phi^*_{(1)k'}}\right)\right\}\right]$	(1,1,0,0,0,1,0,1,1)
$t_{mix(256)}=\left[\bar{y}_2+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\exp\left\{\sum_{k=1}^{q_5}\alpha_{5k}\left(\frac{\bar{w}_{(2)k}-\bar{w}^*_{(1)k}}{\bar{w}_{(2)k}+\bar{w}^*_{(1)k}}\right)+\sum_{k'=1}^{q_6}\alpha_{6k'}\left(\frac{\phi_{(2)k'}-\phi^*_{(1)k'}}{\phi_{(2)k'}+\phi^*_{(1)k'}}\right)\right\}\right]$	(0,1,0,0,1,1,0,1,1)
$t_{mix(156)}=\left[\bar{y}_2+\sum_{i=1}^{q_1}\alpha_{1i}\left(\bar{x}^*_{(1)i}-\bar{x}_{(2)i}\right)\right]\left[\exp\left\{\sum_{k=1}^{q_5}\alpha_{5k}\left(\frac{\bar{w}_{(2)k}-\bar{w}^*_{(1)k}}{\bar{w}_{(2)k}+\bar{w}^*_{(1)k}}\right)+\sum_{k'=1}^{q_6}\alpha_{6k'}\left(\frac{\phi_{(2)k'}-\phi^*_{(1)k'}}{\phi_{(2)k'}+\phi^*_{(1)k'}}\right)\right\}\right]$	(1,0,0,0,1,1,0,1,1)
$t_{mix(124)}=\left[\bar{y}_2+\sum_{i=1}^{q_1}\alpha_{1i}\left(\bar{x}^*_{(1)i}-\bar{x}_{(2)i}\right)+\sum_{i'=1}^{q_2}\alpha_{2i'}\left(\tau^*_{(1)i'}-\tau_{(2)i'}\right)\right]\left[\prod_{j'=1}^{q_4}\left(\frac{\omega^*_{(1)j'}}{\omega_{(2)j'}}\right)^{h\alpha_{4j'}}\right]$	(1,1,1,0,0,0,1,0,1)

3. Empirical study

In this section we have empirically compared special cases that are discussed in sections 4.1 and 4.2. For this comparison, the data of Census report of district Jhang (1998), Pakistan (see Ahmad et al. (2009)) is used. The population size is 368 (N). From N , 276(N_1) are considered as respondent group of population and the remaining 92(N_2) are non-respondent group. A sample of 160(n_1) is selected at the first phase sample and from first phase, a sample of 90(n_2) is selected as the second phase sample. From the second phase sample, a sub-sample of 10(r) is selected as the re-contacted sample and it is assumed that there is full response from this sample.

The data set, which is considered for the empirical study, consists of three quantitative and three qualitative auxiliary variables along with one response variable. The variables description is given in Table 3. The variances and co-variances and the mean of all variables for both complete and non-respondent

populations are given in Table 4 and Table 5 respectively. The bias and MSE's of all special cases given in Table 1 and Table 2 are given in Table 6 and Table 7 respectively.

Table 3. Description of variables *(each variable is taken from rural locality)*

Description of Variables	
Y	Literacy ratio
x	Household size
z	Population of both sexes
w	Household characteristics
τ	Male above and below average education
ω	Female above and below average education
ϕ	Persons below and above average age

Table 4. Variance co-variances and mean of complete population N = 368

		Variance Co-variance Matrix						
Variable	**Mean**	y	x	z	w	τ	ω	ϕ
y	29.831	62.860	0.548	5780.0000	593.123	1.936	0.5630	1.440
x	6.372	0.548	0.266	138.7860	-7.139	0.015	0.0020	0.0110
z	5901.46	5780.0	138.786	26660000.0	1262000.0	711.523	1191.0	1173.0
w	897.71	593.123	-7.139	1262000.0	211500.0	113.702	162.718	157.043
τ	0.35	1.936	0.015	711.5230	113.702	0.227	0.1060	0.1560
ω	0.42	0.563	0.002	1191.0	162.718	0.106	0.2440	0.1740
ϕ	0.43	1.440	0.011	1173.0	157.043	0.156	0.1740	0.2450

Table 5. Variances co-variances and mean of non-respondent population $N_2 = 92$

	Variance Co-variance Matrix							
Variable	y	x	z	w	τ	ω	ϕ	**Mean**
y	40.0700	0.1680	7508.0	1176.00	1.5780	1.2480	1.5370	25.1467
x	0.1680	0.1270	-17.55	-21.4730	0.0030	-0.0190	-0.0080	6.2130
z	7508.00	-17.550	6239000.0	999300.0	668.870	1001.0	991.744	6164.5761
w	1176.00	-21.4730	999300.0	163700.0	105.675	163.453	160.103	994.3152
τ	1.5780	0.0030	668.87	105.6750	0.1950	0.0870	0.1360	0.2609
ω	1.2480	-0.0190	1001.0	163.4530	0.0870	0.2510	0.1750	0.5435
ϕ	1.5370	-0.0080	991.7440	160.1030	0.1360	0.1750	0.2430	0.4022

From Table 6, biases of estimators show that some estimators are overestimating and some are underestimating the population mean of study

variable except the regression estimator that is unbiased and similar information is described from Table 7 for estimators based on biases.

From Table 6, the ranked absolute biases show that the estimator $t_{mix(13)}$ has larger bias as compared to others whereas $t_{mix(26)}$ has smaller bias. The estimators at rank 2, 3, 4 and 5 have a very small amount of bias whereas the remaining ones have large amount of bias. Considering the ranks of MSE, the estimator $t_{mix(25)}$ is more efficient then all others whereas $t_{mix(34)}$ is the least efficient. However, the differences in MSE's for all estimators are very small but there is a lot of variation in biases. Considering the trade-off between biases and MSE's, the sum of ranks of bias and MSE suggests that the biased estimator $t_{mix(26)}$ is useful for practical situations where only qualitative variables are considered, $t_{mix(25)}$ is suitable when there is a mixture of auxiliary variables and $t_{mix(15)}$ can be used for only quantitative auxiliary variables.

Table 6. Bias and MSE of members of generalized class for two factors

Estimators	Bias	Absolute Bias	Ranked Absolute Bias	MSE	Ranked MSE	Sum of Ranks	Ranks of Sum
$t_{mix(23)}$	-8.31E+03	8310	09	15.3249	5	14	7
$t_{mix(25)}$	-1.30E+03	1300	6	15.3217	1	7	3
$t_{mix(15)}$	1.41E+03	1410	7	15.3601	8	15	8.5
$t_{mix(26)}$	1.74E-04	0.000174	2	15.3248	4	6	2
$t_{mix(14)}$	5.39E-04	0.000539	3	15.3620	10	13	6
$t_{mix(12)}$	0	0	1	15.3222	2	3	1
$t_{mix(34)}$	6.64E+04	66400	10	15.3693	11	21	11
$t_{mix(24)}$	1.35E-03	0.00135	5	15.3246	3	8	4
$t_{mix(13)}$	8.31E+04	83100	11	15.3608	9	20	10
$t_{mix(56)}$	-2.17E+03	2170	8	15.3446	7	15	8.5

In the case of estimators based on three components, from Table 7, the ranked absolute biases show that $t_{mix(124)}$ has smaller bias whereas $t_{mix(134)}$ is highly biased. $t_{mix(126)}$ has also a very small amount of bias. Based on ranks of MSE, $t_{mix(256)}$ is more efficient than all others whereas $t_{mix(134)}$ is the least efficient. Considering the trade-off between bias and MSE's, the sum of ranks of bias and MSE suggests

that $t_{mix(124)}$ is useful for practical situations when there is a mixture of auxiliary variables.

Table 7. Bias and MSE of members of generalized class using three components

Estimators	Bias	Absolute Bias	Ranked Absolute Bias	MSE	Ranked MSE	Sum of Ranks	Ranks of Sum
$t_{mix(123)}$	-8.31E+03	8310	6	15.3222	5	11	5
$t_{mix(134)}$	1.34E+08	134000000	8	15.4720	8	16	8
$t_{mix(234)}$	-2.82E+04	28200	7	15.3243	6	13	7
$t_{mix(125)}$	-1.17E+03	1170	3	15.3196	2	5	2.5
$t_{mix(126)}$	3.01E-04	0.000301	2	15.3220	4	6	4
$t_{mix(256)}$	-2.33E+03	2330	4	15.3181	1	5	2.5
$t_{mix(156)}$	-5.72E+03	5720	5	15.3416	7	12	6
$t_{mix(124)}$	1.81E-04	0.000181	1	15.3218	3	4	1

As the suggested class consists of regression, ratio and exponential components, it is obvious that the regression component contributes in terms of reduction of MSE and ratio and exponential components will increase bias and decrease MSE. This statement can be verified from Table 6 and 7. For example, from Table 6 $t_{mix(15)}$ is a regression-cum-exponential estimator with one exponential component having bias 1410 and MSE 15.3601, and by adding another ratio component in this estimator we obtain the regression-cum-exponential $t_{mix(156)}$ (given in Table 7) with bias 5720 and MSE 15.3416 as the result. This type of change can be observed for other estimators with this property.

What is specific to this empirical study is that the qualitative auxiliary variables are performing better than the continuous auxiliary variables. For example, $t_{mix(125)}$ [regression-cum-exponential (quantitative)] has bias 1170 and MSE 15.3196 whereas $t_{mix(126)}$ [regression-cum-exponential (qualitative)] has bias 0.000301 and MSE 15.3220. Similar information can be observed considering other such pairs of estimators.

Summarizing the discussion on both tables, the three-component estimator $t_{mix(124)}$ is better than all others while considering bias and MSE simultaneously. This estimator comprises two regression components of quantitative and qualitative auxiliary variables and one ratio component of qualitative auxiliary variables.

4. Conclusions

In this paper, a general class of regression-cum-ratio-exponential estimators is developed for two-phase sampling in the presence of non-responses at the first phase. Both quantitative and qualitative auxiliary variables are used in the construction of the class to increase the efficiency of the class as well as its members. The general expression of bias and mean square error is also derived. As the proposed class is general in nature, some suitable special cases are deduced along with their bias and mean square errors. On the basis of the empirical study it is concluded that both types of auxiliary variables can play a role in reducing the bias and the mean square error of an estimator. The bias and mean square error can be reduced by increasing the number of auxiliary variables. An increase in ratio or exponential components increases the bias. Our findings show that an estimator based on three components performs better then all others. This estimator comprises two regression components of quantitative and qualitative auxiliary variables and one ratio component of qualitative auxiliary variables.

This paper also fills the gap in the literature as it attempts to estimate the finite population mean using both qualitative and quantitative multi-auxiliary variables in the presence of non-response at the first phase under two-phase sampling. It can also provide an opportunity to the applied survey statisticians if they consider estimation of finite population mean using several qualitative and quantitative auxiliary variables.

REFERENCES

AHMAD, Z., HANIF, M., (2010). Generalized Multi-Phase Multivariate Regression Estimators for Partial Information Case using Multi-Auxiliary Variables. World Applied Sciences Journal, 10(3): 370–379.

AHMAD, Z., HANIF, M., AHMAD, M., (2009a). Generalized Multivariate Ratio Estimators for Multi-Phase Sampling using Multi-Auxiliary Variables. Pakistan Journal of Statistics, 25(4), 615–629.

AHMAD, Z., HANIF, M., AHMAD, M., (2009b). Generalized Regression-Cum-Ratio Estimators for Two-Phase Sampling using Multi-Auxiliary Variables, Pakistan Journal of Statistics, 25(2), 1–14.

AHMAD, Z., HANIF, M., AHMAD, M., (2010). Generalized Multi-Phase Multivariate Ratio Estimators for Partial Information Case using Multi-Auxiliary Variables. Communications of the Korean Statistical Society, 17(5), 625–637.

AHMAD, Z., MAQSOOD, I., HANIF, M., (2013). Generalized estimator of population mean for two phase sampling using multi-auxiliary variables in the presence of non-response at first phase for no information case. Pakistan Journal of Statistics, 29(2), 155–180.

AHMAD, Z., ZAFAR, I., HANIF, M., (2013a). Generalized estimator of population mean for two phase sampling using multi-auxiliary variables in the presence of non-response at first phase for no information case, Journal of Mathematical Sciences and Applied E-Notes. 1(2), 238–256.

AHMAD, Z., ZAFAR, I., BANO, Z., (2013b). Generalized Class of Mean Estimators for Two-Phase Sampling in The Presence of Nonresponse. J. Japan Statist. Soc., 43(2), 163–185.

BOWLEY, A. L., (1926). Measurements of precision attained in sampling. Bull. Inst. Inte. Statist., 22, 1–62.

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of no response in sample surveys. American Statistical Association, 41, 517–529.

HANIF, M., HAQ, I., SHAHBAZ, M. Q., (2010). Ratio Estimators using Multiple Auxiliary Attributes, World Applied Sciences Journal, 8(1), 133–136.

HIDIROGLOU, M. A., SARNDAL, C. E., (1998). Use of auxiliary information for two-phase sampling. Survey Methodology, 24(1), 11–20.

JHAJJ, H. S., SHARMA, M. K., GROVER, L. K., (2006). Dual of ratio estimators of finite population mean obtained on using linear transformation to auxiliary variables. Journal of Japan Statistical Society, 36(1), 107–119.

KHARE, B. B., SRIVASTAVA, S., (1993). Estimation of Population Mean using Auxiliary Character in the Presence of Non Response. National. Acad. Sci. Letters. (India), 16(3), 111–114.

KHARE, B. B., SRIVASTAVA, S., (1995). Study of Conventional and Alternative Two Phase Sampling Ratio, Product and Regression Estimators in the Presence of Non-Response. Proc. Nat. Acad. Sci. (India), 65(A), 195–203.

MOHANTY, S., (1967). Combination of regression and ratio estimate, J. Ind. Statist. Assoc., 5, 16–19.

NEYMAN, J., (1934). On the two different aspects of representative method: The method of stratified sampling and the method of purposive selection. J. Roy. Statist. Soc., 97, 558–606.

NEYMAN, J. (1938). Contribution to the theory of sampling human populations. J. Amer. Statist. Assoc., 33, 101–116.

NAIK, V. D., GUPTA, P. C., (1996). A note on estimation of mean with known population of an auxiliary character. J. Ind. Soc. Agri. Statist., 48(2), 151–158.

RAO, J. N. K., (1973). On double sampling for stratification and analytic surveys. Biometrika, 6, 125–133.

ROY, D. C., (2003). A regression type estimates in two-phase sampling using two auxiliary variables. Pak. J. Statist., 19(3), 281–290.

SAMIUDDIN, M., HANIF, M., (2007). Estimation of Population Means in Single and Two-Phase Sampling with or without Additional Information. Pakistan Journal of Statistics, 23 (2), 99–118.

SHABBIR, J., GUPTA, S., (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. Pakistan Journal of Statistics, 23(1), 1–9.

SINGH, H. P., KUMAR, S., (2008). Estimation of Mean in Presence of Non Response using Two-Phase Sampling Scheme. Statist. Pap. DOI 10.1007/s00362-008-040-5.

SINGH, R., CHAUHAN, P., SAWAN, N., (2010). Ratio-Product Type Exponential Estimator for Estimating Finite Population Mean Using Information on Auxiliary Attribute. Unpublished manuscript.

TABASUM, R., KHAN, I. A., (2004). Double Sampling for Ratio Estimation with Non-Response. J. Ind. Soc. Agril. Statist. 58(3), 300–306.

WU, C., LUAN, Y., (2003). Optimal Calibration Estimators under Two-Phase Sampling. Journal of Official Statistics, 19(2), 119–131.

APPENDIX

Result 1. Inverse of matrix of matrices:

Let $\mathbf{T}$ be a matrix of matrices of order 4×4,

$$\mathbf{T} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} & \mathbf{T}_{13} & \mathbf{T}_{14} \\ \mathbf{T}_{21} & \mathbf{T}_{22} & \mathbf{T}_{23} & \mathbf{T}_{24} \\ \mathbf{T}_{31} & \mathbf{T}_{32} & \mathbf{T}_{33} & \mathbf{T}_{34} \\ \mathbf{T}_{41} & \mathbf{T}_{42} & \mathbf{T}_{43} & \mathbf{T}_{44} \end{bmatrix}.$$

The inverse of $\mathbf{T}$ is

$$\mathbf{T}^{-1} = \begin{bmatrix} \mathbf{B}_{33} & \mathbf{B}_{31} \\ \mathbf{B}_{13} & \mathbf{B}_{11} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{B}_{33}^{-1} + \left(\mathbf{B}_{33}^{-1}\mathbf{B}_{31}\right)\mathbf{G}_{11}^{-1}\left(\mathbf{B}_{13}\mathbf{B}_{33}^{-1}\right) & -\left(\mathbf{B}_{33}^{-1}\mathbf{B}_{31}\right)\mathbf{G}_{11}^{-1} \\ -\mathbf{G}_{11}^{-1}\left(\mathbf{B}_{13}\mathbf{B}_{33}^{-1}\right) & \mathbf{G}_{11}^{-1} \end{bmatrix},$$

where $\mathbf{G}_{11}^{-1} = \left(\mathbf{B}_{11} - \mathbf{B}_{13}\mathbf{B}_{33}^{-1}\mathbf{B}_{31}\right)^{-1}$,

$$\mathbf{B}_{33} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} & \mathbf{T}_{13} \\ \mathbf{T}_{21} & \mathbf{T}_{22} & \mathbf{T}_{23} \\ \mathbf{T}_{31} & \mathbf{T}_{32} & \mathbf{T}_{33} \end{bmatrix},\ \mathbf{B}_{31} = \begin{bmatrix} \mathbf{T}_{14} \\ \mathbf{T}_{24} \\ \mathbf{T}_{34} \end{bmatrix},\ \mathbf{B}_{13} = \begin{bmatrix} \mathbf{T}_{41} & \mathbf{T}_{42} & \mathbf{T}_{43} \end{bmatrix} \text{ and } \mathbf{B}_{11} = \mathbf{T}_{44}.$$

and

$$\mathbf{B}_{33}^{-1} = \begin{bmatrix} \mathbf{A}_{22} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{11} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}_{22}^{-1} + \left(\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)\mathbf{H}_{11}^{-1}\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\right) & -\left(\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)\mathbf{H}_{11}^{-1} \\ -\mathbf{H}_{11}^{-1}\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\right) & \mathbf{H}_{11}^{-1} \end{bmatrix},$$

where $\mathbf{H}_{11}^{-1} = \left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1}$,

$$\mathbf{A}_{22} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix},\ \mathbf{A}_{21} = \begin{bmatrix} \mathbf{T}_{13} \\ \mathbf{T}_{23} \end{bmatrix},\ \mathbf{A}_{12} = \begin{bmatrix} \mathbf{T}_{31} & \mathbf{T}_{32} \end{bmatrix} \text{ and } \mathbf{A}_{11} = \mathbf{T}_{33}.$$

and

$$\mathbf{A}_{22}^{-1} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{T}_{11}^{-1} + \left(\mathbf{T}_{11}^{-1}\mathbf{T}_{12}\right)\mathbf{R}_{22}^{-1}\left(\mathbf{T}_{21}\mathbf{T}_{11}^{-1}\right) & -\left(\mathbf{T}_{11}^{-1}\mathbf{T}_{12}\right)\mathbf{R}_{22}^{-1} \\ -\mathbf{R}_{22}^{-1}\left(\mathbf{T}_{21}\mathbf{T}_{11}^{-1}\right) & \mathbf{R}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{R}_{22}^{-1} = \left(\mathbf{T}_{22} - \mathbf{T}_{21}\mathbf{T}_{11}^{-1}\mathbf{T}_{12}\right)^{-1}$.

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 525–534

ON CONDITIONAL SIMPLE RANDOM SAMPLE

Janusz L. Wywiał [1]

ABSTRACT

Estimation of the population average in a finite and fixed population on the basis of the conditional simple random sampling design dependent on order statistics of the auxiliary variable is studied. The sampling scheme implementing the sampling design is proposed. The inclusion probabilities are derived. The well known Horvitz-Thompson statistic under the conditional simple random sampling designs is considered as the estimator of population mean. Moreover, it was shown that the Horvitz-Thompson estimator under some particular cases of the conditional simple random sampling design is more accurate than the ordinary mean from the simple random sample.

Key words: simple random sample, conditional sampling design, sampling scheme, inclusion probabilities, auxiliary variable, order statistics.

1. Introduction

The sampling designs dependent on an auxiliary variables are constructed in order to improve accuracy of population parameters estimation. Rao (1985) considered problems of conditional statistical inference in survey sampling. Applications of auxiliary information to construction of the conditional versions of sampling designs were discussed in the literature, for instance by Tillé (1998, 2006). This paper was inspired by Royall and Cumberland (1981) proposition of conditional simple sampling design.

Let U be a fixed population of size N. The observation of a variable under study and an auxiliary variable are identifiable and denoted by y_i and $x_i, i = 1, \ldots, N$, respectively. We assume that $x_i \leq x_{i+1}, i = 1, \ldots, N-1$. Our main purpose is to estimate the population average: $\bar{y} = \frac{1}{N}\sum_{i\in U} y_i$.

Let us consider the sample space $\mathbf{S}$ of the samples s of the fixed effective size $1 < n < N$. The sampling design is denoted by $P(s)$ where $P(s) > 0$ for all $s \in \mathbf{S}$ and $\sum_{s\in\mathbf{S}} P(s) = 1$. As it is well known the simple sampling design is defined as follows:

$$P_0(s) = \binom{N}{n}^{-1} \qquad \text{for all} \quad s \in \mathbf{S}. \tag{1}$$

[1]Department of Statistics, University of Economics in Katowice, 1 Maja 50, 40-287 Katowice, Poland. E-mail:janusz.wywial@ue.katowice.pl.

Royall and Cumberland (1981) considered drawing the simple random sample s until the inequality $|\bar{x}_s - \bar{x}| \leq c$ where $\bar{x} = \frac{1}{N}\sum_{i\in U} x_i$, $c > 0$, is fulfilled. This sampling scheme can be called the conditional simple random sampling. Of course conditions can be stated by means of other inequalities, see e.g. Wywiał (2003) using computer simulation analysis because the inclusion probabilities of the conditional simple random sampling design are not known. Derivation of those probabilities is one of our purposes. In the considered case the condition is defined on the basis of the properties of the order statistics of the auxiliary variable. In the next section, the conditional simple random sampling design is defined and the inclusion probabilities are derived. The sampling scheme described in the third section. In the fourth section the Horvitz-Thompson estimator is considered. Next we can find some general conclusions about the properties of the considered estimation strategy. The proof of the theorems is in the Appendix.

Let $s = \{s_1, i, s_2\}$ where $s_1 = \{i_1, ..., i_{r-1}\}$, $s_2 = \{i_{r+1}, ..., i_n\}$, $i_j < i$ for $j = 1, ..., r$, $i_r = i$ and $i_j > i$ for $j = r+1, ..., n$. Hence, x_i is one of the possible observations of the order statistic $X_{(r)}$ of the rank r $(r = 1, ..., n)$ from the sample s. Let $\mathbf{S}(r,i) = \{s : X_{(r)} = x_i\}$ be the set of all samples whose r -th order statistic of the auxiliary variable is equal to x_i where $r \leq i \leq N - n + r$. Hence, $\bigcup_{i=r}^{N-n+r} \mathbf{S}(r,i) = \mathbf{S}$. The size of the set $\mathbf{S}(r,i)$ is denoted by $g(r,i) = Card(\mathbf{S}(r,i))$ and

$$g(r,i) = \binom{i-1}{r-1}\binom{N-i}{n-r}, \qquad \sum_{i=r}^{N-n+r} g(r,i) = \binom{N}{n}.$$

The probability that the r -th order statistic from simple random sample of an auxiliary variable takes value x_i is as follows (see Wilks (1962), pp. 243-244 or Guenther (1975) or Hogg and Craig (1970)):

$$P\left(X_{(r)} = x_i\right) = \frac{g(r,i)}{\binom{N}{n}}, \quad i = r, ..., N - n + r.$$

$$E\left(X_{(r)}\right) = \sum_{i=r}^{N-n+r} x_i P\left(X_{(r)} = x_i\right) = \frac{1}{\binom{N}{n}} \sum_{i=r}^{N-n+r} x_i g(r,i).$$

The sample quantile of order $\alpha \in (0;1)$ is defined as $Q_{s,\alpha} = X_{(r)}$. The rank r can be determined as follows: $r = [n\alpha] + 1$ where $[.]$ is the integer part of the value $n\alpha$. Hence, $r = 1, 2, ..., n$ and $X_{(r)} = Q_{s,\alpha}$ for $\frac{r-1}{n} \leq \alpha < \frac{r}{n}$. So, it will be more convenient to consider the order statistics than the quantiles.

The conditional (truncated) version of the order statistic distribution is as follows:

$$P\left(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w\right) = \frac{P\left(X_{(r)} = x_i\right)}{P\left(x_u \leq X_{(r)} \leq x_w\right)} = \frac{g(r,i)}{z(r,u,w)} = \\ = P\left(X_{(r)} = x_i | r, u, w\right)$$

where

$$P\left(x_u \leq X_{(r)} \leq x_w\right) = \frac{z(r,u,w)}{\binom{N}{n}},$$

$$z(r,u,w) = \sum_{t=u}^{w} g(r,t). \tag{2}$$

2. Sampling design

On the basis of the previous section we have obtained:

$$P_0\left(s \in \mathbf{S}(r,i)\right) = \sum_{s \in \mathbf{S}(r,i)} P_0(s) = \frac{z(r,u,w)}{\binom{N}{n}} = P\left(x_u \leq X_{(r)} \leq x_w\right).$$

Hence,

$$P_0\left(s|s \in \mathbf{S}(r,i)\right) = \frac{P_0(s)}{P_0\left(s \in \mathbf{S}(r,i)\right)} = \frac{1}{z(r,u,w)} =$$
$$= \frac{P_0(s)}{P\left(X_{(r)} = x_i|r,u,w\right)} = P_0\left(s|x_u \leq X_{(r)} \leq x_w\right) = P_0\left(s|r,u,w\right). \tag{3}$$

Definition 2.1. *The sampling design expressed by the equations (3) and (2) will be called the conditional simple random sampling design.*

So, the introduced sampling design provides such the simple random samples where r-the order $X_{(r)}$ takes value from the interval $[x_u;\ x_u]$ where $u \leq r \leq w$.

The inclusion probability of the first and second orders are defined by the following equation: $\pi_k = \sum_{\{s:k \in s\}} P(s)$ and $\pi_{k,t} = \sum_{\{s:k \in s, t \in s\, k \neq t\}} P(s)$, respectively where $k,t = 1,...,N$. Let us assume that if $x \leq 0$, $\delta(x) = 0$ else $\delta(x) = 1$. Let us note that $\delta(x)\delta(x-1) = \delta(x-1)$.

In the Appendix the following theorem is proved on the basis of Wywial's (2008) results.

Theorem 2.1. *The inclusion probabilities of the first order for the conditional simple random sampling design $P_0(s|r,u,w)$ are as follows: if $k < u$,*

$$\pi_k^{(r)}(u,w) = \frac{\delta(r-1)\delta(w-1)\delta(u-1)}{z_r(u,w)} \sum_{i=u}^{w} \binom{i-2}{r-2}\binom{N-i}{n-r},$$

If $u \le k \le w$,

$$\pi_k^{(r)}(u,w) = \\ = \frac{1}{z_r(u,w)}\left(\delta(n-r)\delta(k-u)\delta(k-1)\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1}+\right. \\ \left.+\binom{k-1}{r-1}\binom{N-k}{n-r}+\delta(r-1)\delta(w-k)\sum_{i=k+1}^{w}\binom{i-2}{r-2}\binom{N-i}{n-r}\right),$$

if $k > w$,

$$\pi_k^{(r)}(u,w) = \frac{\delta(n-r)\delta(N-w)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1},$$

The inclusion probabilities of the second order for the conditional simple random sampling design $P_0(s|r,u,w)$ *are as follows:*
If $k < u$, $t < u$ *and* $t \ne k$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(r-2)\delta(w-2)\delta(u-2)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-3}{r-3}\binom{N-i}{n-r}.$$

If $k > w$, $t > w$ *and* $t \ne k$,

$$\pi_{k,t}^{(r)}(u,w) = \\ = \frac{\delta(n-r-1)\delta(N-w-1)\delta(N-u-1)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-1}{r-1}\binom{N-i-2}{n-r-2}.$$

If $k < u$ *and* $t > w$ *or* $t < u$ *and* $k > w$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(r-1)\delta(n-r)\delta(u-1)\delta(N-w)}{z_r(u,w)}\sum_{i=u}^{w}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}.$$

If $k < u$ *and* $u \le t \le w$ *or* $t < u$ *and* $u \le k \le w$,

$$\pi_{k,t}^{(r)}(u,w) = \\ = \frac{\delta(r-1)}{z_r(u,w)}\left(\delta(n-r)\delta(t-u)\delta(t-2)\sum_{i=u}^{t-1}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}+\right. \\ +\delta(t-1)\binom{t-2}{r-2}\binom{N-t}{n-r}+ \\ \left.+\delta(r-2)\delta(w-t)\delta(w-2)\delta(t-1)\sum_{i=t+1}^{w}\binom{i-3}{r-3}\binom{N-i}{n-r}\right).$$

If $u \le k \le w$ *and* $t > w$ *or* $u \le t \le w$ *and* $k > w$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(n-r)}{z_r(u,w)}\Bigg(\delta(n-r-1)\delta(k-u)\delta(N-k)\delta(k-1)\delta(N-u-1)\cdot$$

$$\cdot\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-2}{n-r-2} + \delta(N-k)\binom{k-1}{r-1}\binom{N-k-1}{n-r-1}+$$

$$+\delta(r-1)\delta(w-k)\delta(N-w)\delta(w-1)\delta(N-k-1)\sum_{i=k+1}^{w}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}\Bigg).$$

If $u \le k < t \le w$ *or* $u \le t < k \le w$,

$$\pi_{k,t}^{(r)}(u,w) = \frac{\delta(w-u)}{z_r(u,w)}\Bigg(\delta(n-r-1)\delta(k-u)\delta(N-k)\delta(k-1)\delta(N-u-1)\cdot$$

$$\cdot\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-2}{n-r-2} + \delta(n-r)\delta(N-k)\binom{k-1}{r-1}\binom{N-k-1}{n-r-1}+$$

$$+\delta(r-1)\delta(n-r)\delta(t-k-1)\delta(t-2)\delta(N-k-1)\sum_{i=k+1}^{t-1}\binom{i-2}{r-2}\binom{N-i-1}{n-r-1}+$$

$$+\delta(r-1)\delta(t-1)\binom{t-2}{r-2}\binom{N-t}{n-r}+$$

$$+\delta(r-2)\delta(w-t)\delta(w-2)\delta(t-1)\sum_{i=t+1}^{w}\binom{i-3}{r-3}\binom{N-i}{n-r}\Bigg).$$

Example 2.1. *Let us assume that* $N = 11$, $n = 5$, $r = 3$.
When $u = 4$ *and* $w = 8$, *then* $\pi_k = 0.431$ *and* $\pi_t = 0.483$ *for* $k = 1, 2, 3, 9, 10, 11$ *and* $t = 4, 5, 6, 7, 8$.
When $u = 5$ *and* $w = 7$, *then* $\pi_k = 0.411$ *and* $\pi_t = 0.571$ *for* $k = 1, 2, 3, 4, 8, 9, 10, 11$ *and* $t = 5, 6, 7$.
When $u = 6$ *and* $w = 6$, *then* $\pi_6 = 1$*and* $\pi_t = 0.4$ *for* $t \neq 6$.
Finally, when $u = 3$ *and* $w = 9$, *then* $\pi_k = \frac{5}{11} = 0.45(45)$ *for* $k = 1, ..., 11$. *In this case the conditional simple random sampling design reduces to the simple random sample drawn without replacement.*

Hence, when the parameters u *and* w *are closer and closer to each other then the probability of selecting to the sample the central population elements increases.*

3. Sampling scheme

The sampling scheme implementing the conditional simple random sampling design $P_0(s|r,u,w)$, where $r \le u \le w \le N-n+r$ is as follows. Firstly, population

elements are ordered according to increasing values of the auxiliary variable. Next, the i-th element of the population where $i = u, u+1, ..., w$, is drawn with the probability

$$P(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w) = \frac{P(X_{(r)} = x_i)}{P(x_u \leq X_{(r)} \leq x_w)} = \frac{g(r,i)}{\sum_{j=u}^{w} g(r,j)} \tag{4}$$

where $r = [n\alpha] + 1$.

Finally, two simple samples $s_1(i)$ and $s_2(i)$ are drawn without replacement from the subpopulations $U_1 = \{1, ..., i-1\}$ and $U_2 = \{i+1, i+2, ..., N\}$, respectively. The sample $s_1(i)$ is of the size $r-1$ and the sample $s_2(i)$ is of the size $n-r$. The sampling designs of these samples are independent and

$$P_0(s_1(i)) = \frac{1}{\binom{i-1}{r-1}}, \quad P_0(s_2(i)) = \frac{1}{\binom{N-i}{n-r}} \tag{5}$$

Hence, the selected sample is: $s = \{s_1(i), i, s_2)(i)\}$ and its probability is:

$$P(X_{(r)} = x_i | x_u \leq X_{(r)} \leq x_w) P_0(s_1(i)) P_0(s_2(i)) = P_0(s|r,u,w)$$

where $r = u, u+1, ..., w$.

4. The Horvitz-Thompson estimator

The well-known Horvitz-Thompson (1952) estimator is given by:

$$\bar{y}_{HT,s} = \frac{1}{N} \sum_{k \in s} \frac{y_k}{\pi_k} \tag{6}$$

The estimation strategy $(\bar{y}_{HT,s}, P(s))$ is unbiased for the population mean $\bar{y}$ if $\pi_k > 0$ for $k = 1, ..., N$, where π_k is the inclusion probability of the sampling design $P(s)$. The variance of the strategy is:

$$V_0\left(\bar{y}_{HT,s}, P(s)\right) = \frac{1}{N^2} \left(\sum_{k \in U} \sum_{l \in U} \Delta_{k,l} \frac{y_k y_l}{\pi_k \pi_l} \right), \qquad \Delta_{k,l} = \pi_{k,l} - \pi_k \pi_l \tag{7}$$

Particularly, under the simple random sampling design $P_0(s)$ the strategy $(t_{HT,s}, P(s))$ reduces to simple random sample mean denoted by $(\bar{y}_s, P_0(s))$, where

$$\bar{y}_s = \frac{1}{n} \sum_{k \in s} y_k. \tag{8}$$

It is an unbiased estimator of the population mean and its variance is given by:

$$V_0(\bar{y}_s) = \frac{N-n}{Nn} v_*(y), \qquad v_*(y) = \frac{1}{N-1} \sum_{k \in U} (y_k - \bar{y})^2.$$

Example 4.1. *Let us assume that in the population of the size* $N = 11$ *the following values* (x, y) *of the two dimensional variable is observed* $\{(1,2), (2,6), (3,10), (4,14), (5,15), (6,16), (7,17), (8,18), (9,22), (10,26), (11,30)\}$. *Let the sample*

of size $n = 5$ be selected from that population according to the conditional sampling design. The variance of the simple sample mean is: $V(\bar{y}_s) = V(\bar{y}_{HTs}, P_0(s)) = V(\bar{y}_{HTs}, P_0(s|3,3,9)) = 7.353$. The variances of the Horvitz-Thompson estimator under the conditional design of simple sample are: $V(\bar{y}_{HTs}, P_0(s|3,4,8)) = 5.954$, $V(\bar{y}_{HTs}, P_0(s|3,5,7)) = 4.918$, $V(\bar{y}_{HTs}, P_0(s|3,6,6)) = 3.694$. The inclusion probabilities of the conditional simple random sample are shown in the Example 2.1. Hence, the accuracy of estimation of the population mean on the basis of the Horvitz-Thompson statistic under the considered variants of the conditional simple random sample is better than the accuracy of the mean from the unconditional simple random sample.

5. Conclusions

The sampling design belonging to the class of the sampling designs dependent on the sample parameters of an auxiliary variable was proposed. It is the conditional version of the simple random sampling design explained by Definition 2.1 and denoted by $P_0(s|x_u \leq X_{(r)} \leq x_w)$. Let M_s be the sample median of the auxiliary variable. So, when we assume that the distribution of an auxiliary variable is symmetric then $\bar{x} = M$, where M is the population median of the auxiliary variable. When we assume that the distribution of the sample median is approximation of the distribution of the sample mean $\bar{x}_s$ then the simple random sample design $P_0(s|x_u \leq M_s \leq x_w)$ can be treated as approximation of the simple random sampling design $P_0(s|x_u \leq \bar{x}_s \leq x_w)$, defined by Royall and Cumberland (1981). Our consideration can be generalized to the case when the distribution of the auxiliary variable is not necessary symmetric. It is possible to find such rank r of the order statistic $|E(X_{(r)}) - \bar{x}| = minim$. So, when we assume that the distribution of the sample mean $\bar{x}_s$ is sufficiently approximated by the distribution of the order statistic $X_{(r)}$ then the sampling design $P_0(s|x_u \leq \bar{x}_s \leq x_w)$ can be approximated by the sampling design $P_0(s|x_u \leq X_{(r)} \leq x_w)$.

We can expect that the sampling design can be useful in the case of censored observations of the auxiliary variable as well as when the outliers exist. The precision of the Horvitz-Thompson estimator depends on the parameters u and w through probabilities of the inclusion of the first and second order.

The derived properties of the sampling designs lead to the conclusion that without an additional extensive analysis it is not possible to determine precisely how the sampling strategies depend on the parameters of the conditional simple random sampling design as well as on the joint distribution of the variable under study and the auxiliary variable. This problem will be considered on the basis of computer simulation analysis in another papers. Moreover, such analysis makes it possible to compare the accuracy of the proposed estimation strategies with accuracy of the strategies typically used in statistical research.

Acknowledgement

The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

REFERENCES

GUENTER W. (1975). The inverse hypergeometric - a useful model. Statistica Neerlandica, Vol. 29, pp. 129–144.

HOGG, R. V., CRAIG, A. T., (1970). Introduction to Mathematical Statistics, 3rd edition. MacMillian, New York.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of the sampling without replacement from finite universe. Journal of the American Statistical Association, Vol. 47, pp. 663–685.

ROYALL, R. M., Cumberland W. G., (1981). An empirical study of the ratio estimator and estimators of its variance. Journal of the American Statistical Association, Vol. 76.

TILLÉ, Y., (1998). Estimation in surveys using conditional inclusion probabilities: Simple random sampling. International Statistical Review, 66, 303–322.

TILLÉ, Y., (2006). Sampling Algorithms. Springer.

WILKS, S. S., (1962). Mathematical Statistics. John Wiley and Sons, Inc. New York, London.

WYWIAŁ, J. L., (2003). On conditional sampling strategies. Statistical Papers, Vol. 44, 3, pp. 397–419.

WYWIAŁ, J. L., (2008). Sampling design proportional to order statistic of auxiliary variable. Statistical Papers, Vol. 49, No. 2/April, pp. 277–289.

APPENDIX

The theorems 4.1 formulated in the previous sections is proved here.
Let $\mathbf{S}(U(1,...,i-1),s_1(i))$ and $\mathbf{S}(U(i+1,...,N),s_2(i))$ be the sample spaces of the samples $s_1(i)$ and $s_2(i)$ and $s = s_1(i) \cup \{i\} \cup s_2(i)$, defined in Section 1. Hence,

$$\mathbf{S}(r,i) = \mathbf{S}(U(1,...,i-1),s_1(i)) \times \{i\} \times \mathbf{S}(U(i+1,...,N),s_2(i))$$

and

$$\mathbf{S}(r;u,w) = \mathbf{S}(r,u) \times \mathbf{S}(r,u+1) \times ... \times \mathbf{S}(r,i) \times ... \times \mathbf{S}(r,w)$$

where $\mathbf{S}(r,i)$ was defined in Section 1.

Wywiał (2008) proposed the following conditional sampling design:

Definition 6.1. *The conditional sampling design proportional to the values* x_i, $i = u,...,w \leq N-n+r$, $u \geq r$, *of the order statistic* $X_{(r)}$ *is as follows:*

$$P_r(s|u,w) = \frac{x_i}{\sum_{j=u}^{w} x_j g(r,j)}$$

where $\quad i \in s \in \boldsymbol{S}(r,i), \quad r \leq u \leq i \leq w \leq N-n+r.$

Moreover, Wywiał (2008) proved the theorem:

Theorem 6.1. *The inclusion probabilities of the first order for the conditional simple random sampling design* $P_r(s|u,w)$ *are as follows:*
If $k < u$,

$$\pi_k^{(r)}(u,w) = \frac{\delta(r-1)\delta(w-1)\delta(u-1)}{z_r(u,w)} \sum_{i=u}^{w} \binom{i-2}{r-2}\binom{N-i}{n-r} x_i,$$

If $u \leq k \leq w$,

$$\pi_k^{(r)}(u,w) =$$

$$= \frac{1}{z_r(u,w)}\left(\delta(n-r)\delta(k-u)\delta(k-1)\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1}x_i + \right.$$

$$\left. + \binom{k-1}{r-1}\binom{N-k}{n-r}x_k + \delta(r-1)\delta(w-k)\sum_{i=k+1}^{w}\binom{i-2}{r-2}\binom{N-i}{n-r}x_i\right),$$

if $k > w$,

$$\pi_k^{(r)}(u,w) = \frac{\delta(n-r)\delta(N-w)}{z_r(u,w)} \sum_{i=u}^{w} \binom{i-1}{r-1}\binom{N-i-1}{n-r-1} x_i,$$

When we replace x_i by 1 for all $i = 1, ..., N$, then the above definition 6.1 and the expression (3) lead to the Definition 2.1. The same operation and the above Theorem 6.1 lead straightforward to the derivation of the first order inclusion probabilities of the conditional simple random sampling design $P_0(s|r, u, v)$, given by expressions (2) and (3). The inclusion probabilities of the second order presented by Theorem 6.1. can be straightforward derived in the same way but on the basis of the appropriate theorem proven by Wywiał (2008).

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 535-544

STATISTICAL ANALYSIS OF A QUESTIONNAIRE: VOLUNTARY HEALTH INSURANCE IMPLEMENTATION AMONG PATIENTS SUFFERING FROM ALLERGY AND ASTHMA

Marta Zalewska[1]
Wojciech Zieliński[2]

ABSTRACT

We consider statistical analysis of multiple answers in a questionnaire. We propose a new method of calculating simultaneous confidence regions. In a communication presented at the European Academy of Allergy and Clinical Immunology the authors (Borowicz et al. (2009)) reported the proportions of respondents which gave one of three possible exclusive answers in a questionnaire concerning the role of voluntary health insurance. There were three possible answers. Apart from percentages of answers confidence intervals of every single answer have been reported. Unfortunately inference about the population based on such intervals may lead to imprecise conclusions.

The inference about the respective population suffering from allergy and asthma proportions requires the construction of two-dimensional confidence region. We propose the use of a simultaneous confidence intervals to inference about true population proportions.

Most of our attention is given to the case of three possible answers but the results may be generalized to any questionnaire with more than two excluding answers.

Key words: confidence region, health insurance, multiple responses, questionnaire

[1] Department of the Prevention of Environmental Hazards and Allergology, Medical University of Warsaw, Banacha 1a, 02-097 Warszawa,e-mail: zalewska.marta@gmail.com

[2] Department of Econometrics and Statistics, Warsaw University of Life Sciences, Nowoursynowska 159, 02-776 Warszawa, e-mail: wojciech_zielinski@sggw.pl

1. Introduction

We consider statistical analysis of multiple answers in a questionnaire. We propose a new method of calculating simultaneous confidence regions. In our article we concentrate on the example of voluntary health insurance implementation among patients suffering from allergy and asthma. There is an obligatory health insurance system in Poland. Unfortunately this system does not work efficiently, mainly because of incorrect diversification of funds. For these reasons together with the obligatory health insurance system we have the optional voluntary health insurance system (VHI) based on voluntary premium. Quite a large number of people in Poland participate in VHI system but the reasons for participating in this system are different. Epidemiology of Allergic Disease survey in Poland (presented during European Academy of Allergy and Clinical Immunology congress in 2009) included a question about the reasons for participating in VHI with three possible answers: additional, supplementary and substitutive (question number 566, and answers 566_1, 566_2 and 566_3 respectively). The results of the questionnaire given in Borowicz et al. (2009) are presented in Table 1.

Table 1. Results of the questionnaire (Borowicz et al. (2009))

The role of voluntary health insurance (question 566)	Frequency	Percentage
additional-increasing health service standard (answer 566_1)	1653	36.5
supplementary-expanding range of health service (answer 566_2)	1668	36.9
substitutive-enabling abandonment of public health care (answer 566_3)	1205	26.6

The results were obtained on the basis of the questionnaire based on the International Study of Asthma and Allergies in Childhood and the European Community Respiratory Health Survey II ECRHS II. All investigated subjects were randomly selected from PESEL (Personal Identification Number). Data acquisition was done by the Computer Assisted Personal Interviewing with GSM transmission to update the main database at the Medical University of Warsaw (http://ecap.pl/eng_www).

The question is: what are the population suffering from allergy and asthma percentages π_1, π_2 and π_3 of the Polish citizens participating in VHI system from the appropriate reasons (additional, supplement and substitutive). The standard approach is to construct individual confidence intervals. Unfortunately this approach may lead to wrong conclusions. Therefore, in what follows we propose to construct a confidence region for percentages π_1, π_2 and π_3 simultaneously.

2. Statistical model

Let X denote the random variable describing answers. It may be assumed that X is multinomially distributed:

$$P_{\boldsymbol{\pi}}\{X=1\}=\pi_1, P_{\boldsymbol{\pi}}\{X=2\}=\pi_2, P_{\boldsymbol{\pi}}\{X=3\}=\pi_3,$$

where $\boldsymbol{\pi}=(\pi_1,\pi_2,\pi_3)$ and $0<\pi_1,\pi_2,\pi_3<1$, $\pi_1+\pi_2+\pi_3=1$. Values of X symbolize answers to questions in the questionnaire (i.e. $X=1$ means that the answer is 566_1, $X=2$ means that the answer is 566_2 and $X=3$ - the answer 566_3). Probabilities π_1, π_2 and π_3 are (multiplied by 100%) population suffering from allergy and asthma (population to be short) percentages of obtaining answers to the questions.

Assume that in a sample of size n, value 1 was observed n_1 times, value 2 - n_2 times and value 3 - n_3 times. Of course $n_1+n_2+n_3=n$. It is known that the maximum likelihood estimator of $\boldsymbol{\pi}$ is: $\hat{\pi}_1=n_1/n$, $\hat{\pi}_2=n_2/n$ and $\hat{\pi}_3=n_3/n$. The problem is in the interval estimation of $\boldsymbol{\pi}$, the vector comprising the probabilities of answers.

In standard approach, each of the probabilities is estimated separately. It means, that three confidence intervals are obtained, usually on the basis of normal approximation, i.e. a confidence interval of the form is built for π_i (at the confidence level $1-\alpha$)

$$\left(\hat{\pi}_i-\frac{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}{\sqrt{n}}z, \hat{\pi}_i+\frac{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}{\sqrt{n}}z\right),$$

where z is the quantile of the order $1-\alpha/2$ of the standard normal distribution (i.e. $N(0,1)$ distribution) and

$$\hat{\pi}_i=\frac{n_i}{n}.$$

This approach gives the results (at 95% confidence level, i.e. $1-\alpha=0.95$) presented in Table 2.

Table 2. Individual confidence intervals for percentages

	Frequency n_i	Estimated Percentage $\hat{\pi}_i$	Left end	Right end
π_1 (answer 566_1)	1653	36.5	35.12	37.93
π_2 (answer 566_2)	1668	36.9	35.45	38.28
π_3 (answer 566_3)	1205	26.6	25.34	27.94

Classical inference is such that the population percentage of the answers to the first question is any number between 35.12% and 37.93%; to the second

question - the number from the interval $(35.45\%, 38.28\%)$ and to the third one is from the interval $(25.34\%, 27.94\%)$. But this kind of inference may lead to wrong conclusions. Namely, it may be stated, that the percentage of population answers to the question 566_1 is 36%, to the second question (566_2) is also 36% and to the third question is 26% (i.e. $(\pi_1, \pi_2, \pi_3) = (0.36, 0.36, 0.26)$). Summing up those three values one obtains 98% of the population instead of expected 100% (2% of population is "missed"!). The other situation is also possible, i.e. stated population percentages may give more than 100% (for example: the percentage of answers to the first question is 37%, to the second - 38% and to the third question 27%). It appears also that the real confidence level of such conclusion is less than the nominal 95%. It means that the risk of wrong conclusions is too high: it is greater than the nominal 5%.

We are interested in simultaneous interval estimation of probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$.

3. Confidence region

There are a lot of papers devoted to the problem of simultaneous confidence intervals for probabilities of multinomial distribution. An extensive review of construction methods may be found in Biszof and Mejza (2004), Correa (2001), May and Johnson (1997). The general rule of construction is based on the set of inequalities

$$\frac{|\hat{\pi}_i - \pi_i|}{\sqrt{\pi_i(1-\pi_i)}} \leq c, \quad i = 1, 2, 3,$$

where c is a constant such that the following equality holds

$$P_{\boldsymbol{\pi}} \left\{ \frac{|\hat{\pi}_i - \pi_i|}{\sqrt{\pi_i(1-\pi_i)}} \leq c, \quad i = 1, 2, 3 \right\} = 1 - \alpha, \qquad \forall \boldsymbol{\pi}.$$

Those confidence regions are easy to calculate. However, simultaneous confidence intervals have two disadvantages. Firstly, the obtained confidence intervals may go out of $(0, 1)$ interval and secondly, in their construction the condition $\pi_1 + \pi_2 + \pi_3 = 1$ was not exploited.

For example, let the following sample be given: $n_1 = 1$, $n_2 = 1$, $n_3 = 48$. In Table 3 the limits of some of known simultaneous confidence intervals ($1 - \alpha = 0.95$) are given.

Table 3. Simultaneous confidence intervals

	QH		GM		NB		FS	
$\hat{\pi}_1$=0.02	0.0025	0.1402	0.0026	0.1361	-0.1493	0.1893	-0.1303	0.1703
$\hat{\pi}_2$=0.02	0.0025	0.1402	0.0026	0.1361	-0.1493	0.1893	-0.1303	0.1703
$\hat{\pi}_3$=0.96	0.8300	0.9916	0.8340	0.9914	0.7907	1.1293	0.8097	1.1103

QH denotes Quesenberry and Hurst (1964) construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \chi^2(\alpha, 2)\pi_i(1-\pi_i), \quad i = 1, 2, 3.$$

GM denotes Goodman (1965) construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \chi^2(\alpha/3, 1)\pi_i(1-\pi_i), \quad i = 1, 2, 3.$$

NB denotes naive binomial construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \chi^2(\alpha, 1)(1/4), \quad i = 1, 2, 3.$$

FS denotes Fitzpatrick and Scott (1987) construction:

$$n_i(\hat{\pi}_i - \pi_i)^2 \leq \gamma, \quad i = 1, 2, 3.$$

where $\gamma = 1$ for $\alpha = 0.1$, $\gamma = 1.13$ for $\alpha = 0.05$ and $\gamma = 1.40$ for $\alpha = 0.01$.

Note that the left ends of some of the calculated confidence intervals are negative or the sum of admissible probabilities is greater than one.

In what follows we propose another way of inference. We show how to built a confidence region for all three percentages simultaneously, such that:
1. all percentages in the confidence region will sum up to 100%;
2. the confidence level of conclusion will be equal to the nominal one.

Let us start with the very well known chi-square statistic Bland (2000), Greenwood and Nikulin (1996), Peacock and Peacock (2011) of the Pearson goodness-of-fit test:

$$\chi^2 = n \cdot \left(\frac{\left(\frac{n_1}{n} - \pi_1\right)^2}{\pi_1} + \frac{\left(\frac{n_2}{n} - \pi_2\right)^2}{\pi_2} + \frac{\left(\frac{n_3}{n} - \pi_3\right)^2}{\pi_3} \right).$$

To satisfy the first requirement the statistic above is transformed to

$$\chi^2(\pi_1, \pi_2) = n \cdot \left(\frac{\left(\frac{n_1}{n} - \pi_1\right)^2}{\pi_1} + \frac{\left(\frac{n_2}{n} - \pi_2\right)^2}{\pi_2} + \frac{\left(\frac{n_3}{n} - (1 - \pi_1 - \pi_2)\right)^2}{(1 - \pi_1 - \pi_2)} \right).$$

This statistic may be used in the construction of the confidence region in the following way. Let $\chi^2(\alpha; 2)$ denote the chi-square critical value with two degrees of freedom and the confidence level $1 - \alpha$. Then the confidence region for $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ is obtained as a solution with respect to $\boldsymbol{\pi}$ of the inequality $\chi^2(\pi_1, \pi_2) < \chi^2(\alpha; 2)$:

$$\left\{ (\pi_1, \pi_2, \pi_3) : \chi^2(\pi_1, \pi_2) < \chi^2(\alpha; 2), \pi_3 = 1 - \pi_1 - \pi_2 \right\}.$$

The theoretical background for constructing a confidence region for probabilities $\boldsymbol{\pi}$ may be found in Harton and Zieliński (2005) and Zieliński (2008). Explicit formulae for the confidence region may also be found in those papers. Note that the above construction tacitly assumes that the population in question is infinite. Of course, this is not exactly true because the population of adults in Poland is finite, but it is sufficiently large to accept this assumption as a reasonable approximation. Some remarks on the application of statistical methods devoted to the analysis of infinite populations to finite ones may be found in Zieliński (2011).

4. Results

We apply the constructed above confidence region to the problem of estimating the role of voluntary health insurance (question 566). In the questionnaire the $n = 4526$ answers were obtained. Among them there were $n_1 = 1653$ answers to the first question, $n_2 = 1668$ answers to the second question and $n_3 = 1205$ answers to the third one. As a confidence level 95% were taken, so the critical value of the chi-square distribution with two degrees of freedom equals 5.99. After some calculations the confidence region for π_1 and π_2 was obtained and is presented in Figure 1 (all computations were done using R-project with statistical computing (R Development Core Team (2008)); the computer codes in R were written by ourselves - see Appendix).

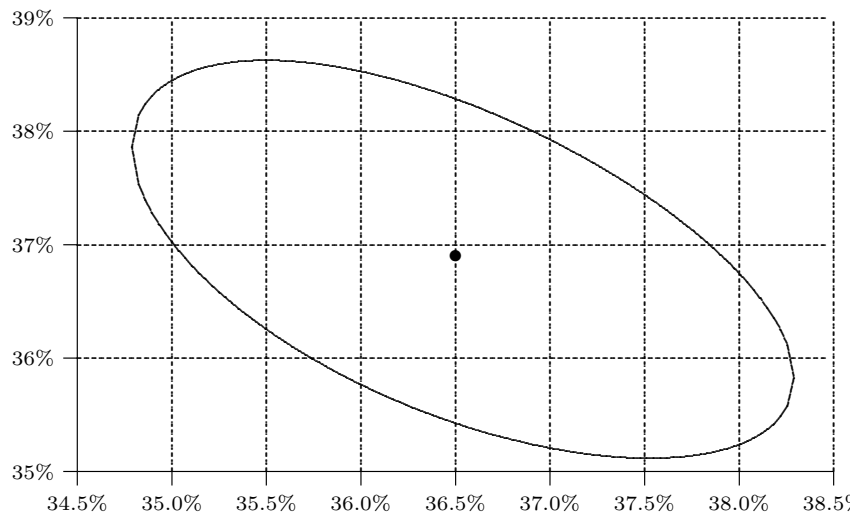

Figure 1. Confidence region for frequency of opinions of the role of health insurance

The letters π_1, π_2 and π_3 denote the proportion of answers to the first, second and third question, respectively in the population of interest (the graph

shows only π_1 and π_2 because $\pi_3 = 1 - \pi_1 - \pi_2$). Recall that the first answer is "additional-increasing health service standard"; the second answer - "supplementary-expanding range of health service" and the third answer is "substitutive - enabling abandonment of public health care".

The confidence region for π_1 and π_2 is inside the contour presented in Figure 1. We have to remember that this two dimensional graphs in fact inform us about three proportions (three possible answers to a given question). The dot "in the center of the confidence region" corresponds to the proportions in the sample: $\hat{\pi}_1 = 0.365$ (36.5%), $\hat{\pi}_2 = 0.369$ (36.9%)- these two are presented in the graph - and $\hat{\pi}_3$ given by $\hat{\pi}_3 = 1 - \hat{\pi}_1 - \hat{\pi}_2 = 0.266$ (26.6%).

The interpretation of this confidence region is similar to that of a confidence interval but two-dimensional. Roughly speaking we trust that combinations of values π_1 and π_2 lie inside the region. More precisely, we make statements with probability of error 0.05 for all three proportions together. For example, a combination of proportions $\pi_1 = 0.375$ (37.5%) and $\pi_2 = 0.36$ (36%) may be true with high confidence because the point with coordinates 0.375 (37.5%) and 0.36 (36%) lies inside the contour shown in Figure 1 (i.e. $(\pi_1, \pi_2, \pi_3) = (0.36, 0.375, 0.265)$). On the other hand the combination $\pi_1 = 0.355$ and $\pi_2 = 0.36$ we treat as extremely unlikely because the point with coordinates 0.355 (35.5%) and 0.36 (36%) lies outside the contour. This last statement is true in spite of the fact that $\pi_1 = 0.355$ (35.5%) considered separately is possible and $\pi_2 = 0.36$ (36%) considered separately is also possible, but both these values together are unlikely (see Figure 2).

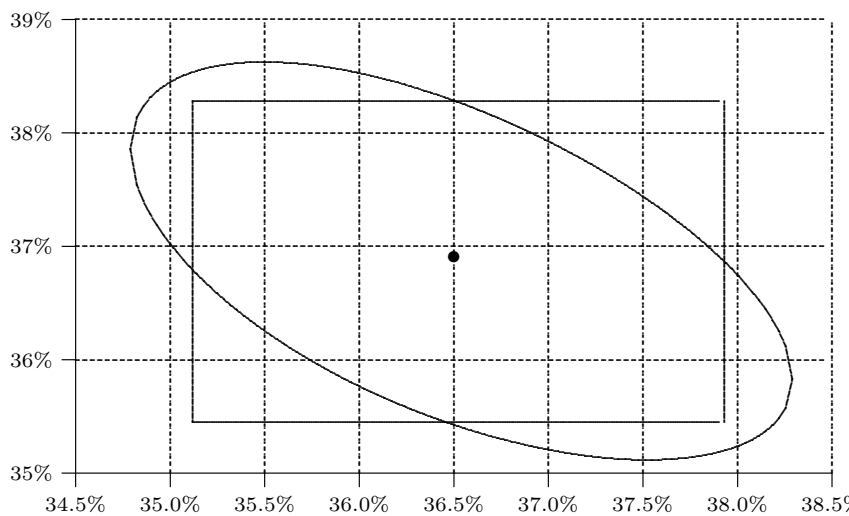

Figure 2. Confidence region for frequency of opinions of the role of health insurance. The rectangle shows two confidence intervals separately for π_1 and π_2.

In Figure 2 the elliptical confidence region is drawn along with the rectangular region of standard approximate confidence intervals. The analysis we present gives more precise information than conventional onedimensional confidence intervals. It might be argued that the interpretation of two-dimensional confidence regions is more difficult than that of one dimensional confidence intervals. However, easily accessible modern computer graphics allows us to show the relations between two or three variables and to understand and explain to the users the meaning of the confidence region.

5. Conclusions

The construction of the confidence region for three probabilities may be easily generalized to the problem of estimating more than three percentages. Of course, if there is a problem of estimating more than three proportions, the graphical illustration is impossible. For k possible mutually excluding answers it is sufficient to consider the statistic

$$\chi^2(\pi_1, \ldots, \pi_k) = n \cdot \sum_{i=1}^{k} \left(\frac{\left(\frac{n_i}{n} - \pi_i\right)^2}{\pi_i} \right)$$

and as the confidence region at the confidence level $1 - \alpha$

$$\left\{ (\pi_1, \ldots, \pi_k) : \chi^2(\pi_1, \ldots, \pi_k) < \chi^2(\alpha; k-1), \pi_1 + \cdots + \pi_k = 1 \right\}.$$

The interpretation is to some extent more complicated than in the case of individual confidence intervals, but it avoids the errors of inference.

In many allergological questionnaires there are numerous questions with multiple answers. We show that simultaneous inference is more appropriate and more informative than the one-dimensional ones. The latter can lose some relevant information while multidimensional analysis is more accurate.

REFERENCES

BISZOF, A., MEJZA, S., (2004). Jednoczesne przedziały ufności dla prawdopodobieństwa w rozkładzie wielomianowym, Colloquium Biometryczne, 34, 77-84.

BLAND, M., (2000). An introduction to medical statistics, Oxford University Press, Third edition.

BOROWICZ, J., SAMOLIŃSKI, B., FURMAŃCZYK, K., WALKIEWICZ, A., LUSAWA, A., MARSZAŁKOWSKA, J. et al., (2009). Attitudes towards the idea of voluntary health insurance implementation among patients suffering from allergy and asthma. Allergy 64 (Suppl. 90): 435 (abstract 1140)

CORREA, J. C., (2001). Interval Estimation of the Parameters of the Multinomial Distribution, ip.statjournals.net:2002/InterStat/ARTICLES/2001/articles/O01001.pdf.

FITZPATRICK, S., SCOTT, A., (1987). Quick simultaneous confidence intervals for multinomial proportions, Journal of the American Statistical Association, 82, 875-878.

GOODMAN, L. A., (1965). On simultaneous confidence intervals for multinomial proportions, Technometrics, 7, 247-254.

GREENWOOD, D. E., NIKULIN, M. S., (1996). A guide to chi-squared testing, Willey.

HARTON, A., ZIELIŃSKI, W., (2005). Confidence region for probabilities of a multinomial distribution. Colloquium Biometricum, 35, 141-145.

MAY, W. L., JOHNSON, W. D., (1997). Properties of simultaneous confidence intervals for multinomial proportions, Communications in Statistics - Simulations, 26, 495-518.

PEACOCK, J. L., PEACOCK, P. J., (2011). Oxford handbook of medical statistics. Oxford University Press.

QUESENBERRY, C. P., HURST, D. C., (1964). Large sample simultaneous confidence intervals for multinomial proportions,Technometrics, 6, 191-195.

ZIELIŃSKI, W., (2008). A remark on interpretation of pooling results. Folia Oeconomica Stetinensia, 7(15), 56-62.

ZIELIŃSKI, W., (2011). Comparison of confidence intervals for fraction in finite populations, Quantitative Methods in Economics, XII, 177-182.

R DEVELOPMENT CORE TEAM, (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, http://www.R-project.org.

http://ecap.pl/eng_www/index_home.html.

Appendix. The computer code in R which was employed to draw Figure 1.

```
n1=1653 #input n1
n2=1668 #input n2
n3=1205 #input n2
n=n1+n2+n3 #the overall number of observations
alpha=0.05 #confidence level
pu1=as.vector(prop.test(n1,n,conf.level =1-alpha/2)$conf.int)
pu2=as.vector(prop.test(n2,n,conf.level =1-alpha/2)$conf.int)
prop.test(n3,n,conf.level =1-alpha/2)
p1=n1/n #estimated proportion π1
p2=n2/n #estimated proportion π2
p3=n3/n #estimated proportion π3
chi=qchisq(1-alpha,2)/n #chi-square critical value
# assistant functions
delta= function(pi1){
delta=(chi*(-1+pi1)*pi1+pi1^2+p1^2+2*pi1*(p1*(-1+p2)-p2))^2+
4*(-1+pi1)*pi1*(pi1+chi*pi1-p1^2)*p2^2
delta
}
p2L=function(pi1){
p2L=-(chi*(-1+pi1)*pi1+pi1^2+p1^2-2*pi1*(p1+p2-p1*p2)+
sqrt(delta(pi1)))/ (2*(pi1+chi*pi1-p1^2))
p2L
}
p2P=function(pi1){
p2P=-(chi*(-1+pi1)*pi1+pi1^2+p1^2-2*pi1*(p1+p2-p1*p2)-
sqrt(delta(pi1)))/(2*(pi1+chi*pi1-p1^2))
p2P
}
P1L=(chi+2*p1-sqrt(chi)*sqrt(chi+4*p1-4*p1^2))/(2*(1+chi))
P1P=(chi+2*p1+sqrt(chi)*sqrt(chi+4*p1-4*p1^2))/(2*(1+chi))
c(P1L,P1P)
pi1=seq(P1L,P1P,length.out =250)
pi2L=p2L(pi1)
pi2P=p2P(pi1)
#Plot of confidence regions for frequencies (π1,π2)
ymax=max(pi2P)
ymin=min(pi2L)
plot(pi1,pi2L,xlim=c(P1L,P1P),ylim=c(round(ymin,2),round(ymax,2)),type="l",las=1,
ylab=expression(pi[2]),xlab=expression(pi[1]),lwd=2)
lines(pi1,pi2P,lwd=2)
points(p1,p2,pch=16)
```

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 545–558

PROPOSITION OF STOCHASTIC POSTULATES FOR CHAIN INDICES

Jacek Białek[1]

ABSTRACT

This article presents and discusses a proposition of stochastic postulates for chain indices. The presented postulates are based on the assumption that prices and quantities are stochastic processes and we consider also the case when price processes are martingales. We define general conditions which allow the chain indices to satisfy these postulates.

Key words: chain indices, price index theory, stochastic processes, martingales.

1. Introduction

The idea of chain index construction, with weights changed every year, was probably first suggested by Alfred Marshall (1887). Marshall was concerned only with the practical problem of allowing for introduction of new commodities into an index of prices. He thought that the index would be greatly facilitated if weights were changed every year and the successive yearly indices linked or changed together by simple multiplication. Francois Divisia (1925) also postulated that the price index should depend not only on prices and quantities at considered moments $t = 0$ and $t = T$ but also on the movement of prices and quantities throughout the interval $[0, T]$. Divisia defined the index of prices by using a differential equation the solution of which was a curvilinear integral, and under assumptions that all functions $p_i(t)$ and $q_i(t)$, describing values of (respectively) prices and quantities of the considered N commodities ($i \in \{1,2,...,N\}$), exist at any point in time. Divisia's approach seems to be related to chain indices although it has a more general character (see Hulten (1973), Banerjee (1979)). Some authors treat Divisia's approach as some kind of justification for chain indices (see von der Lippe (2007). In fact, in some authors' opinion, all index formulas used in practice should approximate the Divisia index and chain indices should naturally translate the Divisia index into the reality of

[1] Department of Statistical Methods, University of Łódź, Poland. E-mail: jbialek@uni.lodz.pl.

price observations at discrete points in time (see Feenstra, Reinsdorf (2000)). It should be added that there are infinitely many discrete approximations to the Divisia continuous time index because the value of this index depends on the path connecting moments $t = 0$ and $t = T$ (see Samuelson and Swamy (1974), Vogt (1978)). The ideas of chain indices and Divisia's approach have many supporters and opponents (see Forsyth and Fowler (1981)) but we must admit with absolute certainty that chain indices play important role in practice (see for example Cho (2006)) and are recommended for deflation by the revised System of National Accounts[1] (see Von der Lippe (2001)). In this paper we try to supplement the theoretical background for chain indices by adding some new, stochastic postulates for them. According to NSA (News Stochastic Approach[2]) we treat the prices (and also quantities) of commodities as random variables. We claim that on the one hand these postulates are quite natural requirement but on the other hand some of them rules out known chain index formulas.

2. Chain indices and their properties

In the monograph of Von der Lippe (2007, p. 133) we can read: "*A chain index is essentially a specific type of aggregation (over intervals in time) and description of a time series rather than a comparison of two states taken in isolation; it provides a measure of the cumulated effect of successive steps (and the shape of the path) from 0 to 1, 1 to 2, …, $t-1$ to t*". Let us denote by $P_{\tau-1,\tau}$ a *direct* price index formula (like Paasche, Laspeyres, Fisher or others). The chain index $\overline{P}_{0,t}$ calculated for the considered time interval $[0,t]$ can be expressed as a product of "links" $P_{\tau,\tau+1}$

$$\overline{P}_{0,t} = \prod_{\tau=0}^{t-1} P_{\tau,\tau+1}\,, \tag{1}$$

where each price index $P_{\tau,\tau+1}$ compares moment $\tau+1$ with the preceding moment τ.

In the literature we can meet a few major arguments for using chain indices in practice (Von der Lippe (2007)). We can list these arguments as follows: a) the "base" to which a time series of indices or of year-to-year growth rates refers is more relevant and realistic in the case of chain indices than in the case of traditional direct indices; b) some advantages are derived from a superior flexibility and adaptability as regards the structure of weights and the appearance

[1] According to the recommendations of System of National Accounts (SNA 1993) the chained Fisher price index should be used for both price level measurement and deflation (see also Von der Lippe (2007), p. 365).

[2] For more details see Clements and Izan (1987) and Selvanathan and Prasada Rao (1994).

of new and disappearance of old goods[1]; c) on the one hand chain indices are found unfavourable in the case of high price fluctuations or when cycles exist but on the other hand these indices are useful (in terms of desirable numerical results, like low inflation) "if individual prices and quantities tend to increase or decrease monotonically over time[2]"; d) in many authors' opinion the chain index version of various index formulas will yield less divergent results than the corresponding direct index version (see Von der Lippe (2007)); e) chain indices are recommended for deflation procedure by SNA. It is worth adding that many statements presenting advantages of chain indices over the direct indices focus on the links rather than the chain and they give arguments in favour of the chain index approach from the simple fact that the interval $[0,t]$ is subdivided into a number of sub-intervals and the chain index is derived from multiplications of links. Although in Europe chain indices are made mandatory for official statistics, not everyone shares the above-mentioned opinions and some criticism of the presented arguments can be found in Von der Lippe (2007). For instance, the Boskin Commission (1996) did not recommend chain indices but rather direct "superlative" indices (like Fisher formula), with weights from periods 0 and t.

However, we should also discuss axiomatic properties of chain indices to have a full list of arguments for or against these indices. It has to be mentioned here that from the axiomatic point of view the chain indices have many drawbacks. It is not only an easy theoretical possibility (see Von der Lippe (2007)) that chain indices may fail *the mean value test*[3], this has been shown empirically already at least once (Szulc (1983)). It means that a chain index may exceed the greatest individual price relative or can be smaller than the smallest price relative. The list of unsatisfied tests is longer – chain indices fail also *identity*, *monotonicity*, or *transitivity*. Many of arguments advanced to justify chain indices suffer from a lack of axiomatic tools to evaluate their properties. Note that only the link is an index in the sense of axiomatic approach. In fact, a chain is not and index and can violate many of tests despite all indices, playing the role of links, satisfy them all. Moreover, as it was above-mentioned in the introduction, chain indices depend on how an interval is subdivided. Hence, chain indices provide a summary description of a process rather than a comparison of two moments. In the next part of the paper we propose quite natural postulates for this process, where the last of them comes from finance. In our opinion the above-mentioned postulates can play an axiom role and we show its connection with the traditional *mean value property*.

[1] The revised SNA 93 treats chain indices as „indices whose weighting structures are as up-to-date and relevant as possible". The SNA also found that chain indices make it "possible to obtain a much better match between products in consecutive time periods (…), given that products are continually disappearing from markets to be replaced by new product, or new qualities".

[2] SNA 93, para. 16.44.

[3] To read more about tests and axioms for price indices see Balk (1995).

3. New postulates and their interpretation

3.1. Stochastic model

Let us consider a group of N commodities. We observe them in discrete moments $\{t = 0,1,2,....,T\}$. Let us define a probability space $(\Omega, \Im, P)$. Let $F = \{\Im_t : t = 0,1,2,...,T\}$ be a filtration, i.e. each $\Im_t$ is an $\sigma-$algebra of Ω with $\Im_0 \subseteq \Im_s \subseteq \Im_t \subseteq \Im$ for any $s < t$. Without loss of generality, we assume $\Im_0 = \{\emptyset, \Omega\}$. The filtration F describes how the information about the market is revealed to the observer. We consider the following state-variables:

$p_i(t)$ - a price of the $i-$th commodity at time t,

$q_i(t)$ - a quantity of the $i-$th fund at time t,

$v_i(t) = p_i(t)q_i(t)$ - value of the $i-$th commodity at time t,

$v(t) = \sum_{i=1}^{N} v_i(t)$,

$v_i^*(t) = v_i(t)/v(t)$ - the percentage of a relative value of the $i-$th commodity at time t.

Here and subsequently, the symbol $X = Y$ means that the random variables X and Y are defined on $(\Omega, \Im, P)$ and it holds that $P(X = Y) = 1$. We assume that each $p_i(t)$ and $q_i(t)$ is adapted to $F = \{\Im_t : t = 0,1,2,...,T\}$, which means that each $p_i(t)$ and $q_i(t)$ is measurable with respect to $\Im_t$.

3.2. Stochastic postulates

As an initial stage of the discussion on postulates for chain indices we present the idea behind its definition. According to our best knowledge, the axiomatic price index theory is based on the deterministic approach and no test for indices is constructed for the case when prices and quantities are random. It would be quite interesting to rebuild the axioms on the stochastic case. For example, from the axiomatic approach (Balk (1995)) we know that one of the basic requirements for price indices is the so-called *proportionality*, which means that if all prices change $\lambda-$fold (from moment 0 to t) then the value of price index $P_{0,t}$ is also changed by λ. In other words, from

$$\frac{p_i(t)}{p_i(0)} = \lambda, \text{ for } i = 1,2,....,N, \tag{2}$$

we implicate $P_{0,t} = \lambda$.

The natural question is whether we should rebuild this axiom in stochastic case and require the following implication (let us call it *stochastic proportionality)*

$$E(\frac{p_i(t)}{p_i(0)}) = \lambda\ ,\ i = 1,2,....,N \ \Rightarrow\ E(P_{0,t}) = \lambda\ , \tag{3}$$

where $E(X)$ denotes the expected value of a random variable X .

Let us notice that in the special case, when $\lambda = 1$, we obtain the stochastic version of *identity* (constant prices test)

$$E(\frac{p_i(t)}{p_i(0)}) = 1,\ i = 1,2,....,N \ \Rightarrow\ E(P_{0,t}) = 1. \tag{4}$$

On the basis of the implication (4) we construct the first postulate for chain indices:

Postulate 1

The chain index $\overline{P}_{0,t}$ should satisfy

$$E(\frac{p_i(\tau+1)}{p_i(\tau)}) = 1,\ i = 1,2,....,N\ ,\tau = 0,1,...,t-1\ \Rightarrow\ E(\overline{P}_{0,t}) = 1. \tag{5}$$

As we know, for any random variables X and Y the condition $E(X) = E(Y)$ does not have to mean that $E(X/Y) = 1$. Thus, we propose another postulate, which on the one hand seems to be natural but on the other hand may be very restrictive:

Postulate 2

The chain index $\overline{P}_{0,t}$ should satisfy

$$E(p_i(\tau)) = p_i = const\ ,\ i = 1,2,....,N\ ,\tau = 0,1,...,t\ \Rightarrow\ E(\overline{P}_{0,t}) = 1. \tag{6}$$

If we assume that prices of commodities are martingales (see Williams (1991)), which means that each $E\left|p_i(t)\right| < \infty$ and additionally

$$E(p_i(t)/\Im_s) = p_i(s),\ \ i = 1,2,....,N\ , \tag{7}$$

we get the following conditional expected value of the partial index

$$E(\frac{p_i(t)}{p_i(s)}/\Im_s) = \frac{1}{p_i(s)}E(p_i(t)/\Im_s) = \frac{p_i(s)}{p_i(s)} = 1. \tag{8}$$

For $s = 0$ the equality (8) corresponds to the condition from the left side of the implication (4). Thus, we could expect that if the equality (7) holds for each price process, then the chain index should also behave like martingale and thus have the expected value constant in time. In other words, we form the following postulate for chain indices:

Postulate 3

If each process $\{p_i(t): t = 0,1,2,...,T\}$ is a $F-$martingale[1] for $i \in \{1,2,...,N\}$, then $\{\overline{P}_{0,t}: t = 0,1,2,...,T\}$ is also a $F-$martingale.

The postulate 3, although regarded as very important, seems to be less restrictive than postulates 1 and 2. In our opinion it has even axiomatic character because martingales have the expected value constant in time. Thus, the postulate 3 can play a role of a minimum requirement for chain indices. It is worth adding that the concept of martingale in probability theory is quite old and it was introduced by Paul Lévy in 1934, though he did not name it: the term *martingale* was introduced later by Ville (1939), who also extended the definition to continuous martingales. However, martingales play important role in modern probability, statistics and finance (Mansuy (2009)). In finance, in the case of measures of price dynamics on the given time interval, it is a very desirable property (see for example Gajek, Kałuszka (2000, 2001), Białek (2008)).

4. Some general remarks on the proposed postulates

In this section we discuss the general conditions for satisfying the presented postulates. In particular, we show some connections between traditional tests (postulates) for direct price indices and our postulates. We start our consideration from the theorem connected with the most fundamental postulate 3.

Theorem 1

If the direct price index formula (link) satisfies *the mean value test*[1] then the chain index $\overline{P}_{0,t}$, which is based on this link, satisfies the postulate 3.

[1] In probability theory, a martingale is a model of a fair game where knowledge of past events never helps predict the mean of the future winnings. In particular, a martingale is a sequence of random variables (i.e., a stochastic process) for which, at a particular time in the realized sequence, the expectation of the next value in the sequence is equal to the present observed value even given knowledge of all prior observed values at a current time. The assumption that prices are martingales is quite strong because it rules out any trends in relative prices. However, in this paper we discuss chain indices not only from the angle of official statistics but also from the angle of financial markets, where it is commonly considered assumption (see Samuelson (1965), Longstaff & Schwartz (2001), Mansuy (2009)). Moreover, we can use chain indices to construct measures of pension funds' efficiency (Białek (2012, 2013)), where the martingale pricing is one of the theoretical approaches (see for instance Gajek, Kałuszka (2001)).

Proof

We need to show that under the assumption that each process $\{p_i(t): t = 0,1,2,...,T\}$ is a $F-$martingale we would get for any moment t

$$E(\overline{P}_{0,t} / \Im_{t-1}) = \overline{P}_{0,t-1}. \tag{9}$$

Let us notice that from the fact that each $p_i(t)$ and $q_i(t)$ is measurable with respect to $\Im_t$ we conclude that also each stochastic process $P_{0,t}$ is measurable with respect to $\Im_t$. Thus we have

$$E(\overline{P}_{0,t} / \Im_{t-1}) = E(\prod_{\tau=0}^{t-1} P_{\tau,\tau+1} / \Im_{t-1}) = \prod_{\tau=0}^{t-2} P_{\tau,\tau+1} \cdot E(P_{t-1,t} / \Im_{t-1}). \tag{10}$$

From the assumption about *the mean value test* we have that

$$1 = E(\min_{i\in\{1,2,...,N\}} \frac{p_i(t)}{p_i(t-1)} / \Im_{t-1}) \le E(P_{t-1,t} / \Im_{t-1}) \le E(\max_{i\in\{1,2,...,N\}} \frac{p_i(t)}{p_i(t-1)} / \Im_{t-1}) = 1, \tag{11}$$

and hence

$$E(P_{t-1,t} / \Im_{t-1}) = 1. \tag{12}$$

From (10) and (12) we confirm (9), namely

$$E(\overline{P}_{0,t} / \Im_{t-1}) = E(\prod_{\tau=0}^{t-1} P_{\tau,\tau+1} / \Im_{t-1}) = \prod_{\tau=0}^{t-2} P_{\tau,\tau+1} = \overline{P}_{0,t-1}. \tag{13}$$

Remark 1

As it was mentioned above, the postulate 3 should be treated as a fundamental requirement. Let us notice that, by contraposition, if the chain index does not satisfy this postulate then the direct price index (link) does not fulfil *the mean value property*. It is worth adding that according to Pfouts (1966) *the mean value test* is one of the most essential properties of the index function. This fact is in conformity with our intuitive notion of an index to be a measure of a "representative" aggregated change. Moreover, *the mean value test* is included in systems of minimum requirements for price indices (see Eichhorn and Voeller, 1976). The immediate conclusion from the theorem 1 is that all used in practice price indices (like Laspeyres, Paasche, Fisher, Törnqvist, Walsh and other formulas – see Appendix) fulfil the postulate 1.

[1] The mean value test denotes that a value of the price index formula lies between minimum and maximum price relative. For instance, the Laspeyres price index can be expressed as follows: $P^{La}_{\tau,\tau+1} = \sum_{i=1}^{N} v_i^*(\tau) \frac{p_i(\tau+1)}{p_i(\tau)}$ and thus, being a convex combination of partial indices, this index fulfils the mean value property.

Remark 2

In Gajek and Kałuszka (2001) authors propose the stochastic definition of the average rate of return of a group of N open pension funds. Their measure is as follows: $R(0,T)=\prod_{\tau=0}^{T-1}(1+\sum_{i=1}^{N}v_i^*(\tau)r_i(\tau,\tau+1))-1$, where $r_i(\tau,\tau+1)$ denotes the rate of return of $i-$th fund.

The major result of these authors is the theorem which allows one to state that $R(0,T)$ is martingale provided that unit prices are also martingales. It is easy to show (see Białek (2012)) that the measure of Gajek and Kałuszka can be expressed as a Laspeyres chain price index[1] and thus the links satisfy *the mean value test*. In other words, the thesis of the theorem by Gajek and Kałuszka is simply a consequence of the theorem 1.

Theorem 2

If any links $P_{s-1,s}$ and $P_{t-1,t}$ are independent (for $s\neq t$) and each link satisfies *the mean value test* then the chain index $\overline{P}_{0,t}$ fulfils the postulate 1.

Remark 3

The proof of the theorem 2 is quite obvious and it is omitted. The thesis of this theorem is a simple consequence of the known fact that independent links allow one to write

$$E(\overline{P}_{0,t})=E(\prod_{\tau=0}^{t-1}P_{\tau,\tau+1})=\prod_{\tau=0}^{t-1}E(P_{\tau,\tau+1}). \tag{14}$$

Remark 4

Let us notice that for any random variables X and Y we have (provided that the below expected values and standard deviations exist and $P(X=0)=0$)

$$\rho(X,\frac{Y}{X})=\frac{E(Y)-E(X)E(\frac{Y}{X})}{D(X)D(\frac{Y}{X})}, \tag{15}$$

where $\rho(X,Y/X)$ denotes the correlation coefficient between random variables X and Y/X and $D(X)$ denote the standard deviation of X. From (15) we get (if $E(X)\neq 0$)

[1] To read more about connections between measures of funds' efficiency and chain indices see Białek (2013).

$$E(\frac{Y}{X}) = \frac{E(Y)}{E(X)} - \frac{\rho(X,\frac{Y}{X})D(X)D(\frac{Y}{X})}{E(X)} = \frac{E(Y)}{E(X)} - \frac{\mathrm{cov}(X,\frac{Y}{X})}{E(X)}, \qquad (16)$$

where $\mathrm{cov}(X,\frac{Y}{X})$ denotes a covariance between random variables X and Y/X.

Thus, in the case of uncorrelated[1] X and Y/X, we obtain the equality

$$E(\frac{Y}{X}) = \frac{E(Y)}{E(X)}. \qquad (17)$$

The immediate conclusion is the following: if price processes fulfil[2]

$$\mathrm{cov}(p_i(\tau), \frac{p_i(\tau+1)}{p_i(\tau)}) = 0, \qquad (18)$$

then

$$E(\frac{p_i(\tau+1)}{p_i(\tau)}) = \frac{E(p_i(\tau+1))}{E(p_i(\tau))}. \qquad (19)$$

In other words we have the equivalence for each $i = 1,2,....,N$

$$\forall\ \tau = 0,1,...,t-1\ \ E(\frac{p_i(\tau+1)}{p_i(\tau)}) = 1 \ \Leftrightarrow\ \forall\ \tau = 0,1,...,t\ \ E(p_i(\tau)) = const \qquad (20)$$

and it leads to the final conclusion that if the condition (18) holds then the postulates 1 and 2 are equivalent. Moreover, we can formulate the following theorem:

Theorem 3

If the direct price index formula (link) satisfies *the mean value test and the circular test*[3] and, moreover, $p_i(0)$ and $p_i(t)/p_i(0)$ are uncorrelated for each $i \in \{1,2,...,N\}$, then the chain index $\overline{P}_{0,t}$, which is based on this link, satisfies the postulate 2.

Proof

Let us assume, according to the assumptions from the postulate 2, that

$$E(p_i(\tau)) = p_i = const,\ i = 1,2,....,N\ ,\tau = 0,1,...,t. \qquad (21)$$

[1] For instance, such a theoretical situation was considered in Frishman (1971).

[2] It can be quite natural assumption because it requires that prices and relative price changes are uncorrelated.

[3] The circular test denotes that for any moments $s < t < v$ it holds that $P_{s,v} = P_{s,t}P_{t,v}$. The circularity is one of the most restrictive tests in price index theory but often considered in theoretical papers and monographs (see Balk (1995), von der Lippe (2007)).

The satisfied *circular test* leads to the following equality

$$E(\overline{P}_{0,t}) = E(\prod_{\tau=0}^{t-1} P_{\tau,\tau+1}) = E(P_{0,1} \cdot P_{1,2} \cdot ... \cdot P_{t-1,t}) = E(P_{0,t}). \tag{22}$$

The satisfied *the mean value test* leads to the following relation

$$E(\frac{p_n(t)}{p_n(0)}) = E(\min_{i\in\{1,2,...,N\}} \frac{p_i(t)}{p_i(0)}) \le E(P_{0,t}) \le E(\max_{i\in\{1,2,...,N\}} \frac{p_i(t)}{p_i(0)}) = E(\frac{p_m(t)}{p_m(0)}), \tag{23}$$

for some $n, m \in \{1,2,...,N\}$.

From the fact that $\mathrm{cov}(p_n(0), \frac{p_n(t)}{p_n(0)}) = 0$ and $\mathrm{cov}(p_m(0), \frac{p_m(t)}{p_m(0)}) = 0$ we conclude that

$$E(\frac{p_n(t)}{p_n(0)}) = \frac{E(p_n(t))}{E(p_n(0))} = \frac{p_i}{p_i} = 1, \tag{24}$$

and analogically

$$E(\frac{p_m(t)}{p_m(0)}) = \frac{E(p_m(t))}{E(p_m(0))} = \frac{p_i}{p_i} = 1. \tag{25}$$

From (23), (24) and (25) we obtain $E(P_{0,t}) = 1$, which confirms that the postulate 2 is fulfilled.

5. Conclusions

In the paper three stochastic postulates for chain indices are proposed, as an alternative for the classic axiomatic price index theory. The novelty of the presented approach is due to treating the prices and quantities as stochastic processes. The presented postulates have different nature – the postulates 1 and 2 are quite restrictive and we can treat them as some desirable properties but the postulate 3, connected with *the mean value property*, has axiomatic character. Under some additional condition the postulates 1 and 2 are equivalent (see Remark 4). If these postulates are not equivalent we can still show conditions which allow one to fulfil each of the postulates. The most restrictive assumption is in the theorem 3 because it requires *the circularity*. However, there are price index formulas satisfying *the circular test*, like the Walsh price index (see Von der Lippe (2007)). This discussion serves also as a kind of introduction to the author's future research agenda on chain index theory. In our opinion the theorems 1, 2 and 3 are a good starting point because all consideration begins from the basic *proportionality* and *identity*.

REFERENCES

BALK, M., (1995). Axiomatic Price Index Theory: A Survey, International Statistical Review 63, 69−95.

BANARJEE, K. S., (1979). An Interpretation of the Factorial Indexes in the Light of Divisia's Integral Indexes, Statistische Hefte, 20, 261−269.

BIAŁEK, J., (2008). New definition of the average rate of return of a group of pension funds, [in:] Financial Markets: Principles of Modelling, Forecasting and Decision-Making, vol. 6, 126−135, Łódź.

BIAŁEK, J., (2012). The use of statistical chain indices to evaluate the average return of OFE, [in:] Financial investments and insurance - global trends and the Polish market (ed. Krzysztof Jajuga, Wanda Ronka-Chmielowiec), Scientific Papers of the Wrocław University of Economics, 23−32, Wrocław.

BIAŁEK, J., (2013). Measuring Average Rate of Return of Pensions: A Discrete, Stochastic and Continuous Price Index Approaches, International Journal of Statistics and Probability, Vol. 2, No. 4, 56−63.

BOSKIN, M. J., DULBERGER, E. R., GORDON, R. J., GRILICHES, Z., JORGENSON, D., (1996). Toward a More Accurate Measure of the Cost of Living, Final Raport to the Senate Finance Committee from the Advisory Commission to Study the Consumer Price Index.

CHO, D., (2006). A Chain-Type Price Index for New Business Jet Aircraft, Business Economics, vol. 41, 1, 45−52.

CLEMENTS, K. W., IZAN, H. Y., (1987). The Measurement of Inflation: A Stochastic Approach, Journal of Business and Economic Statistics 5, 339−350.

DIVISIA, F., (1925). L'indice montaire et la theorie de la monnaie, Revue d'Economique Politique.

EICHHORN, W., VOELLER, J., (1976). Theory of the Price Index. Fisher's Test Approach and Generalizations, Berlin, Heidelberg, New York: Springer-Verlag.

FEENSTRA, R. C., REINSDORF, M. B., (2000). An exact price index for the almost ideal demand system, Economic Letters 66, 159−162.

FORSYTH, F. G., FOWLER, R. F., (1981). The Theory and Practice of Chain Price Index Numbers, Journal of the Royal Statistical Society, 144, Part 2, 224−246.

FRISHMAN, F., (1971). On the arithmetic means and variances of products and ratios of random variables, [in:] A Modern Course on Statistical Distributions in Scientific Work, Army Research Office, Durham, North Carolina, 330−345 (chapter 8).

GAJEK, L., KAŁUSZKA, M., (2000). On the average return rate for a group of investment funds, Acta Universitas Lodziensis, Folia Oeconomica 152, 161−171, Łódź.

GAJEK, L., KAŁUSZKA, M., (2001). On some properties of the average rate of return – a discrete time stochastic model, (working paper).

HULTEN, C. R., (1973). Divisia Index Numbers, Econometrica 41:6, 1017−1025.

LONGSTAFF, F. A., SCHWARTZ, E. S., (2001). Valuing American options by simulation: a simple least squares approach, Review of Financial Studies 14, 113−148.

MARSHALL, A., (1887). Remedies for Fluctuations of General Prices, Contemporary Review 51, 355−375.

MANSUY, R., (2009). The origins of the Word "Martingale", Electronic Journal for History of Probability and Statistics 5 (1).

PFOUTS, R. W., (1966). An Axiomatic Approach to Index Numbers, Review of the International Statistical Institute, 34 (2), 174−185.

SAMUELSON, P. A., SWAMY, S., (1974). Invariant economic index numbers and canonical duality: survey and synthesis, American Economic Review 64 (4), 566−593.

SAMUELSON, P. A., (1965). Proof That Properly Anticipated Prices Fluctuate Randomly, Industrial Management Review, 6 (2), Spring, 41−49.

SELVANATHAN, E. A., PRASADA RAO, D. S., (1994). Index Numbers: A Stochastic Approach, Ann Arbor: The University of Michigan Press.

SZULC, B. J., (1983). Linking Price Index Numbers, 537−566 in Price Level Measurement, W.E. Diewert and C. Montmarquette (eds.), Ottawa: Statistics Canada.

WILLIAMS, D., (1991). Probability with Martingales, Cambridge University Press.

VILLE, J., (1939). Étude critique de la notion de collectif, Monographies des Probabilités (in French) 3, Paris: Gauthier-Villars.

VOGT, A., (1978). Divisia Indices on Different Paths, In Eichorn, W., Henn, R. and S. Opitz, Hg., Theory and Applications of Economic Index Numbers, Physica: Würzburg, 297−305.

VON DER LIPPE, P., (2001). Chain Indices, A Study in Price Index Theory, Stuttgart: Federal Statistics Office of Germany, vol. 16.

VON DER LIPPE, P., (2007). Index Theory and Price Statistics, Peter Lang, Frankfurt, Germany.

Appendix 1.

For example, according to the thesis of the theorem 1, the following direct price indices (links) guarantee that postulate 1 is satisfied:

- the Laspeyres price index

$$P_{\tau,\tau+1}^{La} = \frac{\sum_{i=1}^{N} q_i(\tau)\, p_i(\tau+1)}{\sum_{i=1}^{N} q_i(\tau)\, p_i(\tau)};$$

- the logarithmic Laspeyres price index

$$P_{\tau,\tau+1}^{LLa} = \prod_{i=1}^{N} \left(\frac{p_i(\tau+1)}{p_i(\tau)}\right)^{v_i^*(\tau)};$$

- the Paasche price index

$$P_{\tau,\tau+1}^{Pa} = \frac{\sum_{i=1}^{N} q_i(\tau+1)\, p_i(\tau+1)}{\sum_{i=1}^{N} q_i(\tau+1)\, p_i(\tau)};$$

- the logarithmic Paasche price index

$$P_{\tau,\tau+1}^{LPa} = \prod_{i=1}^{N} \left(\frac{p_i(\tau+1)}{p_i(\tau)}\right)^{v_i^*(\tau+1)};$$

- the Törnqvist price index

$$P_{\tau,\tau+1}^{T} = \prod_{i=1}^{N} \left(\frac{p_i(\tau+1)}{p_i(\tau)}\right)^{\frac{1}{2}(v_i^*(\tau)+v_i^*(\tau+1))};$$

- the Fisher price index

$$P_{\tau,\tau+1}^{F} = \sqrt{P_{\tau,\tau+1}^{La} P_{\tau,\tau+1}^{Pa}};$$

- the Walsh price index

$$P_{\tau,\tau+1}^{W} = \frac{\sum_{i=1}^{N} p_i(\tau+1)\sqrt{q_i(\tau) q_i(\tau+1)}}{\sum_{i=1}^{N} p_i(\tau)\sqrt{q_i(\tau) q_i(\tau+1)}}.$$

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 559–572

LAG LENGTH SPECIFICATION IN ENGLE-GRANGER COINTEGRATION TEST: A MODIFIED KOYCK MEAN LAG APPROACH BASED ON PARTIAL CORRELATION

Oluokun Kasali Agunloye[1], **Dahud Kehinde Shangodoyin**[2], **Raghunath Arnab**[2]

ABSTRACT

The Engle-Granger cointegration test is highly sensitive to the choice of lag length and the poor performance of conventional lag selection criteria such as standard information criteria in selecting appropriate optimal lag length for the implementation of the Engle-Granger cointegration test is well-established in the statistical literature. Testing for cointegration within the framework of the residual-based Engle-Granger cointegration methodology is the same as testing for the stationarity of the residual series via the augmented Dickey-Fuller test which is well known to be sensitive to the choice of lag length. Given an array of candidate optimal lag lengths that may be suggested by different standard information criteria, the applied researchers are faced with the problem of deciding the best optimal lag among the candidate optimal lag lengths suggested by different standard information criteria, which are often either underestimated or overestimated. In an attempt to address this well-known major pitfall of standard information criteria, this paper introduces a new lag selection criterion called a modified Koyck mean lag approach based on partial correlation criterion for the selection of optimal lag length for the residual-based Engle-Granger cointegration test. Based on empirical findings, it was observed that in some instances over-specification of lag length can bias the Engle-Granger cointegration test towards the rejection of a true cointegration relationship and non-rejection of a spurious cointegration relationship. Using real-life data, we present an empirical illustration which demonstrates that our proposed criterion outperformed the standard information criteria in selecting appropriate optimal truncation lag for the implementation of the Engle-Granger cointegration test using both augmented Dickey-Fuller and generalized least squares Dickey-Fuller tests.

Key words: modified Koyck mean lag, partial correlation criterion, Engle-Granger cointegration test, optimal truncation lag, information criteria, augmented Dickey-Fuller test, generalized least square Dickey-Fuller test.

[1] Department of Statistics, University of Botswana, Gaborone, Botswana.
E-mail: k.agunloye@yahoo.com.
[2] Department of Statistics, University of Botswana, Gaborone, Botswana.

1. Introduction

The residual-based Engle-Granger cointegration methodology is arguably the most widely used bivariate cointegration test in empirical analysis. One of the major specification decisions that poses a big challenge to analysts and applied researchers is selection of appropriate lag length for the implementation of unit root test for the estimated residuals from cointegrating regressions. A number of previous studies have demonstrated a strong influence of lag selection on the outcome of the Engle-Granger cointegration test. Gutierrez et al. (2009) show that misspecification of appropriate lag length may greatly affect the cointegration results such that under-specification of lag length could invalidate the cointegration test and over-specification of lag length could result in a loss of power. Hall (1991) pointed out that the choice of lag structure in the error correction model (ECM) is a vital specification decision because too few lags may lead to serial correlation problem, whereas too many lags specified in the ECM will consume more degree of freedoms leading to small sample problem. Li et al. (2009) also corroborated Hall (1991) position by arguing that appropriate specification of lag length is one of the most important specification decisions concerning implementation of the error correction process. Johansen (1991) proposed the use of appropriate information criterion or a sequence of likelihood ratio tests for the determination of lag length.

This paper is primarily concerned with appropriate specification of lag length for the cointegration test as well as the error correction process (ECP) within the context of the Engle-Granger cointegration methodology. Standard information criteria such as Akaike Information Criterion (AIC), Akaike Final Prediction Error (FPE), the Bayesian Information Criterion (BIC) and Hannan-Quinn Information Criterion (HQIC) that are commonly employed for the choice of optimal lag structure have been shown to exhibit a strong tendency to either over-specify or under-specify the lag length. Nishi (1988) and Lutkepohl (1993) showed that both Akaike Information Criterion (AIC) and Final Prediction Error (FPE) are not consistent estimators of the truncation lag order but the Bayesian Information Criterion (BIC) is strongly consistent. Bewley and Yang (1998) evaluated the performance of standard information criteria such as AIC and BIC in selecting appropriate lag structure for the cointegration test and showed that these conventional lag selection criteria appear to have problem of underestimation and overestimation of the lag structure. Clarke and Mirza (2006) argue that both AIC and FPE cannot be recommended as lag selection procedures since both criteria are well known to have a positive probability of overestimating the true lag order.

In general, the major drawback of the commonly used standard information criteria lies with problem of underestimation and overestimation of lag length which are regarded as undesirable in cointegration analysis as demonstrated in Cheung and Lai (1993) and Gonzalo (1994). Given this demonstrated weaknesses of the standard information criteria, we therefore propose an alternative lag

selection criterion called a modified Koyck mean lag approach based on partial correlation criterion (MK-PCC) for the purpose of lag specification in the residual-based Engle-Granger cointegration test proposed by Engle and Granger (1987).

The remaining part of this paper is organized as follows. Section 2 discusses specification of augmented Dickey-Fuller (ADF) and generalized least squares Dickey-Fuller (DF-GLS) tests for the implementation of the residual-based Engle-Granger cointegration test. Section 3 introduces lag specification procedure based on the modified Koyck mean lag approach using partial correlation criterion. Section 4 presents preliminary data description and unit root tests. Section 5 discusses the Engle-Granger cointegration tests, residual analysis and estimation of error correction models. Finally, section 6 concludes.

2. Specification of Engle-Granger cointegration test

Consider two non-stationary time series variables that are integrated of the same order, say order 1, $I(1)$ variables. Following Engle and Granger (1987), two variables, say x and y are said to be cointegrated of order $CI(1,1)$ if there exists a long-run equilibrium relationship between the two integrated variables such that the residuals of the estimated regression are stationary or integrated of order zero , $I(0)$.

The long-run equilibrium relationship is captured by the following regression models:

$$y_t = \alpha_0 + \alpha_1 x_t + w_t \tag{1}$$

$$x_t = \beta_0 + \beta_1 y_t + u_t \tag{2}$$

where x and y are $I(1)$ variables, α_0, α_1, β_0 and β_1 are cointegrating parameters, w_t and u_t are OLS residuals which capture divergences between the variables from an assumed equilibrium long-run relationship.

The use of the Engle-Granger (EG) cointegration methodology requires pair-wise comparison of two cointegrating regressions because the EG method produces just only one cointegrating vector. We distinguish between the pair of cointegration regressions (1) and (2) above because unlike Johansen cointegration methodology, the Engle-Granger cointegration procedure is sensitive to the choice of dependent variable (see Dickey et al., 1991). Testing for the presence of cointegration in the context of the bivariate Engle-Granger cointegration test is essentially equivalent to testing for the presence of a unit root in the estimated residual series $\{\hat{u}_t\}$ and $\{\hat{w}_t\}$ for the

cointegrating regressions (1) and (2) where the Engle-Granger (EG) tests (which are akin to the standard Dickey-Fuller tests) used for testing the stationarity of the residuals are specified as follows:

$$\Delta\hat{u}_t = \rho_1\hat{u}_{t-1} + \varepsilon_t \tag{3}$$

$$\Delta\hat{w}_t = \rho_2\hat{w}_{t-1} + \in_t \tag{4}$$

The first difference of the residuals is regressed on the lagged level of the residuals without a constant, where ρ_1 and ρ_2 are parameters of interest representing the slope of each line, $\Delta\hat{u}_t$ and $\Delta\hat{w}_t$ are the first difference of the estimated residual series $\{\hat{u}_t\}$ and $\{\hat{w}_t\}$ respectively, $\hat{u}_{t-1}$ and $\hat{w}_{t-1}$ are the estimated lagged residuals, ε_t and $\in_t$ are error terms which are expected to be serially uncorrelated. Equations (3) and (4) do not include intercept terms because the estimated residual series $\{\hat{u}_t\}$ and $\{\hat{w}_t\}$ are obtained from regression equations (1) and (2) respectively. The EG test requires that error terms be serially uncorrelated. Due to the problem of serial correlation in standard EG test, it is a common practice to use the augmented Engle-Granger (AEG) test which accommodates more lags of the first difference of the residuals to eliminate the serial correlation problem that is associated with standard EG test. The corresponding AEG tests for (3) and (4) are specified as follows:

$$\Delta\hat{u}_t = \rho_1\hat{u}_{t-1} + \sum_{i=0}^{p}\xi_i\Delta\hat{u}_{t-i} + \varepsilon_t \tag{5}$$

$$\Delta\hat{w}_t = \rho_2\hat{w}_{t-1} + \sum_{j=0}^{q}\Omega_j\Delta\hat{w}_{t-j} + \in_t \tag{6}$$

where ρ_1 and ρ_2 are parameters, ξ_i and Ω_j are coefficients of lagged difference of the estimated residuals, $\Delta\hat{u}_t$ and $\Delta\hat{w}_t$ are first difference of the estimated residual series $\{\hat{u}_t\}$ and $\{\hat{w}_t\}$ respectively, $\hat{u}_{t-i}$ and $\hat{w}_{t-j}$ are lags of the estimated residuals, ε_t and $\in_t$ are error terms, p and q are optimal truncation lag parameters to be determined to whiten the error terms. AEG test can be utilized to perform unit root test on the estimated coefficients ρ_1 and ρ_2 individually to establish the existence or non-existence of long-run equilibrium relationship. Any unit root test involving ADF is sensitive to the choice of lag length which is the number of lagged differences with which the regression is augmented. Since AEG test is a modification of ADF test, it also inherits the lag

selection problem that is commonly associated with ADF test due to its sensitivity to the choice of lag length. The main criticism of the Augmented Dickey-Fuller (ADF) test is that the power of the test is very low if the time series under test is nearly non-stationary which implies that the time series is stationary but with a root close to 1 (see Brooks 2002). The focus of our present study is to employ the modified Koyck mean lag approach based on partial correlation criterion (MK-PCC) for lag selection required for the implementation of AEG tests since enough lags need to be chosen for the error terms ε_t and $\in_t$ to be serially uncorrelated. In applying the MK-PCC, we consider a distributed lag re-parameterization of the augmented Engle-Granger (AEG) tests as follows:

CASE 1: When y is the dependent variable for the cointegrating regression, we have the following representation:

$$y^{(*)} = \Delta\hat{u}_t - \rho_1\hat{u}_{t-1} = \sum_{i=0}^{p}\xi_i\Delta u_{t-i} \tag{7}$$

CASE 2: When x is the dependent variable for the cointegrating regression, we have the following representation:

$$x^{(*)} = \Delta\hat{w}_t - \rho_2\hat{w}_{t-1} = \sum_{j=0}^{q}\Omega_j\Delta\hat{w}_{t-j} \tag{8}$$

Using generalized least squares Dickey-Fuller (DF-GLS) test as an alternative unit root test to ADF, we repeat the same distributed lag re-parameterization for the DF-GLS test as follows:

CASE 1: When y is the dependent variable for the cointegrating regression, we have the following representation:

$$y^{(*)} = \Delta\hat{u}_t^d - \rho_1\hat{u}_t^d = \sum_{i=0}^{p}\xi_i\Delta u_{t-i}^d \tag{9}$$

CASE 2: When x is the dependent variable for the cointegrating regression, we have the following representation:

$$x^{(*)} = \Delta\hat{w}_t^d - \rho_2\hat{w}_{t-1}^d = \sum_{j=0}^{q}\Omega_j\Delta\hat{w}_{t-j}^d \tag{10}$$

Interpretation of notations is the same as earlier given above except that the residual series are subjected to generalized least squares detrending.

3. Modified Koyck mean lag approach based on partial correlation criterion for lag selection (MK-PCC)

Following Koyck (1954) mean lag model, we can assume the Koyck postulations as follows

$$\bar{L}_{(i)} = \frac{R_{(i)}}{1-R_{(i)}},\ i = 1,...,4 \tag{11}$$

where $\bar{L}_{(i)}$ is the mean lag for a particular unit root test, $R_{(i)}$ is the partial correlation coefficient computed for each of the model in equation (7) through equation (10) between $y^{(*)}$ and lagged differences $\Delta y_{t-1}, \Delta y_{t-2}, \ldots\ldots, \Delta y_{t-12}$ (in case of monthly dataset) and it measures the rate at which $y^{(*)}$ depends on these lagged differences. The main idea of MK-PCC is based on fitting simple linear regression model to the left-hand side of equation (7) through equation (10) to generate the parameters needed and to compute the partial correlation between the parameter on the left-hand side of equation (7) through equation (10) and different choices of lagged differences from the set of lagged differences $\Delta y_{t-1}, \Delta y_{t-2}, \ldots\ldots, \Delta y_{t-12}$ on the right-hand side of equation (7) through equation (10) while controlling for the effects of other remaining lagged differences. For the first computation we compute partial correlation between $y^{(*)}$ and Δy_{t-1} while controlling for $\Delta y_{t-2}, \ldots\ldots, \Delta y_{t-12}$. For the second computation we compute partial correlation between $y^{(*)}$ and the first two lagged differences (i.e. Δy_{t-1} Δy_{t-2}) while controlling for $\Delta y_{t-3}, \ldots\ldots, \Delta y_{t-12}$ and so on like that. We also repeat the same procedure for other specification of unit root tests as shown above. The partial correlation coefficient denoted by $R_{(i)}$ is computed and adjusted for maximum lag until it gives a values less than 0.3 which is equivalent to lag 0 since for $R_{(i)} < 0.3$, the mean lag will be assumed to be zero since the mean lag specified by $\bar{L}_{(i)} = \frac{R_{(i)}}{1-R_{(i)}}$ for which $R_{(i)} < 0.3$ is a fraction not up to 0.5. It should be noted that to have a reasonable mean lag length we expect the absolute value of $R_{(i)}$ to be in the interval $[0.5, 0.999)$ (see Agunloye et al., 2013).The same procedure is repeated for DF-GLS test by fitting simple linear regression model to the left-hand side of equations (9) and (10) to generate the parameters needed and to compute the partial correlation between the parameter on the left-hand side of equation (9) through equation (10) as earlier explained above for ADF test.

As indicated earlier in the introductory part of this paper, the residual-based Engle-Granger cointegration test is very sensitive to the choice of truncation lag parameters p and q. The problem of bias in cointegration is due to misspecification of lag length. Since the Engle-Granger cointegration test is equivalent to testing for the presence of unit root in the estimated residuals from the cointegrating regression it also shares the problem of low power that is commonly associated with unit root test when the estimated residual is closed to being a unit root process but not exactly a unit root process. For the purpose of the present study, we consider a situation when the estimated parameters of interest (i.e. ρ_1 and ρ_2) assume any of the following values: 0.9, 0.95 and 0.999 in equations (7), (8), (9) and (10) respectively. Our choice of these parameter values is informed due to the fact that the power of test for Augmented Dickey-Fuller (ADF) is very low if the process is nearly non-stationary, which means the process is stationary but with a root close to the non-stationary boundary (Brooks 2002).

4. Data description and unit root test

For empirical analysis, we use two sets of data. One real dataset and one simulated dataset. The real dataset are US 3-Month Treasury Bills (USMTB) for short-term money market interest rate series and US 10-Month Government Security (USMGS) for long-term interest rate series. The data cover the period from January 1962 through February 2014 and are obtained from IMF Monthly Bulletin. A total of 626 observations are collected for USMGS and USMTB series respectively.

This paper adopts the residual-based Engle-Granger (EG) cointegration test for empirical analysis. The implementation of EG methodology is carried out in two steps. The first step tests for the order of integration of time series variables. The order of integration of a variable is the number of times a variable is required to be differenced to attain stationarity. A condition applicable to EG test is that the variables entering the cointegrating equation should be integrated of the same order which is assumed to be order 1 in the context of EG test. To test for degree of integration of the USMGS and USMTB series two well-known tests are used in this paper. The first test is the Augmented Dickey-Fuller (ADF) (1984) test and the second test is the generalized least squares Dickey-Fuller (DF-GLS) test introduced by Elliot et al. (1996). The optimal lag length were determined using five lag length selection criteria comprising four conventional criteria and newly introduced criterion called MK-PCC. The results for the unit root tests are presented in tables 1 through table 4 below:

Table 1. Summary of results for ADF Unit root test for both series at level

	AIC	FPE	BIC	HQIC	MK-PCC
USMGS	-1.543344(3)	-1.543344(3)	-1.435517(2)	-1.435517 (3)	-1.341465 (0)
USMTB	-2.577255(3)	-2.577255 (3)	-2.577255 (2)	-2.577255 (3)	-2.299976(0)

The null hypothesis of unit root is rejected if the test statistic is less than the 5% critical value

Table 2. Summary of results for DF-GLS unit root test for both series at level

	AIC	FPE	BIC	HQIC	MK-PCC
USMGS	-0.780907(3)	-0.780907 (3)	-0.684594 (2)	-0.684594 (3)	-0.609149(0)
USMTB	-1.649786(3)	-1.649786 (3)	-1.649786 (2)	-1.649786 (3)	-1.406009 (0)

The null hypothesis of unit root is rejected if the test statistic is less than the 5% critical value

Tables 1 and 2 present the results of unit root tests for the level of the two series under investigation using ADF and DF-GLS tests. The ADF test-statistic under different optimal lag lengths is greater than the critical value at 5% level of significance which is -3.417060. Similarly, the DF-GLS test-statistic under different optimal lag lengths is also greater than the critical value at 5% level of significance which is -2.890000. Consequently, we fail to reject the null hypotheses of unit root for the level of the two series. This implies that each of the series is non-stationary at level. In contrast to standard information criteria which had to fit higher lags such as lag 2 or lag 3 in order to establish non-stationarity of both series at levels, MK-PCC lag selection methodology established non-stationarity of both series without fitting any lag.

Table 3. Summary of results for ADF unit root test for both series after first difference

	AIC	FPE	BIC	HQIC	MK-PCC
$\nabla USMGS$	-12.88040(3)	-16.88413 (3)	-12.17434 (2)	-17.17434 (3)	-16.88413 (0)
$\nabla USMTB$	-17.93451(3)	-17.61138 (3)	-17.93451 (2)	-17.93451 (3)	-17.61138 (0)

The null hypothesis of unit root is rejected if the test statistic is less than the 5% critical value

Table 4. Summary of results for DF-GLS unit root test for both series after first difference

	AIC	FPE	BIC	HQIC	MK-PCC
$\nabla USMGS$	-5.409803(3)	-5.409803 (3)	-6.478566 (2)	-5.409803 (3)	-10.67095 (0)
$\nabla USMTB$	-17.95375(3)	-17.63360(3)	-17.95375(2)	-17.95375 (3)	-17.63360 (0)

The null hypothesis of unit root is rejected if the test statistic is less than the 5% critical value

Tables 3 and 4 present the results of unit root tests for the first difference of the two series under investigation using ADF and DF-GLS tests. The ADF test-statistic under different optimal lag lengths is less than the critical value at 5% level of significance which is -3.417060. Similarly, the DF-GLS test-statistic under different optimal lag lengths is also less than the critical value at 5% level of significance which is -2.890000. Consequently, we reject the null hypotheses of unit root for the two series at first difference. This implies that each series is integrated of order 1 since they become stationary after first difference. The empirical results shown in tables 3 and 4 above show that while stationarity of the first difference of both series was achieved at lag zero under MK-PCC lag selection methodology, the standard information criteria had to fit higher lags such as lag 2 or lag 3 in order to achieve the same results.

5. Engle-Granger cointegration test

We fit autoregressive models of order 1 to 12 to the residuals of the cointegrating regressions and the various optimal lag lengths suggested by different lag selection criteria are presented in brackets in table 5 below. The ADF and DF-GLS unit root tests are performed on the residuals from OLS estimation for USMGS and USMTB pairs. All regressions reported are cointegrated at the 5 per cent level. This suggests that the estimated equations reflect a stable long-run relationships.

Table 5. Engle-Granger cointegration test using ADF test

VARIABLE	AIC	FPE	BIC	HQIC	MK-PCC
USMGS-USMTB RESIDUAL	-3.6199(5)	-3.6199(5)	-3.7089(4)	-3.6199(5)	-3.4822(0)
USMTB-USMGS RESIDUAL	-4.7785(10)	-4.5175(10)	-4.6322(4)	-4.2392(4)	-3.6883(0)

The null hypothesis of "no cointegration" is rejected if the test statistic exceeds the 5% critical value.

Table 5 presents the results of the Engle-Granger cointegration test using ADF unit root test for the stationarity of residuals from each regression equation. For cointegrating regression with USMGS as dependent variable, it is observed that the test statistic for the ADF version of the Augmented Engle-Granger (AEG) test at different optimal lag lengths suggested by conventional lag selection criteria and MK-PCC criterion exceeds the critical value at 5% level of significance. Consequently, we reject the null hypotheses of "no cointegration" at these various optimal lags. This implies that USMGS and USMTB series are cointegrated at these optimal lags. However, for cointegrating regression with USMGS as dependent variable, the test statistic for the ADF version of the Augmented Engle-Granger (AEG) test at different optimal lag lengths suggested conventional lag selection criteria is less than the critical value at 5% level of

significance except for MK-PCC for which the test statistic exceeds the critical value. Hence, we fail to reject the null hypothesis of "no cointegration" under optimal lags suggested by AIC, FPE, BIC and HQIC respectively indicating that USMTB and USMGS are not cointegrated at lag 10 and lag 4 that were suggested by standard information criteria but are cointegrated at lag 0 selected by MK-PCC.

Table 6. Engle-Granger cointegration test using DF-GLS test

VARIABLE	AIC	FPE	BIC	HQIC	MK-PCC
USMGS-USMTB RESIDUAL	-3.6722(5)	-3.6722(5)	-4.0692(4)	-3.6722(5)	-3.4563(0)
USMTB-USMGS RESIDUAL	-4.6524(10)	-4.6967(10)	-4.5326(4)	-4.5326(4)	-3.6005(0)

The null hypothesis of "no cointegration" is rejected if the test statistic exceeds the 5% critical value.

Table 6 presents the results of the Engle-Granger cointegration test using DF-GLS unit root test for the stationarity of residuals from each regression equation. For cointegrating regression with USMGS as dependent variable, it is observed that the test statistic for the DF-GLS version of the Augmented Engle-Granger (AEG) test at different optimal lag lengths suggested by conventional lag selection criteria and MK-PCC criterion exceeds the critical value at 5% level of significance except for BIC which suggested optimal lag 4 for which the test statistic is less than critical value. Consequently, we reject the null hypotheses of "no cointegration" at these various optimal lags. This implies that USMGS and USMTB series are cointegrated under optimal lags suggested by AIC, FPE, HQIC and MK-PCC respectively but they are not cointegrated at lag 4 suggested by BIC. However, for cointegrating regression with USMTB as dependent variable, the test statistic for the DF-GLS version of the Augmented Engle-Granger (AEG) test at different optimal lag lengths suggested conventional lag selection criteria is less than the critical value at 5% level of significance except for MK-PCC for which the test statistic exceeds the critical value. Hence, we fail to reject the null hypothesis of "no cointegration"under the optimal lags suggested by AIC, FPE, BIC and HQIC respectively indicating that USMTB and USMGS are not cointegrated at lag 10 and 4 that were suggested by these standard information criteria but are cointegrated at lag 0 selected by MK-PCC.

5.1. Estimation of Engle-Granger error correction model

Following Engle and Granger (1987), we specify error correction model for the cointegrating relationship between USMGS and USMGTB as follows:

$$\Delta\left(usmgs\right)_t = \tau_0 + \sum_{i=1}^{p_1} \gamma_i \Delta\left(usmgs\right)_{t-i} + \sum_{j=1}^{q_1} \lambda_j \Delta\left(usmtb\right)_{t-j} + \alpha_1 \hat{u}_{t-1} + \varepsilon_t \qquad (12)$$

$$\Delta\left(usmtb\right)_t = \delta_0 + \sum_{i=1}^{p_2} \phi_i \Delta\left(usmtb\right)_{t-i} + \sum_{j=1}^{q_2} \Phi_j \Delta\left(usmgs\right)_{t-j} + \alpha_2 \hat{w}_{t-1} + \in_t \qquad (13)$$

where α_1 and α_2 are adjustment coefficients, p_1, q_1, p_2 and q_2 are the optimal lags required to whiten the error terms in (12) and (13) respectively. In equation (12), USMGS is taken as dependent variable and USMTB is explanatory variable. Similarly in equation (13), USMTB is taken as dependent variable and USMGS is taken as explanatory variable. However, in order for valid inferences to be made from ECM models specified in (12) to (13) above, it is necessary that the coefficients of the lagged residuals represented by α_1 and α_2 , which serve as the "speed of adjustment parameters", are significant and their coefficients are negative. Mathematically, deviations from long-run equilibrium relationship between two variables can only be corrected if our cointegrating vector is negative. The value of adjustment parameter is a crucial parameter of interest that is expected to be less than 1 in absolute terms to guarantee the stability of the system and for the variables in the long-run relationship to be cointegrated. The number of lags to be included in the ECM equations is determined by the number of lags required to whiten the error terms. The ECM models constructed for USMGS and USMTB series were both valid based on the aforementioned criteria.

5.2. Residual analysis

Prior to estimation of the Engle-Granger error correction model, a crucial issue is whether the error terms are uncorrelated, homoscedastic and normally distributed. Residual analysis was conducted using Breusch-Godfrey LM test for serial correlation, ARCH-LM for heteroskedasticity and Jarque-Bera for normality test. The appropriate number of lags is 2 which is the optimal lag order required to whiten the error term. Bivariate analysis showed that both pairs of USMGS and USGMTB were cointegrated at 5% significance levels. The results of the diagnostic tests on residuals are presented in table 7 below.

Table 7. Summary of results of diagnostic tests on residuals

Tests	Test Statistic	p-value	Conclusion
Jarque-Bera	21.18518	0.000025	Normally distributed
ARCH-LM	972.3744	0.0000	No Heteroskedaticity
Breusch-Godfrey LM test	4194.212	0.0000	No Serial Correlation

The p-values in table above are compared with 0.05 significance level.

Table 8. The Engle-Granger Error Correction Model Estimates for USMGS- USMTB Pair

	Coefficient	t-value	Probability
$(USMGS)_{t-1}$	0.396037	8.42717	0.04700
$(USMGS)_{t-2}$	-0.267194	-5.60852	0.04764
$(USMTB)_{t-1}$	-0.031421	-1.00462*	0.03128
$(USMTB)_{t-2}$	0.050940	1.62607*	0.03133
Residual	-0.023292	-2.53590*	0.00918
Constant	-0.001496	-0.13809	0.01083
R^2	0.151445		
$Adj.R^2$	0.144557		
Sum of Squares Residual	44.95889		
S.E Equation	0.270158		
F-statistic	21.98793		
AIC	0.229978		
BIC	0.272740		

*indicates significance at 5% level

Table 8 presents the empirical result from the short-run dynamics based on the Engle-Granger error correction model when USMGS is taken as dependent variable in the cointegrating regression. In estimating this ECM model, two lags for the explanatory variable were found to be sufficient to whiten the residuals. In the Engle-Granger cointegration methodology, the coefficient of the lagged residual shown in table 8 is of particular interest because it represents the speed of adjustment as well as stability of the system. The absolute value of the coefficient is 0.023292 which is less than 1 indicating that the system is stable. However, the coefficient is quite small which indicates that about 2.3292% of any deviation from the long-run path is corrected within a month which translates into about 27.95% adjustment per year.

Table 9. The Engle-Granger Error Correction Model Estimates for USMTB-USMGS Pair

	Coefficient	t-value	Probability
$(USMTB)_{t-1}$	0.308307	6.53244	0.04720
$(USMTB)_{t-2}$	-0.115712	-2.44775	0.04727
$(USMGS)_{t-1}$	0.321225	4.52965*	0.04092
$(USMGS)_{t-2}$	-0.213002	-2.96287*	0.04189
Residual	-0.023995	-1.78542*	0.01344
Constant	-0.003360	-0.20555	0.01635

Table 9. The Engle-Granger Error Correction Model Estimates for USMTB-USMGS Pair (cont.)

	Coefficient	t-value	Probability
R^2	0.189927		
$Adj.R^2$	0.183352		
Sum of Squares Residual	102.3760		
S.E Equation	0.407670		
F-statistic	28.88513		
AIC	1.052882		
BIC	1.095644		

*indicates significance at 5% level.

Table 9 presents the empirical result from the short-run dynamics based on the Engle-Granger error correction model when USMTB is taken as dependent variable in the cointegrating regression. In estimating this ECM model, two lags for the explanatory variable were also found to be sufficient to whiten the residuals. In the Engle-Granger cointegration methodology, the coefficient of the lagged residual shown in table 9 above is of particular interest because it represents the speed of adjustment as well as stability of the system. The absolute value of the coefficient is 0.023995 which is less than 1 indicating that the system is stable. However, the coefficient is quite small which indicates that about 2.3995% of any deviation from the long-run path is corrected within a month which translates into about 28.79% adjustment per year.

6. Conclusion

This paper examined the problem of lag length selection within the framework of the Engle-Granger cointegration test. We demonstrated that the conventional lag selection criteria such as AIC, FPE, BIC and HQIC standard information criteria have the problem of over-specification of lag length. We introduced a new criterion called the modified Koyck mean lag approach based on partial correlation criterion (MK-PCC) which outperforms conventional standard information criteria by avoiding over-specification of lag length commonly associated with frequently used conventional lag selection criteria.

REFERENCES

AGUNLOYE, O. K., ARNAB, R., SHANGODOYIN, D. K., (2013). A New Criterion for Lag-Length Selection in Unit Root Tests. American Journal for Theoretical and Applied Statistics, Vol. 2, No. 6, 293–298.

AKAIKE, H., (1969). Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics, 21, 243–247.

AKAIKE, H., (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory. In: B.N. Petrov and F. Csáki, (Eds) (Budapest: Académiai Kiadó), 267–281.

BEWLEY, R., YANG, M., (1998). On the size and power of system tests for cointegration. The Review of Economics and Statistics, 80(4), 675–679.

BROOKS, C., (2002). Introductory Econometrics for Finance, Cambridge University Press.

CHEUNG, Y. W., LAI, K. S., (1993). Finite sample sizes of Johansen's likelihood ratio test for cointegration. Oxford Bulletin of Economics and Statistics, 55, 313–328.

CLARKE, J. A., MIRZA, S., (2005). A comparison of some common methods for detecting Granger noncausality. Journal of Statistical Computation and Simulation in press.

DICKEY, D. A., JANSEN, D. W., THORNTON, D. L., (1991). A Primer on Cointegration with Application to Money and Income. Review, Federal Reserve Bank of St. Louis, issue Mar., 58–78.

ELLIOTT, G., ROTHENBERG, T. J., STOCK, J. H., (1996). Efficient Tests for an Autoregressive Unit Root. Econometrica, 64, 4, 813–836.

ENGLE, R. F., GRANGER, C. W. J., (1987). Cointegration and Error Correction Representation, Estimation and Testing. Econometrica, 55: 251–257.

GONZALO, J., PITARAKIS, J. Y., (1999). Lag length estimation in large dimensional systems. Journal of Time Series Analysis, 23(4), 401–423.

GUTIERREZ, C. E. C., SOUZA, R. C., GUILLEN, O. T. D. C., (2009). Selection of Optimal Lag-length in Cointegrated VAR models with Weak Form of Common Cyclical Features. Brazilian Review of Econometrics, Vol. 29, No. 1, 59–78.

HANNAN, E. J., QUINN, B. G., (1978). "The determination of the order of an autoregression". Journal of Royal Statistical Society, 41, 190–195.

JOHANSEN, S., (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica, 59(6), 1551–1580.

KOYCK, L. M., (1954). Distributed Lags and Investment Analysis, Amsterdam: North-Holland.

LI, J., MOORADIAN, R. M., YANG, S. X., (2009). "The Information Content of the NCREIF Index", Journal of Real Estate Research, 31(1), 93–116.

LÜTKEPOHL, H., (1993). Introduction to Multiple Time Series Analysis (2nd ed.) (Berlin: Springer-Verlag).

NISHI, R., (1988). Maximum likelihood principle and model selection when the true model is unspecified. Journal of Multivariate Analysis, 27, 392–403.

SAID, E. S., DICKEY, D. A., (1984). Testing for a Unit Root in Autoregressive Moving Average Models of Unknown Order. Biometrika, 71, 3, 599–607.

SCHWARZ, G., (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461–464.

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 573–590

VARIABILITY OF HOUSEHOLD DISPOSABLE INCOME *PER CAPITA* BY TYPES OF RESIDENCE IN POLAND

Anna Turczak[1], **Patrycja Zwiech**[2]

ABSTRACT

The dispersion of household disposable income *per capita* in each class of residence (i.e. six) was estimated for households in Poland. Then, the dispersion of income between the classes was analysed. The computation was carried out separately for subsequent years from 1998 to 2012. The study shows that the households in Poland are differentiated with regard to income *per capita* by types of residence, however, the differences within the groups are much bigger than the differences between the groups. What is particularly surprising, the share of between-group variance in total variance in the population under study was negligible small (just a few percent) compared to the share of the mean within-group variance (more than 90 percent).

Key words: disposable income *per capita*, type of residence, within-group and between-group variance.

1. Introduction

The analysis of diversification in household disposable income *per capita* is a significant study area as it helps to understand the inhomogeneous nature of living standard within a certain social group. Undoubtedly, the income level is a key variable varying the living standard of Polish residents. The aim of this article is to estimate the differences in available income *per capita* across households in various classes of residence, as compared to the variation of household's disposable income *per capita* within classes. The nature of this article is the research one.

This article describes separately six classes of residence (hereinafter referred to as classes or groups):

- cities with 500,000 residents and more (on 24th July 2014 were in Poland 5 such cities);

[1] West Pomeranian Business School, Poland. E-mail: aturczak@zpsb.szczecin.pl.
[2] University of Szczecin, Poland. E-mail: patrycjazwiech@tlen.pl.

- cities with 200,000 to 499,999 residents (on 24th July 2014 were in Poland 12 such cities);
- towns with 100,000 to 199,999 residents (on 24th July 2014 were in Poland 23 such towns);
- towns with 20,000 to 99,999 residents (on 24th July 2014 were in Poland 183 such towns);
- towns with less than 20,000 residents (on 24th July 2014 were in Poland 691 such towns);
- and villages (on 1st January 2015 were in Poland 43,068 villages).

The above division is disjoint and exhaustive.

In order to meet the article objective, three research tasks were determined, namely:

1) to compare mean household disposable income *per capita* across certain classes with mean household disposable income *per capita* in Poland;
2) to compare the dispersion of household disposable income *per capita* within certain groups with the income dispersion in Poland;
3) to analyse the between-group variance against the mean within-group variance.

The analysis was carried out separately for each year from 1998 to 2012. The data come from the Household Budget Surveys (HBS) which are conducted annually by the Central Statistical Office of Poland, on a regular basis. The HBS data for the period from 1998 to 2012 were provided by the Central Statistical Office of Poland (GUS) pursuant to Contract No. 20/Z/DI-6-611/632/2013/RM concluded between GUS and the University of Szczecin. The said database includes detailed information on 31,756 Polish households in 1998, 31,428 in 1999, 36,163 in 2000, 31,847 in 2001, 32,342 in 2002, 32,452 in 2003, 32,214 in 2004, 34,767 in 2005, 37,508 in 2006, 37,366 in 2007, 37,358 in 2008, 37,302 in 2009, 37,412 in 2010, 37,375 in 2011 and 37,427 in 2012. The household budget survey was carried out by the Central Statistical Office of Poland with the use of representative method which makes it possible to generalise the results to all the households in Poland (*Budżety*... 2012, 2013, p. 13).

This article tests two research hypotheses. The first one states the highest mean household disposable income *per capita* in Poland is recorded in big cities and the less residents in a town there are, the lower the mean household disposable income *per capita* may be observed. But the lowest household disposable income is typical of villages. The second research hypothesis to be verified in this article states the variation of household disposable income *per capita* within classes of residence (i.e. groups) is significantly higher than the between-group variation.

2. Applied research tools

The subject of analysis in this article is the quantitative characteristic *X* which is household disposable income *per capita*. Household disposable income is defined as a sum of household current incomes from various sources reduced by prepayments on personal income tax made on behalf of a tax payer by a tax-remitter, by tax on income from property, taxes paid by self-employed persons and by social security and health insurance premiums. The disposable income covers both income in cash and in kind, including natural consumption (consumer goods and services taken to satisfy household's needs from self-employment – in and outside farming) as well as goods and services received free of charge. Disposable income is allocated to expenditures and savings increase (*Budżety…* 2012, 2013, p. 18).

The study concerns the distribution of the said variable *X* within the examined statistical population. First of all, in order to describe the structure, the analysis of central tendency was carried out with the use of such a classical measure as the arithmetic mean. Let the mean value of variable *X* be denoted by $\bar{x}$. On the other hand, to analyse the differences between individual observations of variable *X*, the variance will be applied as the classical measure of dispersion. The variance of variable *X* is denoted by $S^2(x)$. The variance is expressed in square units of the examined variable and is not interpreted (Pułaska-Turyna, 2005, p. 71). It is always non-negative (Bielecka, 2001, p. 134).

Standard deviation is the absolute measure of variation and it is calculated as the square root of the variance. It is expressed in the same units as the statistical data and therefore it is interpreted (Aczel, 2005, p. 26). The standard deviation of characteristic X is denoted by $S(x)$.

Based on the value of arithmetic mean $\bar{x}$ and the value of standard deviation $S(x)$, the classical coefficient of variation $V(x)$ may be calculated. It is defined as the quotient of standard deviation and arithmetic mean (Hoseini, Mohammadi, 2012, p. 1). Therefore it can be assumed as the relative measure of dispersion of statistical units in terms of analysed statistical characteristic (Podgórski, 2005, p. 68). The classical coefficient of variation is unitless, however, for interpretation purposes it is expressed as percentage (Kelley, 2007, p. 755). The higher coefficient $V(x)$ is, the more diverse statistical population is (Buga, Kassyk-Rokicka, 2008, p. 47). The coefficient of variation is particularly useful for comparing the level of dispersion of a few variables in the same population or for comparing the level of dispersion of one variable in various populations (Żyżyński, 2000, p. 68).

It is assumed that when the classical coefficient of variation is below 10%, the dispersion of the variable examined is statistically insignificant. On the other hand, in the population with high diversification, the classical coefficient of variance may be even higher than 100% (Kot and others, 2007, p. 179). The manner of determining the dispersion of examined statistical characteristic

depending on the value of classical coefficient of variation is shown in Table 1, but the thresholds determined there are only conventional.

(a)

Table 1. The manner of determining the level of dispersion based on the classical coefficient of variation

Range of coefficient $V(x)$	Interpretation (determining the level of variability)
0 − 10%	very low variability
10 − 20%	low variability
20 − 40%	moderate variability
40 − 60%	high variability
60% and more	very high variability

Source: own compilation based on: (Pułaska-Turyna, 2005, p. 78).

When the arithmetic mean and standard deviation are computed, then the typical data intervals may be determined. They include about 68% of all the observations in the statistical population (Makać, Urbanek-Krzysztofiak, 2001, p. 99). The typical data interval based on the classical measures is determined by the formula below (Liskowski, Tauber, 2003, p. 66):

$$\bar{x} - S(x) < x_{typ.} < \bar{x} + S(x).$$

Let the given population be divided into n separate groups. Then, the mean value of statistical characteristic X for each group may be computed. It is expressed as $\bar{x}_i$ (i = 1, 2, …, n) for the purpose of this article. Thus, the arithmetic mean of all the means in considered groups is expressed as $\overline{\bar{x}_i}$. Its value equals the total mean $\bar{x}$ computed for all the observations from n groups in total (i.e. $\overline{\bar{x}_i} = \bar{x}$).

For each i-th group, the within-group variance $S^2(x_i)$, within-group standard deviation $S(x_i)$ and classical within-group coefficient of variation $V(x_i)$ can be computed – they are the within-group measures of dispersion. If the means of considered groups are not the same, so if $\bar{x}_1 \neq \bar{x}_2 \neq ... \neq \bar{x}_n$, the variance computed for entire statistical population under study (i.e. $S^2(x)$) is higher than the mean within-group variance $\overline{S^2(x_i)}$, the total standard deviation $S(x)$ is higher than the mean within-group standard deviation $\overline{S(x_i)}$ and finally the total coefficient of variation $V(x)$ is higher than the mean within-group coefficient $\overline{V(x_i)}$.

Using between-group measures of variation we can determine the size of average differences between the observations of separate groups, i.e. the

differences between the means in the said groups (i.e. values $\bar{x}_1$, $\bar{x}_2$, …, $\bar{x}_n$). In order to determine the degree of this variability, between-group variance $S^2(\bar{x}_i)$, between-group standard deviation $S(\bar{x}_i)$ and between-group coefficient of variation $V(\bar{x}_i)$ have to be computed. Obviously, the values of $S^2(\bar{x}_i)$ (and also $S(\bar{x}_i)$ and $V(\bar{x}_i)$) are affected by not only the within-group means calculated, but also by the number of units in each group (Zeliaś, 2000, p. 62).

The variance has a property which is very important for the purpose of this article. Namely, the sum of the between-group variance and the mean within-group variance is always the same as the total variance computed for entire statistical population considered (Fabisiak, Kaźmierczak, 2012, p. 46). It may be expressed by the equation below (Western, Bloome, 2009, p. 4):

$$S^2(x) = S^2(\bar{x}_i) + \overline{S^2(x_i)},$$

where:

$S^2(x)$ – variance computed for the entire analysed population consisting of *n* groups;

$\bar{x}_i$ – arithmetic mean computed for *i*-th group (*i* = 1, 2, …, *n*);

$S^2(\bar{x}_i)$ – between-group variance;

$S^2(x_i)$ – within-group variance computed for *i*-th group;

$\overline{S^2(x_i)}$ – mean within-group variance.

The above equation enables drawing a conclusion that if each statistical unit from the *i*-th group was the same value concerning examined variable as the *i*-th group mean, then the within-group variances would equal zero, so the mean within-group variance would equal zero as well, and then the total variance would be the same as the between-group variance.

If total variance $S^2(x)$ is the sum of two components, so by dividing each component by $S^2(x)$ we may compute the shares of $S^2(\bar{x}_i)$ and $\overline{S^2(x_i)}$ in the sum. Therefore, the ratio $\frac{S^2(\bar{x}_i)}{S^2(x)}$ is the share of the between-group variance in the total variance and the ratio $\frac{\overline{S^2(x_i)}}{S^2(x)}$ is the share of the within-group variance in the total variance.

3. Mean disposable income *per capita* in Poland and within the groups under study

Based on information on the level of household disposable income and the size of household, the income *per capita* may be calculated. Such a value may be computed for each household surveyed by the Central Statistical Office of Poland in the household budget surveys. The database information provided by GUS made it possible to assign every household to relevant residence class, which, in turn, enabled computing the mean value of household disposable income *per capita* in each of the six groups. Then, the mean household disposable income *per capita* was computed for all households (i.e. in total regardless of the residence class). Such calculations were repeated fifteen times separately for each year from 1998 to 2012. The obtained results are shown in Table 2.

Table 2. Mean household disposable income *per capita* by class of residence in years 1998-2012 (in PLN)

Years	Town by size in thousands of inhabitants					Rural	Total
	500 and more	200–499	100–199	20–99	less than 20		
1998	744.85	597.49	562.24	535.39	482.76	408.58	512.53
1999	843.59	642.91	613.58	574.83	522.85	436.31	554.87
2000	927.83	733.10	695.54	624.01	567.28	477.71	603.10
2001	960.25	786.35	736.90	689.06	591.06	508.33	649.45
2002	1,001.24	819.06	754.18	708.84	625.69	522.96	673.70
2003	1,068.05	824.53	754.36	721.55	661.49	529.12	693.86
2004	1,115.93	847.02	750.91	764.74	677.18	544.09	717.37
2005	1,124.06	912.08	800.65	775.77	695.16	581.44	731.61
2006	1,258.11	1,019.90	858.50	841.37	766.54	653.14	798.90
2007	1,416.02	1,128.12	978.85	937.18	844.55	742.94	899.20
2008	1,609.84	1,238.35	1,157.61	1,068.21	1,001.76	841.17	1,022.95
2009	1,765.16	1,301.87	1,242.92	1,167.13	1,061.35	904.34	1,099.80
2010	1,912.92	1,417.51	1,294.78	1,243.32	1,130.86	972.44	1,180.55
2011	1,955.99	1,465.86	1,349.00	1,273.10	1,197.04	998.15	1,219.25
2012	2,036.65	1,525.18	1,355.82	1,311.55	1,233.59	1,065.17	1,276.92

Source: own computation based on the household budget surveys carried by the Central Statistical Office of Poland.

Based on the data in respective columns of Table 2, the following conclusion may be drawn: the mean household disposable income *per capita* is higher in cities/towns than in villages and the more residents are, the higher income is. The comparison of the within-group means obtained with the mean of the entire statistical population also allows to state that the mean household disposable income *per capita* in towns with at least 20,000 residents exceeds the total mean income *per capita*, while in the towns with less residents than 20,000 and villages the mean household disposable income *per capita* is lower than mean income *per capita* computed for all the groups in total.

4. Dispersion of disposable income *per capita* in Poland and within the groups under study

As it was already mentioned, the mean value does not provide comprehensive information on the distribution of studied variable within the population. Since the mean is a measure of central tendency, it informs only on the value around which the observations are focused. Therefore – for example – two populations may have the same value of the arithmetic mean, although there are significant differences between the observed values of the variable in the first population, while such differences are very slight or even do not exist at all in the second one. Hence, in order to better know the structure of phenomenon concerned, not only the average was analysed but also the variation of units with regard to the statistical characteristic considered.

The objective is to compare the dispersion within six groups into which the population was divided with the dispersion between the groups. In order to achieve the said objective, relevant measures of variability were computed, namely the variance and the standard deviation, as well as the classical coefficient of variation based on the standard deviation. Table 3 shows the values of variance computed for each group out of six residence classes as well as for the total number of surveyed households.

Table 3. Variance (in PLN^2)

Years	Town by size in thousands of inhabitants					Rural	Total
	500 and more	200–499	100–199	20–99	less than 20		
1998	289,882.3	124,008.3	105,593.3	137,196.1	79,962.0	110,273.8	145,271.6
1999	1,116,661.8	179,247.6	125,859.3	296,836.1	91,282.0	127,638.1	288,905.9
2000	472,804.2	234,433.8	202,372.1	155,563.3	160,070.0	446,940.9	332,198.6
2001	515,795.8	246,234.3	276,130.2	190,050.8	149,157.0	183,937.9	252,889.5
2002	638,100.8	278,382.4	235,645.4	320,345.1	156,969.4	509,405.5	421,817.3
2003	707,300.1	330,482.4	251,999.6	223,476.9	290,597.2	181,028.0	314,861.5
2004	867,142.6	320,211.5	237,952.6	260,458.0	193,048.2	315,556.0	380,900.0
2005	881,191.4	425,348.0	247,677.6	345,058.5	217,323.8	339,987.6	414,627.2
2006	1,198,118.1	534,721.1	288,783.3	287,785.5	246,490.5	311,525.8	439,307.7
2007	2,240,545.7	592,013.4	369,861.8	372,644.8	233,640.4	553,538.5	693,845.6
2008	2,210,778.4	602,373.7	656,123.0	506,675.9	334,292.5	1,381,298.5	1,159,241.9
2009	1,605,506.7	700,795.9	635,647.2	523,957.5	372,284.8	903,046.5	886,726.7
2010	3,287,239.3	5,199,600.5	621,372.9	572,342.3	666,130.4	906,183.7	1,460,063.2
2011	3,238,306.2	898,564.0	663,402.1	599,062.6	510,073.9	896,361.0	1,111,116.6
2012	3,505,695.7	953,054.9	640,851.7	639,113.3	525,278.8	1,212,367.8	1,307,952.1

Source: the same as in Table 2.

Next, the square root of each value of variance was taken to obtain the corresponding values of standard deviation. Table 4 shows computed 105 values of standard deviation.

Table 4. Standard deviation (in PLN)

Years	Town by size in thousands of inhabitants					Rural	Total
	500 and more	200–499	100–199	20–99	less than 20		
1998	538.41	352.15	324.95	370.40	282.78	332.08	381.15
1999	1,056.72	423.38	354.77	544.83	302.13	357.26	537.50
2000	687.61	484.18	449.86	394.42	400.09	668.54	576.37
2001	718.19	496.22	525.48	435.95	386.21	428.88	502.88
2002	798.81	527.62	485.43	565.99	396.19	713.73	649.47
2003	841.01	574.88	502.00	472.73	539.07	425.47	561.13
2004	931.20	565.87	487.80	510.35	439.37	561.74	617.17
2005	938.72	652.19	497.67	587.42	466.18	583.08	643.92
2006	1,094.59	731.25	537.39	536.46	496.48	558.14	662.80
2007	1,496.85	769.42	608.16	610.45	483.36	744.00	832.97
2008	1,486.87	776.13	810.01	711.81	578.18	1,175.29	1,076.68
2009	1,267.09	837.14	797.27	723.85	610.15	950.29	941.66
2010	1,813.07	2,280.26	788.27	756.53	816.17	951.94	1,208.33
2011	1,799.53	947.93	814.49	773.99	714.19	946.76	1,054.10
2012	1,872.35	976.25	800.53	799.45	724.76	1,101.08	1,143.66

Source: own computation based on Table 3.

Once standard deviation values were divided by relevant mean values, the coefficient values, which are relative measures of dispersion, were obtained. Since the numerator (the standard deviation) and the denominator (the mean) of the coefficient of variation are expressed in the same unit (PLN), then the obtained quotient will be a unitless measure, and in order to make the interpretation easier it was multiplied by 100%. The values of the coefficient of variation computed separately for each class of residence and for all statistical units examined are presented in Table 5.

Table 5. Coefficient of variation by type of residence

Years	Town by size in thousands of inhabitants					Rural	Total
	500 and more CV (%)	200–499 CV (%)	100–199 CV (%)	20–99 CV (%)	less than 20 CV (%)	CV (%)	CV (%)
1998	72.3	58.9	57.8	69.2	58.6	81.3	74.4
1999	125.3	65.9	57.8	94.8	57.8	81.9	96.9
2000	74.1	66.0	64.7	63.2	70.5	139.9	95.6
2001	74.8	63.1	71.3	63.3	65.3	84.4	77.4
2002	79.8	64.4	64.4	79.8	63.3	136.5	96.4
2003	78.7	69.7	66.5	65.5	81.5	80.4	80.9
2004	83.4	66.8	65.0	66.7	64.9	103.2	86.0
2005	83.5	71.5	62.2	75.7	67.1	100.3	88.0
2006	87.0	71.7	62.6	63.8	64.8	85.5	83.0
2007	105.7	68.2	62.1	65.1	57.2	100.1	92.6
2008	92.4	62.7	70.0	66.6	57.7	139.7	105.3
2009	71.8	64.3	64.1	62.0	57.5	105.1	85.6
2010	94.8	160.9	60.9	60.8	72.2	97.9	102.4
2011	92.0	64.7	60.4	60.8	59.7	94.9	86.5
2012	91.9	64.0	59.0	61.0	58.8	103.4	89.6

Source: own computation based on Table 2 & 4.

Analysis of data presented in Table 5 allows stating that households in Poland vary significantly as far as household disposable income *per capita* is concerned. The variation is significant not only in entire statistical population studied in this article but also in each of six groups of the population. Special attention should be paid to exceptionally high value of the coefficient computed for villages and the largest cities. Risking a guess, with such a high dispersion, the mean looses its informative value. In order to prove such a conclusion, let us take data for any year within fifteen-year-study, say, 2012. So, lower and upper limits of the typical data intervals in the case of said groups in given year were the following:

- cities with 500,000 residents and more: PLN 164.30 and PLN 3,909.00;
- cities with 200,000 to 499,999 residents: PLN 548.93 and PLN 2,501.43;
- towns with 100,000 to 199,999 residents: PLN 555.29 and PLN 2,156.35;
- towns with 20,000 to 99,999 residents: PLN 512.10 and PLN 2,111.00;
- towns with less than 20,000 residents: PLN 508.83 and PLN 1,958.35;
- and villages: PLN −35.91 and PLN 2,166.25.

Indeed, the households in cities with more than 500,000 residents have the mean household disposable income *per capita* higher by as much as PLN 971.48 than the households in villages. However, the dispersion within the said two groups is so high that, for example, typical households from the cities with more than 500,000 residents are the households with income *per capita* in the amount of PLN 165, while simultaneously typical rural households are the households with income in the amount of even PLN 1,950. It provokes reflection, since the average differences between the households within given classes are much bigger that the differences between the households from various classes. Further part of this article will prove that statement, so the comparison of between-group and within-group variability will be carried out.

5. Dispersion of disposable income *per capita* between groups and within-group dispersion

Table 3 shows the results of computed within-group variances. The mean within-group variance may be calculated based on the above results and size of each group. Then the between-group variance may be estimated based on the means in these groups and the sizes of them. Table 6 shows the information on between-group variances and mean within-group variances in years concerned.

Table 6. Comparison of between-group and within-group variation

Years	Variance (in PLN^2)		Standard deviation (in PLN)		Classical coefficient of variation (in %)		The share in the total variance (in %)	
	between group	mean within-group	between-group	mean within-group	between-group	mean within-group	of the between-group variance	of the mean within-group variance
1998	11,623.8	133,647.8	107.81	365.58	21.0	71.3	8.0	92.0
1999	16,275.2	272,630.7	127.57	522.14	23.0	94.1	5.6	94.4
2000	18,853.5	313,345.0	137.31	559.77	22.8	92.8	5.7	94.3
2001	21,862.8	231,026.7	147.86	480.65	22.8	74.0	8.6	91.4
2002	24,165.4	397,651.9	155.45	630.60	23.1	93.6	5.7	94.3
2003	29,059.7	285,801.9	170.47	534.60	24.6	77.0	9.2	90.8
2004	32,591.5	348,308.5	180.53	590.18	25.2	82.3	8.6	91.4
2005	29,754.2	384,873.0	172.49	620.38	23.6	84.8	7.2	92.8
2006	35,451.9	403,855.8	188.29	635.50	23.6	79.5	8.1	91.9
2007	43,754.0	650,091.6	209.17	806.28	23.3	89.7	6.3	93.7
2008	55,568.7	1,103,672.4	235.73	1,050.56	23.0	102.7	4.8	95.2
2009	66,971.9	819,754.8	258.79	905.40	23.5	82.3	7.6	92.4
2010	80,118.0	1,379,945.2	283.05	1,174.71	24.0	99.5	5.5	94.5
2011	84,680.4	1,026,436.2	291.00	1,013.13	23.9	83.1	7.6	92.4
2012	86,145.9	1,221,806.2	293.51	1,105.35	23.0	86.6	6.6	93.4

Source: own computation based on Table 2 & 3.

Comparing the value of between-group variance with the mean within-group variance in each year concerned makes it possible to state that the dispersion of entries within the classes of residence is significantly higher than the dispersion of entries between the classes. Obviously, the same conclusion may be drawn when comparing relevant values of standard deviation. What is interesting, the ratio of mean within-group standard deviation to between-group standard deviation in each year concerned was almost the same and from 1998 to 2012 the mean within-group standard deviation was about four times higher than the between-group standard deviation.

The mean value of variable in the entire analysed statistical population is the same as the mean of the means in groups into which the population was divided. Hence, the denominator of between-group coefficient of variation and the denominator of mean within-group coefficient of variation are the same and are equal to the denominator of total coefficient of variation (and the denominator is the mean value of examined characteristic), then the quotient of the mean within-group coefficient and the between-group coefficient equals the quotient of relevant standard deviations and it will be around 4. Therefore, the average differences in the household disposable income *per capita* between two households of the same residence class are four times bigger that the average differences between two households with disposable income *per capita* at the mean level of two various classes.

As it was already mentioned in the first chapter of this article, the total variance equals the sum of the between-group variance and the mean within-group variance. So, the total variance consists of two components and determining the structure of total variance makes it possible to know precisely the significance of each component. Hence, the share of the first component in total variance may be computed by dividing the between-group variance by the total variance. By analogy, the share of the second component in total variance may be computed by dividing the mean within-group variance by the total variance. The last two columns in Table 6 show the information on the impact of the between- and within-group variances on the total variance in 15 successive years. So, in each year taken into consideration the mean within-group variance was over 90% of the total variance and the share of the between-group variance was always below 10%. Undoubtedly, the average difference between disposable income *per capita* between households of the same group of residence is very big comparing to the differences between the means for households from various classes.

6. Comparison of between-group variability of household disposable income in EU countries

For statistical purposes a common classification into three disjoint and exhaustive groups of areas was prepared to be used by all of the European Union countries. This classification indicates the character of an area due to the degree of its urbanization. It is based on the share of local population living in urban clusters and in urban centres. The three types of areas are as follows (Eurostat website [1], date of access: 19.01.2015):

- sparsely populated areas (alternate name: rural areas);
- intermediate density areas (alternate name: towns and suburbs or small urban areas);
- densely populated areas (alternate name: cities or large urban areas).

The rules of classifying local administrative units of countries into these three groups were specified precisely. The methodology is based on a combination of criteria of geographical contiguity and minimum population threshold applied to 1 km^2 population grid cells. This approach, based on mapping the territory by a grid square cell of 1 km^2, avoids distortions caused by using local administrative units varying in size and/or shape.

With information on average annual income *per capita* in each class and on the number of people falling within these classes, it is possible to calculate absolute and relative dispersion of income between the classes. Results of calculations on between-group standard deviation, which is the absolute measure of between-group variability, have been presented in Table 7. The said table also featured results obtained for between-group coefficient of variation, which is a relative measure of between-group variability. Calculations were carried out

separately for four consecutive years from 2010 to 2013 inclusive, and for all twenty-eight countries of the European Union.

Table 7. Comparison of between-group measures of income variation in all EU countries

Countries	Between-group standard deviation (in EUR)				Between-group coefficient of variation (in %)			
	2010	2011	2012	2013	2010	2011	2012	2013
Austria	967.96	811.89	480.73	449.09	4.1	3.4	2.0	1.8
Belgium	417.89	286.77	404.27	288.11	2.0	1.3	1.8	1.2
Bulgaria	758.80	777.77	645.67	845.21	21.8	22.7	19.6	23.8
Croatia	474.69	530.81	746.35	831.53	6.4	7.8	12.3	14.2
Cyprus	2,090.70	1,822.52	2,233.53	2,290.97	11.0	9.3	11.0	11.7
Czech Republic	635.68	611.20	680.83	570.33	8.0	7.3	7.8	6.6
Denmark	682.30	1,879.82	614.79	1,029.54	2.5	6.6	2.1	3.5
Estonia	714.57	559.44	693.79	767.27	10.6	8.5	9.7	9.8
Finland	2,015.82	1,912.60	1,848.94	1,418.26	8.6	7.9	7.2	5.5
France	780.34	1,359.16	1,404.84	1,533.12	3.3	5.7	5.7	6.2
Germany	1,097.35	926.79	872.85	770.79	5.1	4.3	4.0	3.4
Greece	1,933.76	1,065.04	1,020.62	1,043.57	13.9	8.5	9.4	10.8
Hungary	627.89	798.16	757.65	791.97	13.5	15.5	14.2	15.4
Ireland	(-)	(-)	2,618.31	(-)	(-)	(-)	11.9	(-)
Italy	1,114.97	1,146.21	1,421.84	1,535.18	6.1	6.3	8.0	8.8
Latvia	587.75	607.74	636.08	639.84	10.8	12.0	11.4	10.9
Lithuania	(-)	(-)	886.99	796.13	(-)	(-)	17.3	14.1
Luxembourg	2,607.49	2,197.65	3,701.28	3,883.74	7.1	5.9	9.9	10.0
Malta	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
Netherlands	468.82	636.90	160.49	420.90	2.1	2.8	0.7	1.8
P o l a n d	**893.59**	**946.62**	**990.11**	**960.81**	**17.5**	**16.3**	**16.7**	**16.0**
Portugal	1,283.11	1,347.27	1,501.21	1,181.01	12.2	13.0	14.6	11.9
Romania	593.58	588.53	655.56	465.08	24.9	24.3	27.1	19.5
Slovakia	658.22	621.67	721.09	504.03	9.9	9.0	9.6	7.0
Slovenia	763.91	745.44	577.46	565.06	6.0	5.8	4.5	4.5
Spain	1,496.50	1,493.21	1,492.21	1,901.01	10.4	10.7	10.8	12.2
Sweden	1,568.46	1,372.77	1,455.33	1,195.35	7.5	5.6	5.3	4.3
United Kingdom	1,008.91	241.86	837.53	672.15	4.9	1.2	3.7	3.1
EU (28)	**2,547.71**	**2,526.74**	**1,436.92**	**1,730.17**	**15.1**	**14.8**	**8.2**	**9.8**

(-) no reliable data disposable

Source: own computation based on Eurostat database: "Mean and median income by degree of urbanisation": http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di17&lang=en [*date of access: 19.01.2015*]*; "Annual population by sex, age, degree of urbanisation and labour status": http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfsa_pgauws&lang=en* [*date of access: 19.01.2015*].

When considering all countries of the European Union jointly, it is possible to see a significant decrease in the dispersion of between-group income *per capita* – coefficient of variation of more than 15% in 2010 dropped in 2013 below 10%. It is also worth noting that during the period under study, most of the twenty-eight

countries reported the value of the coefficient lower than 10%, i.e. this measure was at a level that may indicate very low between-group variability (see Table 1).

Out of all countries of the European Union the smallest income variability *per capita* between regions with different degrees of urbanization could be observed in the case of Belgium. The classical coefficient of between-group variation was 1 to 2% in the case of this country for each of the four years under analysis. The coefficient turned out to be at a similar level also in Austria and the Netherlands.

The data summarized in Table 7 also allows drawing a conclusion that in Poland the fact of living in a given class of residence in a much greater extent affects the size of income achieved than in the case of other countries of the European Union. However, it should be borne in mind that in Poland – like in the entire European Union – dispersion of *per capita* income between regions differing in the degree of urbanization dropped dramatically over the period of 2010–2013. The value of the classical coefficient of variation of 16.0% in 2013 enables an observation that, although mean relative differences between average income of persons from sparsely populated areas, intermediate-density areas and densely populated areas were much higher in Poland than in most other EU countries, eventually, the variability in Poland should be assessed as low.

Between-group dispersion of income higher than in Poland was only recorded in Romania and Bulgaria. Interestingly, in these countries the between-group standard deviation remained at a very low level, which in each year under analysis was lower than the between-group standard deviation observed in Poland. In Poland, however, the average income *per capita* is approximately 70% higher than in Bulgaria and more than 150% higher than in Romania; therefore, in relation to the average level of income in a given country, variability of the investigated variable in Poland was lower than in other two mentioned countries.

7. Discussion on the need to mitigate social inequalities

The following dimensions of inequality can be determined on the basis of social sciences (Wójcik-Żołądek, 2013, p. 2):

- the economic dimension (including categories such as income, property, living conditions);
- the social dimension (concerning education, lifestyle, participation in culture, social prestige);
- the political dimension (referring to differences in participation in power and in civic engagement).

Treating the economic aspect as the only dimension of inequality in society is therefore too much of an oversimplification. Income stratification is, however, construed in the literature on the subject as one of the most important measures of inequality, because the level of income is widely recognized as the most important determinant of social status. It is also stressed that income is a factor which influences the activity of individuals and households in almost all spheres

of life – from the development of material conditions through access to health care, provision of appropriate education, participation in culture, access to technological achievements, up to access to power. Therefore, encountering income limitations does not only narrow down the decision-making field of a household in terms of the size of realized consumption, but also determines the degree of failure to satisfy many other needs, including non-economic needs (Leszczyńska, 2014, p. 410). We may even be tempted to state that the size of income, having an impact on the achievement of a wide range of material and non-material objectives, is a major determinant of a sense of satisfaction with the overall quality of human life (Bal, 2012, p. 252).

Representatives of various trends in economics present different, often radically extreme, approaches to the problem of occurrence of income inequalities in society. The differences in approach are based usually in personal beliefs on philosophical, ethical, sociological and psychological foundations of economics (Umiński, 2013, p. 210). The discussion on consequences of social inequality – especially the stratification of income – takes place not only on the ground of social sciences, but also in the public debate, often causing a lot of emotion. Nevertheless, there is a general consensus among researchers that excessive income inequality infringes the principle of social justice and has a negative impact on economic growth (Pliszka, 2004, p. 354). Often in scientific and political debates, it is also stressed that exceeding a certain threshold of income stratification threatens the maintenance of social cohesion (Kołodko, 2014, p. 32). Thus, determining which income inequalities must be considered excessive and which optimal becomes a key issue. The aim of social policy should be to eliminate only the unjustified, and not all, social inequalities. It seems that helpful in this regard will be addressing the issue of causes of the occurring inequalities. Now, the source of income stratification of society is the differences in environmental and genetic conditions and differences in preferences and ambitions. Reducing inequalities resulting from the first group of conditions is undeniable – it does not arouse much controversy and involves wide social acceptance. In turn, reduction of income disparities related to differing decisions of individuals is at least debatable.

Thus, the basis for answering the question of which social inequalities are justified and which are not should be a distinction between two categories – possibilities and preferences. Justified inequalities are those for which the responsibility is borne exclusively by individuals through their autonomous decisions – whether educational, professional or those related to the degree of commitment to the improvement of their living conditions. On the other hand, unjust social inequalities are those independent of the will of a given individual, ones he or she cannot influence, does not control and is not able to change. There is no doubt that factors such as place of birth, environment of growing up, socio-economic situation of parents, immediate environment, abilities and aptitudes largely influence the size of income that this individual will achieve during his or her adult life, and cause the principle of equal opportunities to be undermined.

Thus, in order for disparities in income to be fully justified, the playing field should be levelled. On the other hand, the way the players will behave on the field depends entirely on them and they alone bear responsibility for their actions (Bartak, 2014, p. 224). We can perceive as justified only a situation where personal effort determines the success in life rather than inherited wealth or favourable family environment in childhood, which equips the child with appropriate cultural capital right from the start and allows him or her to access better education (Woźniak, 2012, p. 27–28).

The subject of analysis in this paper are income inequalities due to different conditions of life in big cities, in small towns and villages. These inequalities should undoubtedly be mitigated through the application of appropriately selected tools. A well-designed social policy should therefore limit inequalities arising from the fact that people do not start at the same position in the race for a better financial situation, a higher social status and the associated convenience. The best way to reduce income inequalities is to provide all social groups with access to modern education adapted to the requirements of a knowledge-based economy. It is also necessary to allow individual entities access to adequate infrastructure, to the use of achievements of technical and technological progress and to the entire spectrum of achievements of civilization. The priority of state policy should always be to give equal opportunities, eliminate barriers, stimulate innovation and ensure fair competition. In the modern economy, government policy cannot be reduced, therefore, to redistributive activities, as it is obvious that it would only strengthen demanding attitudes, reinforce learned helplessness, restrict professional activity and self-responsibility of people (Bartak, 2014, p. 220). Proper state policy as regards reducing income inequalities does not slow down the pace of modernization processes that are being carried out in the economy; on the contrary – it leads to their acceleration. Disparities between large urban agglomerations, small towns and rural areas should therefore be mitigated by supporting well-designed investment in human capital and improvement of infrastructure.

8. Conclusions

The aim of this article was to assess the dispersion of disposable income *per capita* between households in Poland from various classes of residence in comparison to the dispersion of income within these classes. The said objective was achieved by execution of three research tasks.

In the article, two research hypotheses were verified. The first hypothesis stated that the highest household disposable income *per capita* in Poland is recorded in the cities with above 500,000 residents and the amount of the said income decreases with decreasing number of residents as well as the rural households have the lowest mean disposable income *per capita*. The hypothesis was verified positively on the basis of data from 1998 to 2012. The comparison of

the within-group means allowed drawing a conclusion that the said regularity is permanent as it occurred throughout fifteen years.

The second tested hypothesis stated that in terms of disposable income *per capita* the households in Poland vary to a larger extent within all the classes of residence than between the classes. The above hypothesis was verified positively as well. The mean within-group standard deviation was a few times higher than the between-group standard deviation and the share of between-group variance was only a few per cent of the total variance. Hence, without any doubt, the amount of household disposable income is affected by many other factors which are more important than the class of residence.

In conclusion, it should be also emphasized that the location of household (city, small town or village) is clearly significant for the level of household disposable income *per capita*, which has been proven by the occurring differences in the means computed for each group determined in the study. However, the differences between the said means should be considered slight, as compared to the average differences of the observed values between households of the same classes of residence. Therefore, the division for classes of residence proposed by the Central Statistical Office of Poland seems to be not a good one to show the variation of income *per capita* among Polish households because assigned class of residence explains at minimum extent the differences in the income levels. Therefore, a more appropriate way of division should be considered, namely the one better explaining the dispersion of household disposable income *per capita*. The authors of this article have already carried out such a study, and the results will be presented in further articles.

REFERENCES

ACZEL, A. D., (2005). Statystyka w zarządzaniu, [Statistics in management], PWN, Warsaw.

BAL, I., (2012). Marginalizacja i wykluczenie społeczne jako bariera rozwoju regionalnego, [The marginalization and the social exclusion of people as a reason for stopping regional development], Nierówności społeczne a wzrost gospodarczy 28, [Social Inequalities and Economic Rise], Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów.

BARTAK, J., (2014). Uwarunkowania redukcji nierówności dochodowych w Polsce, [Determinants of the reduction of income inequalities in Poland], Nierówności społeczne a wzrost gospodarczy 37, [Social Inequalities and Economic Rise], Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów.

BIELECKA, A., (2001). Statystyka w zarządzaniu. Opis statystyczny, [Statistics in management. Statistical description], Wyższa Szkoła Przedsiębiorczości i Zarządzania im. Leona Koźmińskiego w Warszawie, Warsaw.

Budżety gospodarstw domowych w 2012 r., [Household budget survey in 2012], (2013). Central Statistical Office of Poland, Warsaw.

BUGA, J., KASSYK-ROKICKA, H., (2008). Podstawy statystyki opisowej, [Basics of descriptive statistics], Wyższa Szkoła Finansów i Zarządzania w Warszawie, Warsaw.

Eurostat database:
http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di17&lang=en.

Eurostat database:
http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lfsa_pgauws&lang=en.

Eurostat website [1]:
http://ec.europa.eu/eurostat/web/degree-of-urbanisation/overview.

FABISIAK, A., KAŹMIERCZAK, A., (2012). Ocena poziomu wyżywienia gospodarstw domowych pracowników i rolników w Polsce za pomocą syntetycznego wskaźnika poziomu wyżywienia, [Evaluation of consumption level of employee and farmer households in Poland using synthetic index of consumption level], Journal of Agribusiness and Rural Development 2 (24).

HOSEINI, J., MOHAMMADI, A., (2012). Estimator and tests for coefficient of variation in uniform distribution, Biometrics & Biostatistics, Volume 3, Issue 5: 149.

KELLEY, K., (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach, Behavior Research Methods, The Psychonomic Society, 39 (4).

KOŁODKO, G., (2014). Społeczne i przestrzenne aspekty zróżnicowania dochodów we współczesnym świecie, [Social and spatial aspects of income inequality in the contemporary world], Nierówności społeczne a wzrost gospodarczy 39, [Social Inequalities and Economic Rise], Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów.

KOT, S., JAKUBOWSKI, J., SOKOŁOWSKI, A., (2007). Statystyka. Podręcznik dla studiów ekonomicznych, [Statistics. Handbook for economic studies], DIFIN, Warsaw.

LESZCZYŃSKA, M., (2014). Ocena społecznego zrównoważenia rozwoju w Polsce według kryterium dynamiki dochodów gospodarstw domowych, [Estimation of balanced social evolution from a perspective of incomes development changes in socio-professional groups of households], Nierówności społeczne a wzrost gospodarczy 37, [Social Inequalities and Economic Rise], Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów.

LISKOWSKI, M., TAUBER, R. D., (2003). Podstawy statystyki praktycznej, [Basics of practical statistics], Wyższa Szkoła Hotelarstwa i Gastronomii w Poznaniu, Poznań.

MAKAĆ, W., URBANEK-KRZYSZTOFIAK, D., (2001). Metody opisu statystycznego, [Methods of statistical description], Wydawnictwo Uniwersytetu Gdańskiego, Gdańsk.

PLISZKA, T., (2004). Skutki nierówności społecznych, [The results of social inequalities], Nierówności społeczne a wzrost gospodarczy 5, [Social Inequalities and Economic Rise], Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów.

PODGÓRSKI, J., (2005). Statystyka dla studiów licencjackich, [Statistics for bachelor degree studies], PWE, Warsaw.

PUŁASKA-TURYNA, B., (2005). Statystyka dla ekonomistów, [Statistics for economists], DIFIN, Warsaw.

UMIŃSKI, P., (2013), Nierówności dochodowe w koncepcji Johna K. Galbraitha – wskazanie źródeł i sformułowanie hipotez badawczych, [Income inequality in the concept of John K. Galbraith – an indication of the sources and research hypothesis formulation], Nierówności społeczne a wzrost gospodarczy 30, [Social Inequalities and Economic Rise], Wydawnictwo Uniwersytetu Rzeszowskiego, Rzeszów.

WESTERN, B., BLOOME, D., (2009). Variance function regression for studying inequality, Digital access to scholarship at Harvard, Harvard University, January 2009.

WOŹNIAK, W., (2012). Nierówności społeczne w polskim dyskursie politycznym, [Social inequalities in Polish political discourse], Wydawnictwo Naukowe Scholar, Warsaw.

WÓJCIK-ŻOŁĄDEK, M., (2013). Nierówności społeczne w Polsce, [Social inequalities in Poland], INFOS Zagadnienia społeczno-gospodarcze 20 (157), [INFOS Socio-economic issues], Biuro Analiz Sejmowych, [Parliamentary Research Office].

ZELIAŚ, A., (2000). Metody statystyczne, [Statistical methods], PWE, Warsaw.

ŻYŻYŃSKI, J., (2000). Podstawy statystyki, [Basics of statistics], Wyższa Szkoła Ekonomiczno-Humanistyczna w Skierniewicach, Skierniewice.

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 591–610

EVALUATION OF SELECTED APPROACHES TO CLUSTERING CATEGORICAL VARIABLES

Zdeněk Šulc[1], **Hana Řezanková**[2]

ABSTRACT

This paper focuses on recently proposed similarity measures and their performance in categorical variable clustering. It compares clustering results using three recently developed similarity measures (IOF, OF and Lin measures) with results obtained using two association measures for nominal variables (Cramér's V and the uncertainty coefficient) and with the simple matching coefficient (the overlap measure). To eliminate the influence of a particular linkage method on the structure of final clusters, three linkage methods are examined (complete, single, average). The created groups (clusters) of variables can be considered as the basis for dimensionality reduction, e.g. by choosing one of the variables from a given group as a representative for the whole group. The quality of resulting clusters is evaluated by the within-cluster variability, expressed by the WCM coefficient, and by dendrogram analysis. The examined similarity measures are compared and evaluated using two real data sets from a social survey.

Key words: variable clustering, nominal variables, association measures, similarity measures.

1. Introduction

When dealing with high dimensional data, reduction of the number of variables is often desired. It can spare both the computational time and costs for gathering the information in the future. The use of principal component analysis or factor analysis, as described, for example, by Jolliffe (2002), or their categorical counterparts, such as correspondence analysis Greenacre (2010), is very popular. These methods provide additional information about a data set, variables of which have significant loadings on a shared vector, see Palla et al. (2012). An approach based on multiple correspondence analysis for large data sets

[1] Department of Statistics and Probability, University of Economics, Prague. W. Churchill sq.4, 130 67 Praha 3, Czech Republic. E-mail: zdenek.sulc@vse.cz.
[2] Department of Statistics and Probability, University of Economics, Prague. W. Churchill sq.4, 130 67 Praha 3, Czech Republic. E-mail: hana.rezankova@vse.cz.

is presented by D'Enza and Greenacre (2012). Another way to achieve the dimensionality reduction of a data set can be to create groups of similar variables using cluster analysis. One variable of each group can be chosen as a representative for further analysis. Hierarchical cluster analysis represents the basic approach used for variable clustering, see Gordon (1999), Gan et al. (2007). It is based on a proximity matrix, which contains dissimilarities of analyzed variables taken pairwise. More sophisticated approaches are represented, for example, by model-based clustering, see Chavent et al. (2010); Everitt et al. (2011). In R software, one might find a few variable clustering procedures in a package named ClustOfVar, see Chavent et al. (2012). The practical use of variable clustering can be found in various fields of use, e.g. in questionnaires surveys, actuarial sciences, chemistry, gene expression analysis, see Palla et al. (2012), or in getting rid of redundant variables in predictive models, see Payne and Edwards (1999).

The paper focuses on comparison of two kinds of similarity measures which can be used in variable clustering with binary or nominal variables. The first ones are the association measures, Cramér's V and the uncertainty coefficient, which express the dependency between two variables based on the chi-square statistic and the ANOVA method. The second kind is represented by recently developed similarity measures, IOF, OF and Lin, which were originally proposed for object clustering, but have been adjusted for variable clustering in this paper. Clustering with both kinds of measures is going to be compared with the simple matching coefficient, which is commonly used in categorical data clustering and thus it can serve as a reference measure.

The IOF, OF and Lin measures have never been evaluated for variable clustering; they have only been studied for object clustering so far. Moreover, the evaluations of these measures were performed only with the known cluster membership, see Boriah et al. (2008), Chandola et. al. (2009); thus cluster analysis was treated more like a classification problem with supervised learning. Moreover, both publications were focused on the outlier detection performance of the similarity measures.

In this paper, two data sets from a social survey are analyzed. The quality of clusters, obtained using different similarity measures, is evaluated from aspects of both the within-cluster variability, measured by the WCM (within-cluster mutability) coefficient, and the dendrogram analysis. To minimize the influence of clustering algorithm on clustering performance of the similarity measures, clusters obtained by three linkage methods are compared and evaluated.

The rest of the paper is organized as follows. Section 2 introduces the association and other similarity measures. Section 3 describes evaluation criteria of cluster quality. The application of theoretical approach to real data is presented in Section 4. The final results are summarized in the Conclusion.

2. Nominal variable clustering

A basic approach to variable clustering is to create a dissimilarity matrix, which contains dissimilarities of analyzed variables taken pairwise, and then to apply agglomerative hierarchical cluster analysis. A dissimilarity measure can be derived from a similarity measure. Many similarity measures have been proposed for categorical data. One can use association measures for nominal variables, see Anderberg (1973), or similarity measures determined for objects characterized by nominal variables. There are also several other approaches, for example, in Chavent et al. (2010), where the adjustment of existing centre-based method for categorical variable clustering is presented. It is not possible to compare all approaches or all measures; therefore, we focus only on the selected ones.

Three linkage methods of hierarchical clustering are applied in this paper: *complete method* (CLM), *single method* (SLM) and *average method* (ALM). In CLM, the dissimilarity between the furthest variables from two different clusters is considered as the distance between these clusters. SLM takes the dissimilarity between the nearest variables from two different clusters for this purpose, and ALM takes the average distance of all dissimilarities between variables from two different clusters.

2.1. Association measures

Different types of association measures for nominal variables are used in multivariate analysis. Some of them are based on Pearson's chi-squared statistic, some on the principle of dependence measurement in the ANOVA method.

The measures based on the chi-square statistic compare observed and expected counts under the hypothesis of independence; these counts are frequencies of combinations of categories of two nominal variables. Pearson's coefficient of contingency, Cramér's V and the phi coefficient belong to this group. In this paper, *Cramér's V* is applied because it takes values from the interval [0, 1] and takes into account the numbers of categories. It is calculated according to the formula

$$V = \sqrt{\chi^2 / n(q-1)}\,, \tag{1}$$

where χ^2 is Pearson's chi-squared statistic, n is the number of surveyed objects and q is a minimum number of categories of two analyzed variables. If at least one variable is dichotomous, then values of Cramér's V equal the values of the phi coefficient. Cramér's V can be transformed into a dissimilarity measure by subtracting its value from 1.

In the ANOVA method, a directional dependence is considered. In such a case, a symmetric measure is calculated as the harmonic mean of two asymmetric measures. There are two symmetric coefficients for nominal variables derived from asymmetric measures which are based on the principle of ANOVA: the

lambda coefficient and the uncertainty coefficient. The former one is based only on frequencies of modal categories, the latter one takes into account frequencies of all combinations of categories. Therefore, *the uncertainty coefficient* is applied in our experiments. It takes values from the interval [0, 1] and it is based on the entropy as a variability measure. For the *c*-th and *d*-th variables it is calculated as

$$U_{cd} = \frac{2 \cdot (H_c + H_d - H_{cd})}{H_c + H_d}, \tag{2}$$

where H_c (H_d) is the entropy of the *c*-th (*d*-th) variable and H_{cd} is the within-group entropy. Generally, the entropy H is expressed as

$$H = -\sum_{u=1}^{h} p_u \ln p_u , \tag{3}$$

where p_u is a relative frequency of the *u*-th category and h is the number of categories if for all u $p_u \neq 0$. In the case of $p_u = 0$, the corresponding addend equals 0 for this u. The uncertainty coefficient can be transformed into a dissimilarity measure by subtracting its value from 1.

More association measures for variable clustering can be found in Řezanková (2014).

2.2. Recently developed similarity measures

Compared with association measures, which are based on frequencies in a contingency table, the other similarity measures considered in this paper compare categories taken pairwise for each object individually. The term *the other similarity measures* covers the recently developed similarity measures (IOF, OF and Lin) and the overlap measure, which serves as a reference measure. All these measures have a drawback which is that all analyzed variables must have the same number of categories and the categories must have the same meaning. The reason is as follows: if categories across the variables did not have the same meaning, it would make no sense to compare them. For this reason the same number of categories is considered.

All formulas in this paper are based on the data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, ..., n$ and $c = 1, 2, ..., m$ (n is the total number of objects, m is the total number of variables).

Originally, the *IOF* (inverse occurrence frequency) measure comes from an information retrieval, where it used to serve to determine a relative number of documents containing a specific word, see Sparck-Jones (1972, 2002). The original measure was designed to deal only with binary variables; later, it was adjusted to deal with nominal variables as well. The measure was constructed to assign higher weights to mismatches on less frequent values and lower weights to mismatches on more frequent values. When determining similarity between variables $\mathbf{x}_c$ and $\mathbf{x}_d$ for the *i*-th object, it can be expressed as

$$S_i(x_{ic}, x_{id}) = \begin{cases} 1 \text{ if } x_{ic} = x_{id} \\ \dfrac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{id})} \text{ otherwise} \end{cases}, \tag{4}$$

where $f(x_{ic})$ is a frequency of the category x_{ic} of the i-th object. Dissimilarity between variables $\mathbf{x}_c$ and $\mathbf{x}_d$ is expressed as

$$D(\mathbf{x}_c, \mathbf{x}_d) = \frac{1}{\dfrac{\sum_{i=1}^{n} S_i(x_{ic}, x_{id})}{n}} - 1. \tag{5}$$

The *OF* (occurrence frequency) measure has an opposite system of weights to the IOF measure. It assigns higher weights to mismatches on more frequent values and otherwise, i.e.

$$S_i(x_{ic}, x_{id}) = \begin{cases} 1 \text{ if } x_{ic} = x_{id} \\ \dfrac{1}{1 + \ln \dfrac{m}{f(x_{ic})} \cdot \ln \dfrac{m}{f(x_{id})}} \text{ otherwise} \end{cases}. \tag{6}$$

Dissimilarity can be determined using Equation (5).

The *Lin* measure, which was introduced by Lin (1998), represents an information-theoretic definition of similarity based on relative frequencies. It was derived from theoretic assumptions about similarity. The emphasis was put on the universality of use; thus, it can be used in various situations including determination of similarity between ordinal values. It assigns higher weights to more frequent categories in the case of a match and lower weights to less frequent categories in the case of a mismatch, i.e.

$$S_i(x_{ic}, x_{id}) = \begin{cases} 2 \cdot \ln p(x_{ic}) \text{ if } x_{ic} = x_{id} \\ 2 \cdot \ln(p(x_{ic}) + p(x_{id})) \text{ otherwise} \end{cases}, \tag{7}$$

where $p(x_{ic})$ expresses a relative frequency of the category x_{ic} of the i-th object. The dissimilarity measure is defined as

$$D(\mathbf{x}_c, \mathbf{x}_d) = \frac{1}{\dfrac{\sum_{i=1}^{n} S_i(x_{ic}, x_{id})}{\sum_{i=1}^{n} (\ln p(x_{ic}) + \ln p(x_{id}))}} - 1 \cdot \tag{8}$$

Clustering with the measures mentioned above is compared with results obtained using the *overlap* measure, which takes into account only whether two observations match or not. When determining similarity between variables $\mathbf{x}_c$ and $\mathbf{x}_d$ for the i-th object, it assigns value 1 if the variables match and value 0 otherwise.

$$S_i(x_{ic}, x_{id}) = \begin{cases} 1 \text{ if } x_{ic} = x_{id} \\ 0 \text{ otherwise} \end{cases}. \tag{9}$$

The dissimilarity measure is defined as

$$D(\mathbf{x}_c, \mathbf{x}_d) = 1 - \frac{\sum_{i=1}^{n} S_i(x_{ic}, x_{id})}{n}. \tag{10}$$

Unlike recently developed similarity measures, the overlap measure does not take into account frequency distribution of categories of a given object, which could serve as an important factor for determining similarity between variables. The comparison of the above mentioned coefficients applied for an object clustering with respect to the within-cluster variability is described in Šulc and Řezanková (2014).

3. Evaluation criteria of final clusters

In this paper, the quality of final clusters is evaluated from the aspects of the WCM (within-cluster mutability) coefficient and by the dendrogram analysis.

The within-cluster variability is an important indicator of cluster quality. With an increasing number of clusters, the within-cluster variability decreases, so the clusters become more homogenous. In this paper, the measurement of the within-cluster variability is based on the *Gini coefficient*, which determines the variability (mutability) of nominal variables. It is expressed by the following equation

$$G_{gi} = 1 - \sum_{u=1}^{h} \left(\frac{n_{giu}}{m_g} \right)^2, \tag{11}$$

where m_g is the number of variables in the g-th cluster (g = 1, ..., k), n_{giu} is the number of variables in the g-th cluster by the i-th object with the u-th category (u = 1, 2, ..., h; h is the number of categories). After standardization of this coefficient with the aim to get values from 0 to 1, and its extending for n objects and k clusters, it can be expressed in a form of the *normalized within-cluster mutability coefficient*:

$$WCM(k) = \frac{1}{n} \frac{h}{h-1} \sum_{g=1}^{k} \frac{m_g}{m} \sum_{i=1}^{n} \left(1 - \sum_{u=1}^{h} \left(\frac{n_{giu}}{m_g} \right)^2 \right), \tag{12}$$

where m is the number of variables. The *WCM* coefficient is based on the G' measure, which was proposed by Řezanková et al. (2011) for the purpose of evaluation of object clustering.

When clustering a relatively small number of variables, the dendrogram analysis can be very helpful. Dendrograms visualize the process of agglomerative

hierarchical clustering calculation. They have a form of charts, which have the examined variables, e.g. on the Y axis, and the distance between clusters on the X axis. They can be cut at any point to get a particular cluster solution.

4. Real data application

To illustrate the influence of selected association and other similarity measures on variable clustering, two variable sets, which come from the research *Men and Women with a University Degree*, are chosen. This survey was conducted by the *Institute of Sociology of the Academy of Sciences of the Czech Republic*, see the archives of the institute (http://archiv.soc.cas.cz).

The following software was used for the analysis: Matlab, IBM SPSS Statistics, STATISTICA and MS Excel. In Matlab, proximity matrices for all similarity measures were computed. In IBM SPSS, hierarchical cluster analyses using CLM, SLM and ALM were performed. In STATISTICA, dendrograms were created. In MS Excel, evaluation criteria for cluster quality evaluation were computed.

4.1. Description of the variable sets

Two batteries of questions were chosen for the analysis. The first battery consists of 9 variables; all with two possible answers *yes* or *no*. The questions are: *From family reasons, have you ever:* p27a – *worked part-time*, p27b – *worked in shifts*, p27c – *worked flextime*, p27d – *changed a job*, p27e – *changed a profession*, p27f – *moved*, p27g – *refused a job offer*, p27h – *refused a promotion offer*, p27i – *cheated at work?* The cases with missing values were omitted, so answers from 1,904 respondents were included.

The second battery deals with gender equality. It contains 9 variables, which all have three possible answers: *women have better opportunities than men*, *men and women have approximately equal opportunities* and *men have better opportunities than women*. The variables are the following: p13a – *to get a job*, p13b – *to have better salary for the same job*, p13c – *to get a leadership*, p13d – *to be a director*, p13e – *to be promoted*, p13f – *for a salary increase*, p13g – *to gain benefits*, p13h – *to have authority*, p13i – *to keep a job*. There is one additional variable with the name: p12 – *a chance of success* which has the same categories as the previous battery of questions. For this reason, it can be added to the set of variables. Overall, answers from 1,886 respondents were used.

4.2. Binary variable clustering

Table 1 presents values of the WCM coefficient for the solutions with two to five clusters for CLM, computed for the set of questions with binary answers. The quality of a particular cluster solution can be evaluated according to the within-cluster variability expressed by the WCM coefficient. The lower the value of WCM, the better the cluster solution. For the two-cluster solution, most of the

measures, except for the Lin measure, provide the same results, i.e. 0.366. For cluster solutions for three and more clusters, the best results are provided by the recently developed similarity measures, i.e. IOF, Lin and OF, which have the same results. They are followed by the overlap measure and further by the both association measures.

Table 1. Values of the WCM coefficient for clustering of binary variables (CLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.366	0.320	0.255	0.186
Coefficient U	0.366	0.320	0.254	0.186
IOF measure	0.366	0.297	0.232	0.168
OF measure	0.366	0.297	0.232	0.168
Lin measure	0.375	0.297	0.232	0.168
Overlap measure	0.366	0.301	0.236	0.172

Another approach to evaluate the clustering performance is to use dendrograms, which are presented in Figure 1. When looking at the dendrograms, it is apparent that they can be separated into three groups from the point of view of the clustering structure. The first one comprises both the association measures, the second one includes the recently developed similarity measures and the last one contains only the overlap measure. Similarity measures in a particular group provide similar results. Since data dimension reduction is the primary goal of variable clustering, low-cluster solutions are preferred.

When using SLM, as shown in Table 2, one might see that the results are very different from the results achieved by CLM. Generally, they are all worse. There are apparent interesting changes in behaviour of the similarity measures. Both association measures perform better than the recently developed similarity measures from the point of view of their within-cluster variability and the interpretation of dendrograms. Moreover, using SLM, the advantage of recently developed similarity measures, which is based on taking into account frequency distribution of categories, is not apparent in the results. Thus, their results are very similar to the overlap measure, which is also demonstrated by the similar structure of dendrograms of clustering with these measures in Figure 2. The best clusters are provided by Cramér's V in the three-cluster solution.

Table 2. Values of the WCM coefficient for clustering of binary variables (SLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.378	0.299	0.232	0.186
Coefficient U	0.372	0.333	0.232	0.186
IOF measure	0.376	0.307	0.245	0.172
OF measure	0.376	0.307	0.245	0.190
Lin measure	0.376	0.307	0.245	0.190
Overlap measure	0.376	0.307	0.245	0.190

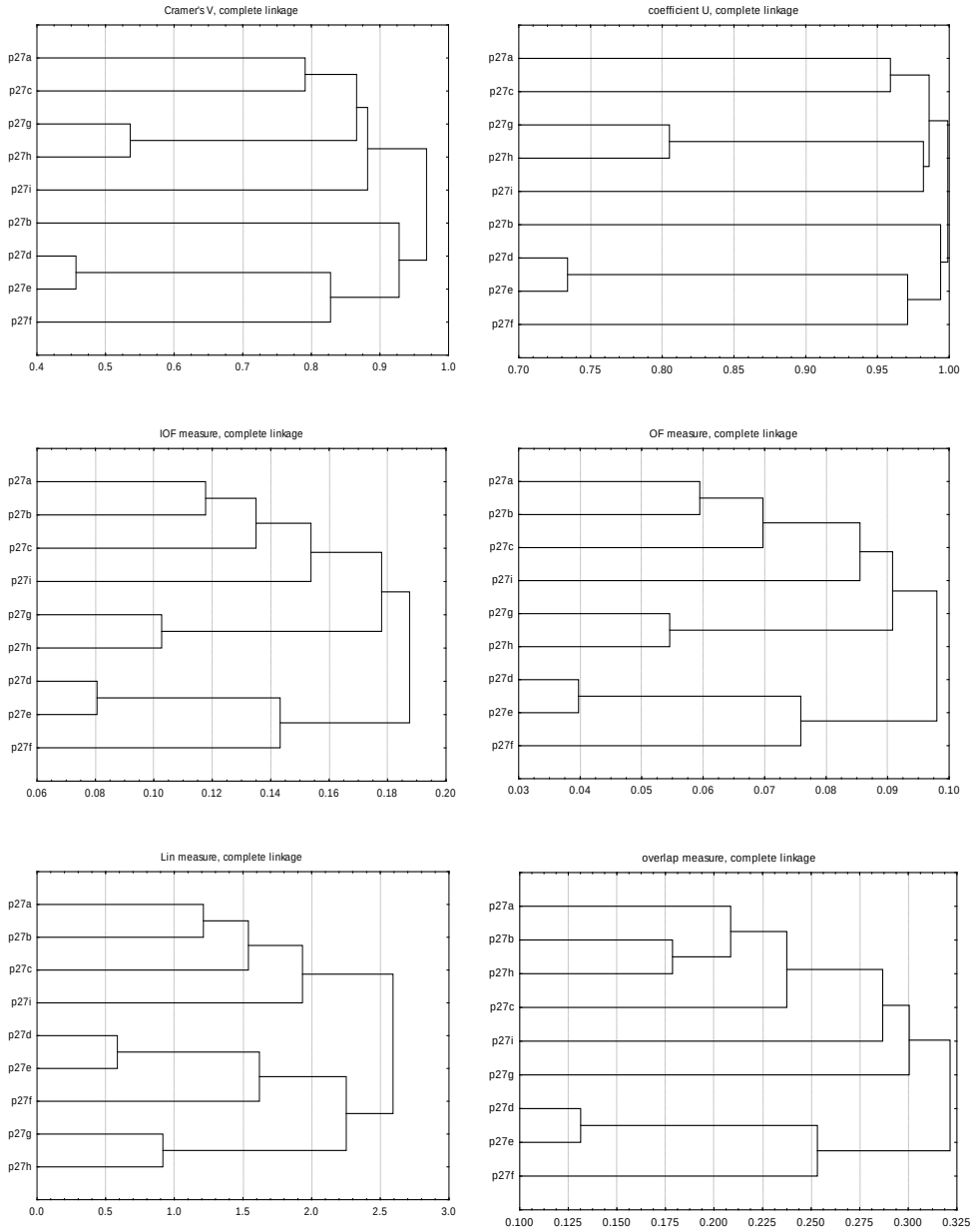

Figure 1. Dendrograms for clustering of binary variables (CLM)

It is important to note that the distances between pairs of variables are differentiated much worse by SLM than by CLM. This fact can cause a bad assignment of clusters into new ones when performing the agglomerative process,

because there are very small differences in their distance by SLM. Especially, such situations are noticeable by the uncertainty coefficient and the IOF measure in Figure 2.

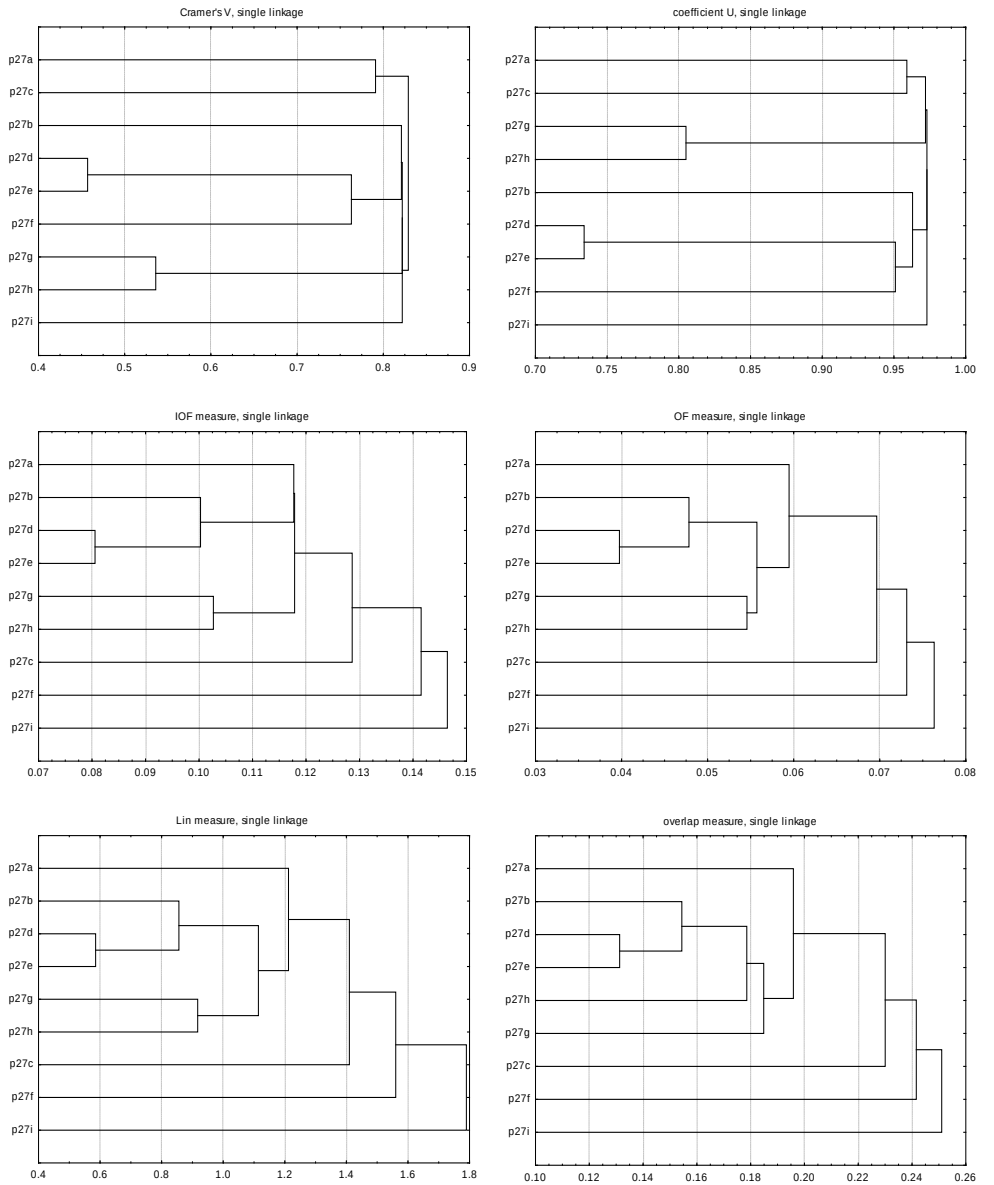

Figure 2. Dendrograms for clustering of binary variables (SLM)

When evaluating the WCM results obtained on the basis of ALM, one can observe that they lie somewhere in between the results of CLM and SLM, as shown in Table 3. However, their structure is much more similar to the one of

CLM, as demonstrated in Figure 3. When examining the dendrograms, one can notice that the distances between clusters are not as large as by CLM, but they are considerably larger than by SLM. The best clusters are provided by the IOF measure. Actually, they are exactly the same as when using CLM.

Table 3. Values of the WCM coefficient for clustering of binary variables (ALM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.366	0.320	0.254	0.186
Coefficient U	0.366	0.333	0.232	0.186
IOF measure	0.366	0.297	0.232	0.168
OF measure	0.375	0.307	0.234	0.172
Lin measure	0.366	0.300	0.232	0.168
Overlap measure	0.375	0.307	0.238	0.177

In the binary variable set, the best clusters are provided by IOF only when using SLM Cramér's V provides better results. Unfortunately, it is not that they are good but because the other measures perform much worse. All the recently developed similarity measures have satisfying results when using CLM or ALM. In the end, the three-cluster solution of the IOF measure by CLM was chosen. The clusters look as follows. In the first cluster, there are variables regarding the kind of work (p27a – *worked part-time*, p27b – *worked in shifts*, p27c – *worked flextime*, p27i – *cheated at work*). The second cluster summarizes variables concerning changing a job (p27d – *changed a job*, p27e – *changed a profession*, p27f – *moved*). The third cluster describes variables regarding a refusal of a good offer in a job (p27g – *refused a job offer*, p27h – *refused a promotion offer*).

4.3. Three-category variable clustering

The within-cluster variability for two- to five-cluster solutions using CLM for three-category variables is contained in Table 4. The results are not as unambiguous as by the binary variables. In the two-cluster solution, the best results provide both the OF and the overlap measure. In the three-cluster solution, there is a different situation; both IOF and Lin have the best results. All the association measures provide worse results in comparison to other similarity measures, which have very similar results of the WCM coefficient.

Table 4. Values of the WCM coefficient for clustering of three-category variables (CLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.416	0.354	0.287	0.208
Coefficient U	0.427	0.352	0.287	0.208
IOF measure	0.385	0.317	0.259	0.208
OF measure	0.381	0.322	0.261	0.196
Lin measure	0.385	0.317	0.259	0.194
Overlap measure	0.381	0.321	0.260	0.195

Looking at the dendrograms in Figure 4, it is apparent that they can be divided into three groups according to the clustering structure. The first group contains both the association measures, Cramér's V and the uncertainty coefficient. These measures have a tendency to create unbalanced clusters; all of them provide at least one cluster comprising only one variable. The second group includes IOF and Lin, and in the last group, there are OF and overlap. According to dendrograms interpretation, the best results are provided by the Lin measure, which has, except for the five-cluster solution, the same results as the IOF measure.

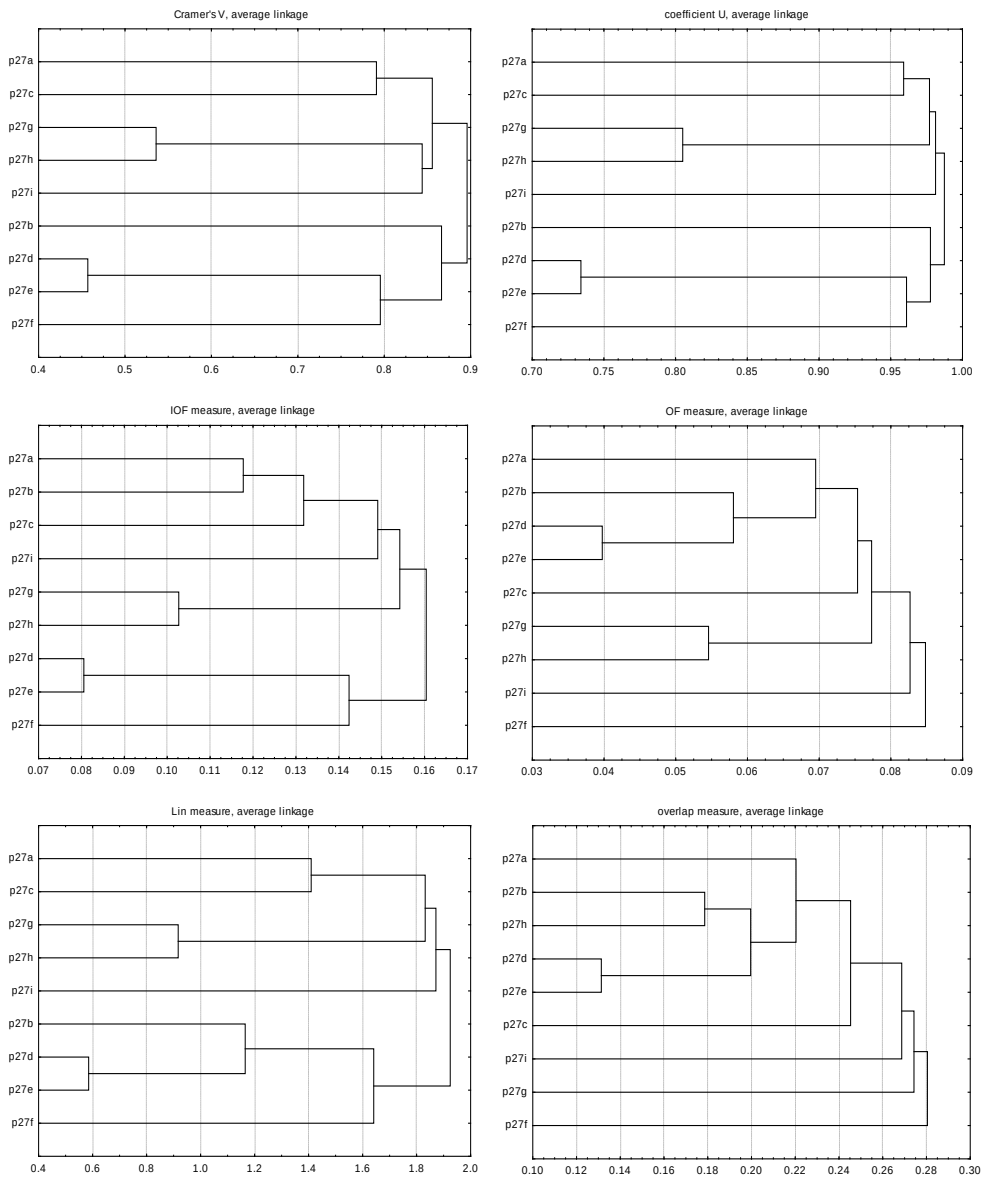

Figure 3. Dendrograms for clustering of binary variables (ALM)

When using SLM, the within-cluster variability of similarity measures in a particular cluster solution is expressed in Table 5. Similarly as by the binary data set, the clustering results are much worse than by CLM. Except for the IOF measure, all other similarity measures provide very unbalanced clusters, which often contain only one variable.

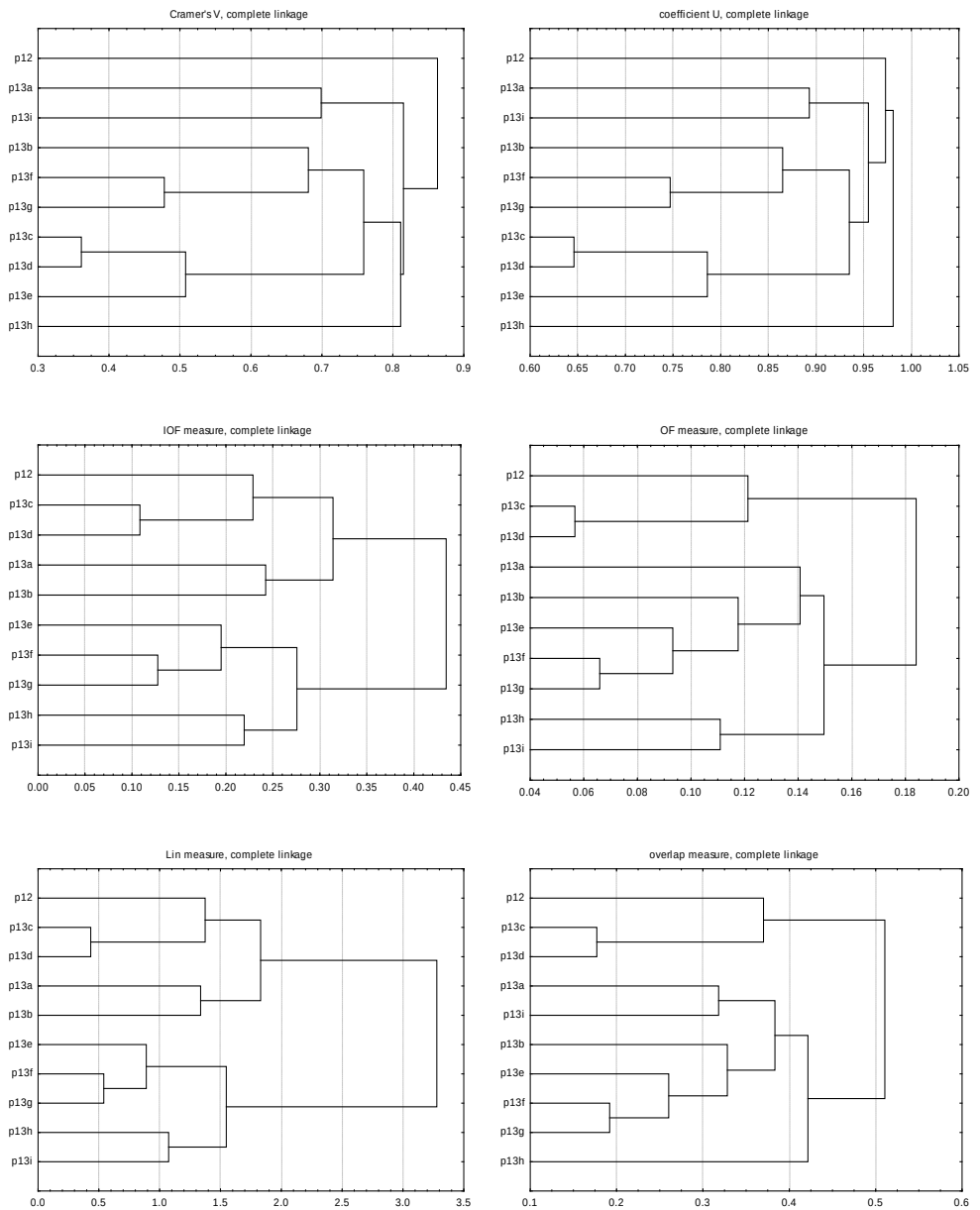

Figure 4. Dendrograms for clustering of three-category variables (CLM)

Table 5. Values of the WCM coefficient for clustering of three-category variables (SLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.416	0.352	0.295	0.248
Coefficient U	0.427	0.352	0.287	0.240
IOF measure	0.381	0.329	0.267	0.202
OF measure	0.416	0.358	0.295	0.198
Lin measure	0.416	0.358	0.257	0.202
Overlap measure	0.416	0.358	0.295	0.198

According to the dendrograms in Figure 5, the OF and overlap measures provide clusters in a similar way. Again, the advantage of recently developed similarity measures, which take into account the frequency distribution of categories, does not seem to have a big importance by SLM. The best clusters are provided by the IOF measure, but they do not reach the quality of the same measure by CLM.

The values of the WCM coefficient for ALM are displayed in Table 6. They are very similar to those provided by CLM; they differ only in details. The overlap measure has the best results across all cluster solutions. It is closely followed by the recently developed similarity measures and then by the association measures.

Table 6. Values of the WCM coefficient for clustering of three-category variables (ALM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.416	0.352	0.295	0.209
Coefficient U	0.427	0.352	0.287	0.208
IOF measure	0.381	0.317	0.267	0.208
OF measure	0.381	0.322	0.263	0.198
Lin measure	0.385	0.323	0.259	0.202
Overlap measure	0.381	0.316	0.256	0.198

When looking at the dendrograms displaying the ALM clustering in Figure 6, one might see that some of them have a similar structure to CLM (the uncertainty coefficient, the overlap measure, and all the recently developed similarity measures). Thus, some measures provide similar results to SLM and some to CLM. The best results are provided by the overlap measure.

Generally, in the three-category variable set, the best results are provided by the IOF measure. Outputs based on this measure are not the best in all cluster solutions; however, they are very robust in most situations. Actually, the best results by CLM, the Lin measure, and by ALM, the overlap measure, were the same as those provided by the IOF measure. By SLM, the IOF measure performed beyond competition.

The two-cluster solution obtained by CLM with the IOF measure was considered to be the best one. The first cluster deals with variables concerning getting a job (p13a – *to get a job*, p13b – *to have better salary for the same job*, p13c – *to get a leadership*, p13d – *to be a director* and p12 – *a chance of success*). The second cluster consists of variables regarding getting a better position in a respondent's job: (p13e – *to promote*, p13f – *for a salary increase*, p13g – *to gain benefits*, p13h – *to have authority*, p13i – *to keep a job*).

5. Conclusion

In this paper, clustering performance of two kinds of similarity measures was examined: the association measures for nominal variables and the other similarity measures originally proposed for objects characterized by nominal variables. There were two main aspects of the comparison. Firstly, the final cluster solutions were evaluated from the point of view of the within-cluster variability; secondly, on the basis of dendrograms and judgments of the researcher. For the analysis, sets of binary and three-category variables were chosen. The influence of different types of linkage methods on resulting clusters was also examined.

Overall, six similarity measures were evaluated in this paper. There were two association measures and four other similarity measures. The association measures, Crammer's V and the uncertainty coefficient, focus on general dependence between two variables when determining their similarity. However, this way of similarity measuring may lead to a loss of some part of information, and thus, to worse dissimilarity determination. The results of the within-cluster mutability (WCM) coefficient and clusters unbalanced by this measures confirmed such a scenario. Therefore, the use of association measures is not suitable for clustering of nominal variables in cases where other possibilities can be considered.

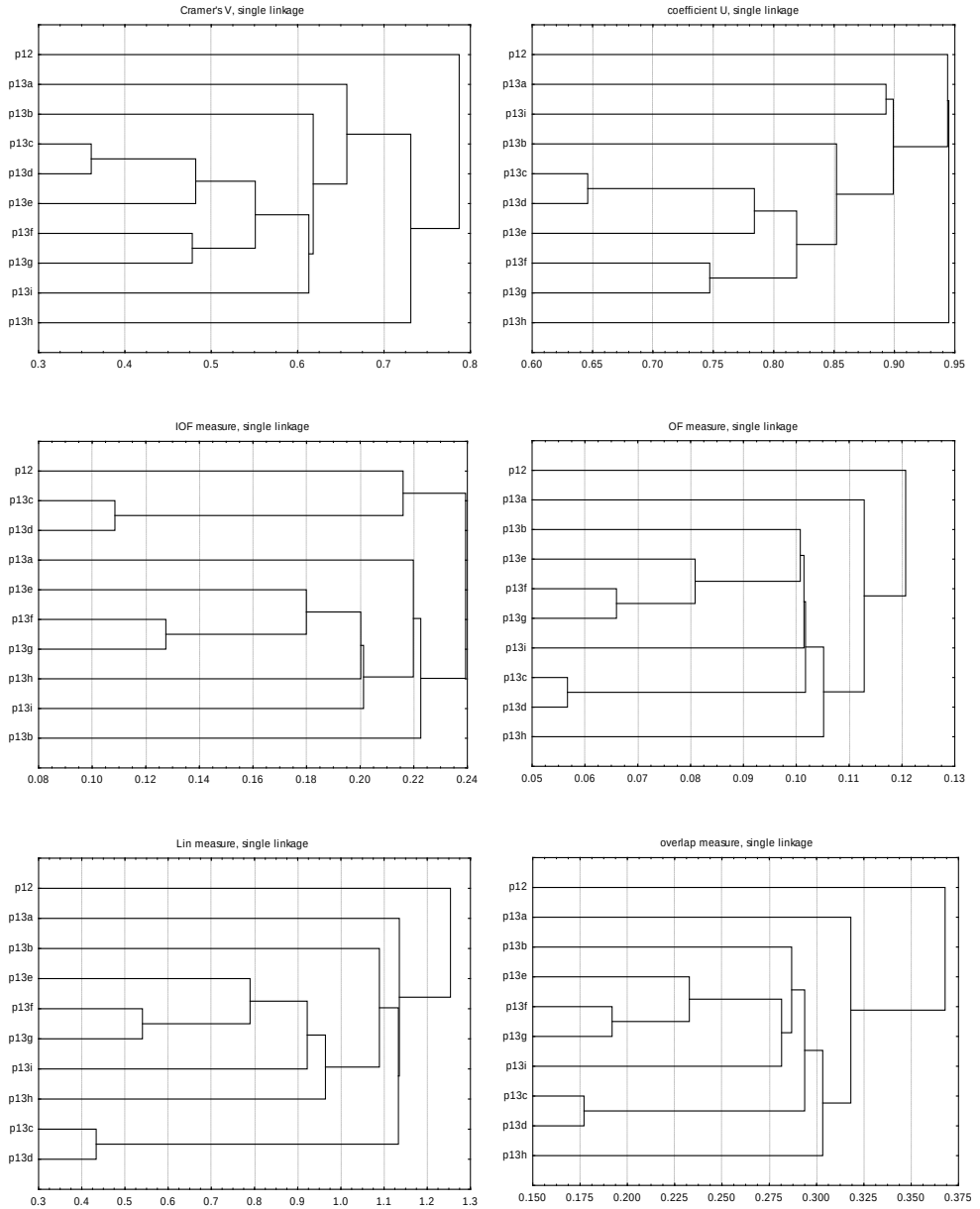

Figure 5. Dendrograms for clustering of three-category variables (SLM)

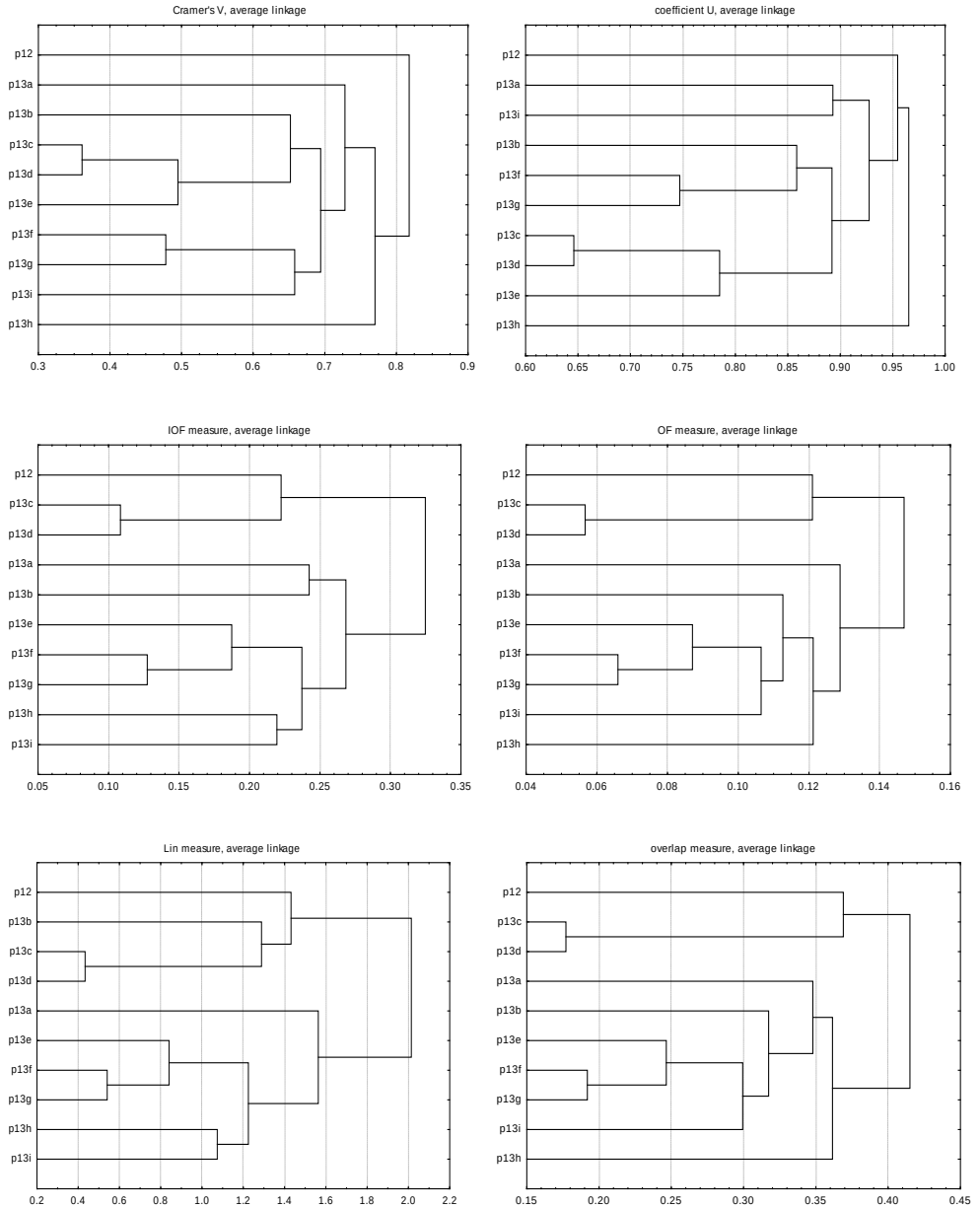

Figure 6. Dendrograms for clustering of three-category variables (ALM)

Four other similarity measures were examined: IOF, OF, Lin and overlap, which compare categories taken pairwise for each object individually, and which all require the same number of categories with the same meaning. Those measures differ mainly in their weight systems. The IOF measure assigns higher weights to

mismatches on less frequent categories which allows it to be more sensitive to outliers in a data set. This approach proved to be more successful in comparison with the OF measure, which uses exactly the opposite weight system, which puts lower weights to those outliers. The Lin measure, as well as the OF measure, assigns lower weights to less frequent categories in the case of a mismatch, but more than that, it also assigns higher weights to more frequent categories in the case of a match. This makes its results very robust in comparison with the OF measure. The overlap measure offers no weight system. This measure provided similar results of the WCM coefficient with the rest of other similarity measures; however, the crucial difference was in cluster quality of resulting clusters. They were unbalanced and their dendrogram interpretation was worse than the rest of the other measures. On the whole, the IOF and Lin measures provided very good clusters of variables in both data sets from the aspects of the WCM coefficient as well as the dendrogram interpretation. Therefore, the use of one of these measures is highly recommended for variable clustering.

When comparing the three linkage methods, the best results are provided by the complete one. It provides good differentiation of clusters; thus, it is easy to cut a dendrogram at a given point. Further, it creates clusters of a similar size, which is in accordance with reduction of a data set. The single linkage method provides very different results in comparison to the complete and average linkage methods. Moreover, the adjustments of recently developed similarity measures, which take into account frequency distribution of categories, do not seem to have any strong influence because of this method. On the whole, this method offers the worst results of all the examined linkage methods; therefore, it cannot be recommended for variable clustering. Thus, the complete or average linkage method should be preferred.

Acknowledgement

This work was supported by the University of Economics, Prague under the project IGS F4/104/2014.

REFERENCES

ANDERBERG, M. R., (1973). Cluster Analysis for Applications. Academic Press, New York.

BORIAH, S., CHANDOLA, V., KUMAR, V., (2008). Similarity measures for categorical data: a comparative evaluation. In: Proceedings of the 8th International Conference on Data Mining. SIAM, pp. 243–254.

CHANDOLA, V., BORIAH, S., KUMAR, V., (2009). A framework for exploring categorical data. In: Proceedings of the 9th International Conference on Data Mining. SIAM, pp. 187–198.

CHAVENT, M., KUENTZ, V., LIQUET, B., SARACCO, L., (2012). ClustOfVar: An R package for the clustering of variables. Journal of Statistical Software, 50(13):1–16. Available at: <http://arxiv.org/abs/1112.0295> [Accessed: 16 October 2014].

CHAVENT, M., KUENTZ, V., SARACCO, J., (2010). A partitioning method for the CLUSTERING of categorical variables. In: Locarek-Junge, H., Weihs, C., eds, Classification as a Tool for Research. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin Heidelberg, pp. 91–99.

D'ENZA, A. I., GREENACRE, M. J., (2012). Multiple correspondence analysis for the quantification and visualization of large categorical data sets. In: Advanced Statistical Methods for the Analysis of Large Data-Sets. Springer, Berlin Heidelberg, pp. 453–463.

EVERITT, B. S., LANDAU, S., LEESE, M., STAHL, D., (2011). Cluster Analysis, 5th edn, Wiley, Chichester.

GAN, G., MA, C., WU, J., (2007). Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM, Philadelphia.

GORDON, A. D., (1999). Classification, 2nd edn, Chapman & Hall/CRC, Boca Raton.

GREENACRE, M. J., (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5):613–619.

JOLLIFFE, I. T., (2002). Principal Component Analysis, 2nd edn, Springer, New York.

LIN, D., (1998). An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning. Morgan Kaufmann, San Francisco, pp. 296–304.

PALLA, K., KNOWLES, D. A., GHAHRAMANI, Z., (2012). A nonparametric variable clustering model. In: Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q., eds, Advances in Neural Information Processing Systems 25. NIPS Foundation. Available at: <http://papers.nips.cc/paper/4579-a-nonparametric-variable-clustering-model.pdf> [Accessed 16 October 2014].

PAYNE, T. R., EDWARDS, P., (1999). Dimensionality reduction through correspondence analysis. Available at: <http://eprints.soton.ac.uk/263091/> [Accessed 16 October 2014].

ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D., (2011). Evaluation of categorical data clustering. In: Mugellini, E., Szczepaniak, P. S., Pettenati, M. C. et al., eds, Advances in Intelligent Web Mastering 3. Springer Verlag, Berlin, pp. 173–182.

ŘEZANKOVÁ, H., (2014). Nominal variable clustering and its evaluation. In: Proceedings of the 8th International Days of Statistics and Economics. Melandrium, Slaný, pp. 1293–1302. Available at: < http://msed.vse.cz/msed_2014/article/276-Rezankova-Hana-paper.pdf > [Accessed 5 November 2014].

SPARCK-JONES, K., (1972, 2002). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1):11–21. Later: Journal of Documentation, 60(5):493–502.

ŠULC, Z., ŘEZANKOVÁ, H., (2014). Evaluation of recent similarity measures for categorical data. In: Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wroclaw, pp. 249–258. Available at: < http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf> [Accessed 5 November 2014].

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 611–626

FUNCTIONAL REGRESSION IN SHORT-TERM PREDICTION OF ECONOMIC TIME SERIES

Daniel Kosiorowski[1]

ABSTRACT

We compare four methods of forecasting functional time series including fully functional regression, functional autoregression FAR(1) model, Hyndman & Shang principal component scores forecasting using one-dimensional time series method, and moving functional median. Our comparison methods involve simulation studies as well as analysis of empirical dataset concerning the Internet users behaviours for two Internet services in 2013. Our studies reveal that Hyndman & Shao predicting method outperforms other methods in the case of stationary functional time series without outliers, and the moving functional median induced by Frainman & Muniz depth for functional data outperforms other methods in the case of smooth departures from stationarity of the time series as well as in the case of functional time series containing outliers.

Key words: functional data analysis, functional time series, prediction.

1. Introduction

A variety of economic phenomena directly leads to functional data: yield curves, income densities, development trajectories, price trajectories, life of a product, and electricity or water consumption within a day (see Kosiorowski et al. 2014). The Functional Data Analysis (FDA) over the last two decades proved its usefulness in the context of decomposition of income densities or yield curves, analyses of huge, sparse economic datasets or analyses of ultra-high frequency financial time series. The FDA enables an effective statistical analysis when the number of variables exceeds the number of observations. Using FDA we can effectively analyse economic data streams, i.e., for example, perform an analysis of non-equally spaced observed time series, which cannot be predicted using, e.g. common moving average or ARIMA framework, by analysing or predicting a whole future trajectory of a stream rather than iteratively predict single observations.

[1] Department of Statistics, Faculty of Management, Cracow University of Economics. E-mail: dkosioro@uek.krakow.pl.

Using a functional regression where both the predictor as well as the response are functions, we can express relations between complex economic phenomena without dividing them into parts. Recently proposed models for functional time series give us a hope for overcoming the so-called *curse of dimensionality* related to nonparametric analysis of huge economic data sets (see Horvath and Kokoszka, 2012). From other perspective, functional medians defined within the *data depth concept* for functional objects may have useful applications in the context of robust time series analysis – in the case of existence paths of outliers in the data.

The analysis of functional time series (FTS) was considered, among others, in the literature in the contexts of: breast cancer mortality rate modelling and forecasting, call volume forecasting, climate forecasting, demographical modelling and forecasting, electricity demand forecasting, credit card transaction and Eurodollar futures (see Ferraty, 2011 for an overview), yield curves and the Internet users behaviours forecasting (Kosiorowski et al. 2014b), extraction of information from huge economic databases (Kosiorowski et al. 2014a).

The FTS undoubtedly brings up conceptually new areas of economic research and provides new methodology for applications. It is not clear, however, which approaches proposed in the FTS literature up to now are the most promising in the context of FTS prediction. The main aim of this paper is to compare main approaches for FTS prediction using real data set related to day and night Internet users behaviours in 2013. Our paper refers to similar simulation studies of the selected FTS prediction methods presented in Didieriksen et al. (2011) and Besse et al. (2000). Additionally, we considered Hyndeman and Shang (2010) nonparametric FTS prediction and moving Frainman & Muniz functional median forecasting methods.

The rest of the paper is organized as follows. In Section 2 we briefly describe selected approaches for FTS prediction. In Section 3 we compare the approaches using empirical examples. We conclude with Section 4 which discusses advantages and disadvantages of the approaches presented in Section 2.

2. Functional time series prediction

2.1. Preliminaries – functional time series

Functions considered within the FDA are usually elements of a certain separable Hilbert space H with certain inner product $\langle \cdot, \cdot \rangle$ which generates a norm $\|\cdot\|$. A typical example is a space $L^2 = L^2([t_0, t_L])$ - a set of measurable real-valued functions x defined on $[t_0, t_L]$ satisfying $\int_{t_0}^{t_L} x^2(t)dt < \infty$. The space L^2 is a separable Hilbert space with an inner product $\langle x, y \rangle = \int x(t)y(t)dt$. We

usually treat the random curve $X = \{X(t), t \in [t_0, t_L]\}$ as a random element of L^2 equipped with the Borel σ algebra. Recently, within a nonparametric FDA, authors have successfully used certain wider functional spaces, i.e. for example, Sobolev spaces (Ferraty and Vieu, 2006).

In order to apply FDA into the economic researches, first we have to transform discrete observations into functional objects using smoothing, kernel methods or orthogonal systems representations. Then we can calculate and interpret functional analogues of basic descriptive measures such as mean, variance and covariance (for details see Ramsay and Silvermann, 2005; Górecki and Krzyśko, 2012).

For the iid observations $X_1, X_2, ..., X_N$ in L^2 with the same distribution as X, which is assumed to be square integrable we can define the following **descriptive characteristics**:

$$\mu(t) = E[X(t)], \text{ mean function,} \tag{1}$$

$$c(t,s) = E\left[(X(t) - \mu(t))(X(s) - \mu(s))\right], \text{ covariancefunction,} \tag{2}$$

$$C = E\left[\langle X - \mu, \cdot \rangle (X - \mu)\right], \text{ covariance operator} \tag{3}$$

and correspondingly their **sample estimators**

$$\hat{\mu}(t) = N^{-1} \sum_{i=1}^{N} X_i(t), \tag{4}$$

$$\hat{c}(t,s) = N^{-1} \sum_{i=1}^{N} (X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(t)), \tag{5}$$

$$\hat{C}(x) = N^{-1} \sum_{i=1}^{N} \langle x_i - \hat{\mu}, x \rangle (x_i - \hat{\mu}),\ x \in L^2, \tag{6}$$

It is worth noting that $\hat{C}$ maps L^2 into a finite dimensional subspace spanned by $X_1, X_2, ..., X_N$.

A **functional analogue** of the **principal component analysis** plays a central role in the FTS. For a covariance operator C, the *eigenfunctions* v_j and the *eigenvalues* λ_j are defined by $Cv_j = \lambda_j v_j$, so if v_j is an eigenfunction, then so is av_j – for any nonzero scalar a. The v_j are typically normalized so that $\|v_j\| = 1$.

In a sample case we define the estimated eigenfunctions $\hat{v}_j$ and eigenvalues by

$$\int \hat{c}(t,s)\hat{v}_j(s)ds = \hat{\lambda}_j \hat{v}_j(t),\ \ j = 1, 2, ..., N, \tag{7}$$

where $\hat{c}(t,s)$ denotes estimated covariance function (see Górecki and Krzyśko, 2012).

Let $y_t(x)$ denote a function, such as monthly income for the continuous age variable x in year t. We assume that there is an underlying smooth function $f_t(x)$ which is observed with an error at discretized grid points of x. **A special case of functional time series $\{y_t(x)\}_{t\in\square}$ is when the continuous variable x is also a time variable**. For example, let $\{Z_w, w\in[1,N]\}$ be a *seasonal time series* which has been observed at N equispaced time points. We divide the observed time series into n trajectories, and then consider each trajectory of length p as a curve rather than p distinct data points. The functional time series is then given by

$$y_t(x)=\{Z_w, w\in(p(t-1), pt]\}\ ,\ t=1,2,...,n\,. \tag{8}$$

The problem of interest is to forecast $y_{n+h}(x)$, where h denotes forecast horizon.

In the context of FTS prediction, several methods have been considered in the literature up to now. Ramsay and Silverman (2005) and Kokoszka (2007) studied several functional linear models. Theoretical background related to the prediction using functional autoregressive processes can be found in Bosq (2000). Functional kernel prediction was considered in Ferraty and Vieu (2006), Ferraty (2011). An application of a functional principal component regression to FTS prediction can be found in Shang and Hyndeman (2011).

For evaluating prediction quality of main approaches for FTS prediction in the case of our empirical data set related to the Internet users of certain services analysis, we refer to frameworks presented in two finite sample studies: Besse et al. (2000) and Didericksen et al. (2011). Within simulation studies, these authors have studied predictions at time n errors E_n and R_n, $1<n<N$, defined in the following way:

$$E_n=\sqrt{\int_{t_0}^{t_L}\left(X_n(t)-\hat{X}_n(t)\right)^2 dt}\,, \tag{9}$$

$$R_n=\int_{t_0}^{t_L}\left|X_n(t)-\hat{X}_n(t)\right|dt\,, \tag{10}$$

for several N=50, 100, 200, several processes models and innovation processes.

2.2. Prediction using fully functional model

In the simple linear regression we consider observations from the following point of view

$$Y_i=\beta_0+\beta_1 x_i+\varepsilon_i\,, i=1,2,...,N\,, \tag{11}$$

where all random variables Y_i as well the regressors x_i are scalars.

In the case of a *functional linear model*, predictors, responses as well as analogues of the coefficients β_0 and β_1 may be curves and have to be appropriately defined.

The **fully functional** model is defined as

$$Y_i(t) = \int \psi(t,s) X_i(s) ds + \varepsilon_i(t), \; i = 1, 2, ..., N \;, \tag{12}$$

where responses Y_i are curves and so are regressors X_i.

The fully functional model can alternatively be written as

$$Y(t) = \int X(s)\beta(s,t) ds + \varepsilon(t), \tag{13}$$

where $\beta(s,t) = \psi(t,s)$, $Y(t) = \left[Y_1(y), ..., Y_N(t)\right]^T$, $X(s) = \left[X_1(s), ..., X_N(s)\right]^T$, and $\varepsilon(t) = \left[\varepsilon_1(t), ..., \varepsilon_N(t)\right]^T$.

Suppose $\{\eta_k, k \geq 1\}$ and $\{\theta_l, l \geq 1\}$ are some bases which need not be orthonormal. Assume that the functions η_k are suitable for expanding the functions X_i and θ_i for expanding the Y_i. For estimating the kernel $\beta(\cdot,\cdot)$, let us consider estimates of the form

$$\beta^*(s,t) = \sum_{k=1}^{K} \sum_{l=1}^{L} b_{kl} \eta_k(s) \theta_l(t), \tag{14}$$

in which K and L are relatively small numbers which are used as smoothing parameters.

We obtain a **least squares estimator** by finding b_{kl} which minimizes the residual sum

$$\sum_{i=1}^{N} \left\| Y_i - \int X_i(s) \beta^*(s,\cdot) \right\|^2 . \tag{15}$$

Derivation of normal equations can be found in Horvath and Kokoszka (2012). Alternative estimators for (14) can be found in Ramsay and Silverman (2005), where authors used large *K* and *L* but introduced a roughness penalty on the estimates.

Effective application of the model (12) relates to fulfilling an assumption that the conditional expectation $E[Y(t) \mid X]$ is a linear function of X. It is worth noting that within the functional regression setup it is possible to perform an analogue of regression diagnostics using **functional residuals** defined as

$$\hat{\varepsilon}_i(t) = Y_i(t) - \int \hat{\psi}(t,s) X_i(s) ds, \; i = 1, 2, ..., N \;, \tag{16}$$

and calculate an analogue of the **coefficient of determination**

$$R^2(t) = \frac{Var\left[E\left[Y(t) \mid X\right]\right]}{Var\left[Y(t)\right]}, \tag{17}$$

note that since $Var\left[E\left[Y(t) \mid X\right]\right] \le Var\left[Y(t)\right]$, $0 \le R^2(t) \le 1$. The coefficient $R^2(t)$ quantifies the degree to which the functional linear model explains the variability of the response curves at a fixed point t. For the global measure we can integrate $R^2(t)$.

2.3. Hyndman & Shang FPC regression

Let $\mathbf{f}(\mathbf{x}) = \left[f_1(x), f_2(x), ..., f_n(x)\right]^T$ denote a sample of functional data. Note that at a population level, a stochastic process denoted by f can be decomposed into the mean function and the products of orthogonal functional principal components and uncorrelated principal component scores. It can be expressed as

$$f = \mu + \sum_{k=1}^{\infty} \beta_k \phi_k , \tag{18}$$

where μ is the unobservable population mean function, β_k is the kth principal component score. Assume that we observe n realizations of f evaluated on a compact interval $x \in [t_0, t_L]$, denoted by $f_t(x)$, for $t = 1, 2, ..., n$. At a sample level, the functional **principal component decomposition** can be written as

$$f_t(x) = \overline{f}(x) + \sum_{k=1}^{K} \hat{\beta}_{t,k} \hat{\phi}_k(x) + \hat{\varepsilon}_t(x), \tag{19}$$

where $\overline{f}(x) = n^{-1} \sum_{t=1}^{n} f_t(x)$ is the estimated mean function, $\hat{\phi}_k(x)$ is the kth estimated orthonormal eigenfunction of the empirical covariance operator

$$\hat{C}(x) = n^{-1} \sum_{t=1}^{n} [f_t(x) - \overline{f}(x)][f_t(x) - \overline{f}(x)]. \tag{20}$$

The coefficient $\hat{\beta}_{t,k}$ is the kth principal component score for year t. It is given by the projection of $f_t(x) - \overline{f}(x)$ in the direction of kth eigenfunction $\hat{\phi}_k(x)$, that is,

$$\hat{\beta}_{t,k} = \left\langle f_t(x) - \overline{f}(x), \hat{\phi}_k(x) \right\rangle = \int_x [f_t(x) - \overline{f}(x)] \hat{\phi}_k(x) dx , \tag{21}$$

where $\hat{\varepsilon}_t(x)$ is the residual, and K is the optimal number of components, which can be chosen for example by cross validation.

By conditioning on the set of smoothed functions $\mathbf{f}(\mathbf{x}) = \left[f_1(x), f_2(x), ..., f_n(x) \right]^T$ and the fixed functional principal components $B = \left[\hat{\phi}_1(x), \hat{\phi}_2(x), ..., \hat{\phi}_K(x) \right]^T$, **the Hyndman and Shangh-step-ahead forecast** of $y_{n+h}(x)$ can be obtained as

$$\hat{y}_{n+h|n}(x) = E[y_{n+h}(x) \,|\, \mathbf{f}(\mathbf{x}), \mathbf{B}] = \overline{f}(x) + \sum_{k=1}^{K} \hat{\beta}_{n+h|n,k} \hat{\phi}_k(k), \tag{22}$$

where $\hat{\beta}_{n+h|n,k}$ denotes **the h-step-ahead forecast of** $\beta_{n+h,k}$ using **univariate time series forecasting** methods (i.e., for example, ARIMA, linear exponential smoothing).

Note: because of orthogonality, the forecast variance can be approximated by the sum of component variances.

2.4. Moving functional median

For one dimensional sample $X^N = \{X_1, X_2, ..., X_N\}$ and empirical cumulative density function (ecdf) $F_N(x) = N^{-1} \sum_{n=1}^{N} I\{X_i \le x\}$ we can define the **halfspace depth** of X_i as

$$HD_N(x_i) = \min\left\{ F_N(x_i), 1 - F_N(x_i) \right\}. \tag{23}$$

We can obtain another one-dimensional depth using the following formula

$$D_N(x_i) = 1 - \left| 1/2 - F_N(x_i) \right|. \tag{24}$$

For N functions $\left\{ X_i(t), t \in [t_0, t_L] \right\}$ and $F_{N,t}(x) = N^{-1} \sum_{n=1}^{N} I\left\{ X_i(t) \le x \right\}$ we can define a **functional depth** by integrating one of the univariate depth (see Zuo and Serfling, 2000 or Kosiorowski, 2012 for a detailed introduction to the data depth concept).

Frainman and Muniz (2001) proposed to calculate the depth of the curve as

$$FD_N(X_i \mid X^n) = \int_{t_0}^{t_L} \left[1 - \left| 1/2 - F_{N,t}(X_i(t)) \right| \right] dt. \tag{25}$$

Frainman and Muniz median is defined as

$$MED_{FM}(X^n) = \arg\max_i FD(X_i \mid X^n). \tag{26}$$

We can **predict next observations** by means of the following formula

$$\hat{X}_{n+1}(t) = MED_{FM}(W_{n,k}), \tag{27}$$

where $W_{n,k}$ denotes a **moving window** of length k ending at moment n, i.e., $W_{n,k} = \{X_{n-k+1}(t), ..., X_n(t)\}$.

3. Empirical example

In order to check properties of the selected method of forecasting FTS we considered an empirical example related to behaviours of the Internet users of two services in 2012 and 2013. The services were considered with respect to the number of unique users and number of page views during an hour. Fig. 1 presents raw data for the year 2013. Fig. 2 presents the main idea of obtaining functional time series on the basis of a periodic one-dimensional time series (in the considered series the period equals 24 hours). Fig. 3 – 6 present obtained functional observations for the corresponding number of users in the first service, the number of users in the second service, the numbers of page views in the first service and the number of page views in the second service. Additionally, we added corresponding functional means and Frainman & Muniz functional medians to the Fig. 3 – 6.

We considered a fully functional model, Hyndman and Shang principal component scores forecasting method, Ferraty and View (2006) functional kernel regression, functional autoregressive FAR(1) model described by Horvath and Kokoszka (2012) and estimated by their improved *estimated kernel* method and using moving Frainman and Muniz median. All calculations were conducted using **fda** (Ramsay et al., 2009), **ftsa** (Shang, 2013), **fda.usc (**Febrero-Bande and Oviedo de la Fuente, 2012) and **DepthProc** (Kosiorowski and Zawadzki, 2014). Below we present selected outputs for the methods which performed best within our empirical analysis. In all the situations we used 7–9 spline basis systems for transforming discrete data to the functional objects.

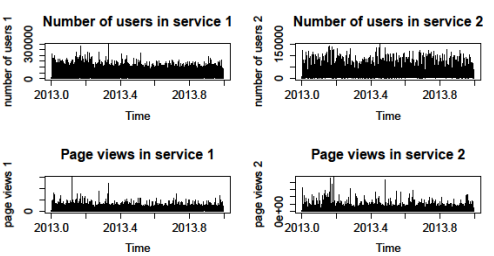

Figure 1. The behaviour of Internet users of two services in 2013

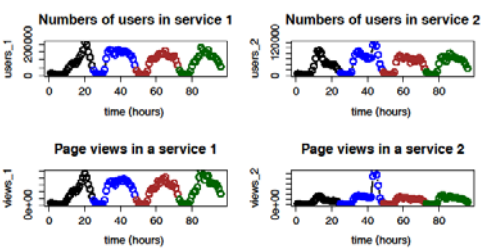

Figure 2. An idea of transformation of the data from univariate to functional time series

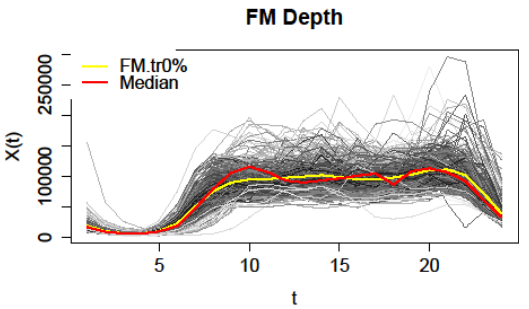

Figure 3. Functional data – number of unique users during 24 hours in service 1

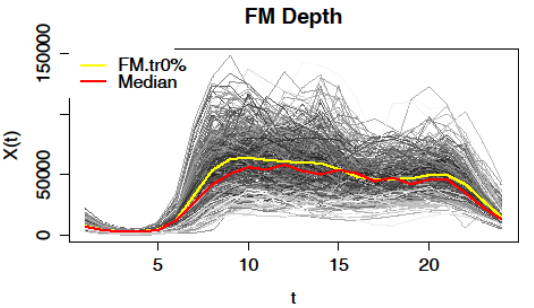

Figure 4. Functional data – number of unique users during 24 hours in service 2

Fig. 7 presents the results of a functional principal component analysis for functional data related to the number of users in the first considered service. We can see there the first two principal component functions and biplots for the observations. It is easy to propose an interpretation according to which the first component relates to using the service at work whereas the second component relates to using the Internet at home. Fig. 7 – 11 present the functional regression method proposed by Hyndman and Shang applied to the corresponding **number of users in the first service, the number of users in the second service, the numbers of page views in the first service** and **the number of page views in the second service**. Each time we used three basis functions (upper panel) and calculated principal component scores (down panel).

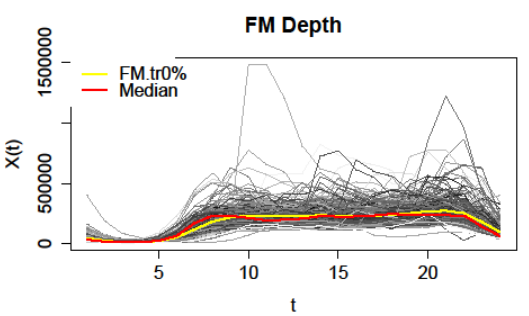

Figure 5. Functional data – number of page views during 24 hours in service 1

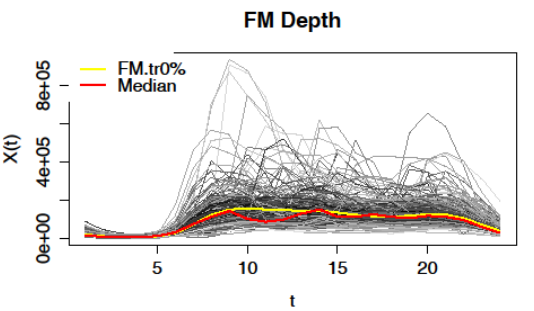

Figure 6. Functional data – number of page views during 24 hours in service 2

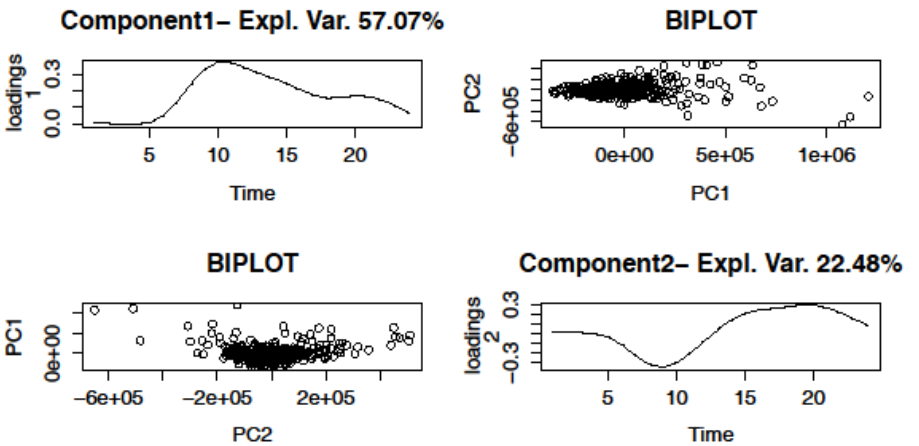

Figure 7. Functional principal components for number of unique users in service 1 in 2013

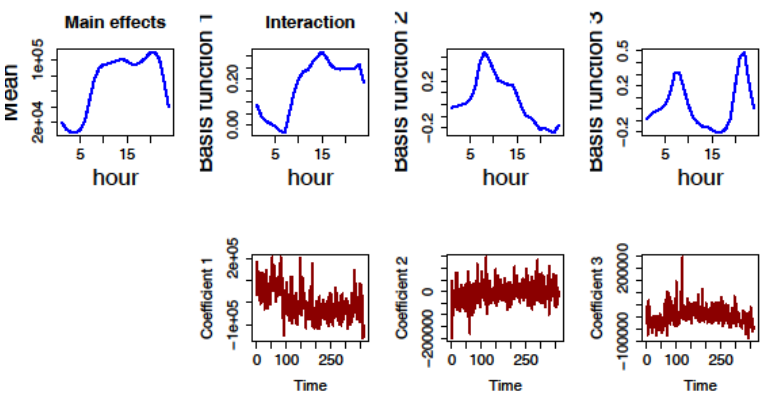

Figure 8. Hyndman & Shang functional PC scores method for number of users in service 1. Three basis function explaining 47%, 18% and 12% variability correspondingly

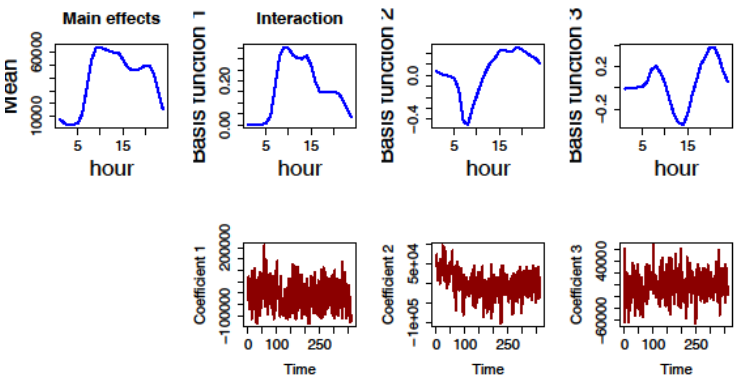

Figure 9. Hyndman & Shang functional PC scores method for number of users in service 2. Three basis function explaining 62%, 15% and 7% variability correspondingly

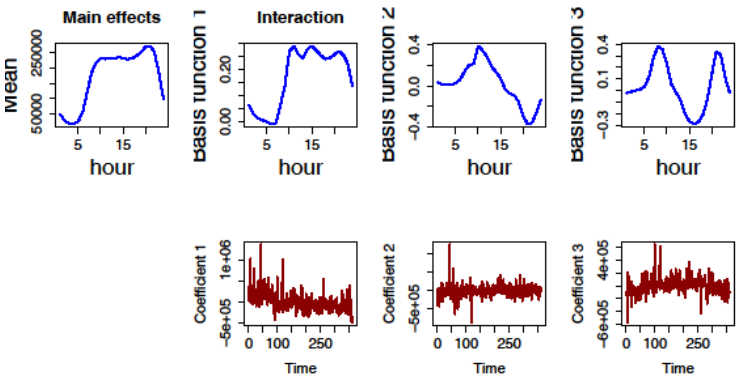

Figure 10. Hyndman & Shang functional PC scores method for number of views in service 1. Three basis function explaining 42%, 20% and 12% variability correspondingly

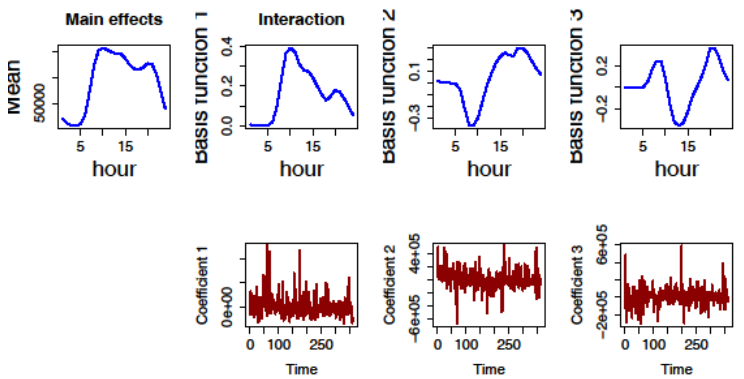

Figure 11. Hyndman & Shang functional PC scores method for number of views in service 2. Three basis function explaining 50%, 20% and 10% variability correspondingly

Fig. 12 – 13 present predictions for the considered examples using Hyndman and Shao method and ARIMA and linear exponential smoothing (ETS) for one-dimensional time series of principal component scores (see Hyndman et al., 2008). Fig. 14 – 15 present observed and predicted values of the number of users in the service 1 and the number of views in the service 1 using moving Frainman and Muniz median calculated from windows consisting of 50 functional observations. Fig. 16 presents observed and predicted values of the number of users in the service 1 calculated using fully linear regression model. Fig. 17 presents residuals in this regression model and Fig. 18 – 19 present an estimated coefficient function for this regression model.

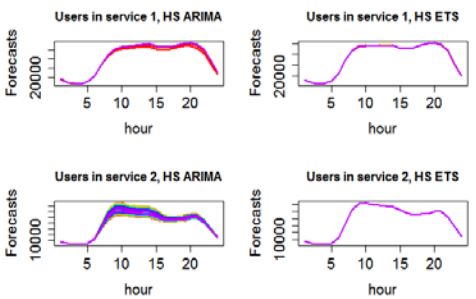

Figure 12. FTS prediction of number of users in the Internet services using Hyndman and Shao FTSA method

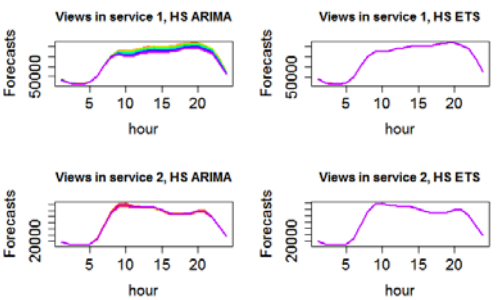

Figure 13. FTS prediction of number of page views in the Internet services using Hyndman and Shao FTSA method

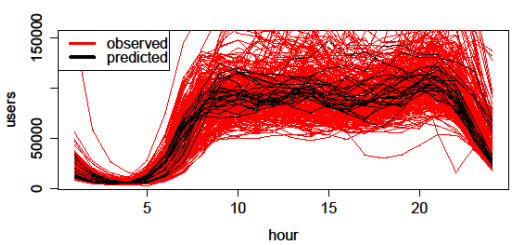

Figure 14. FTS prediction of number of users in the Internet service 1 using moving Frainman & Muniz median

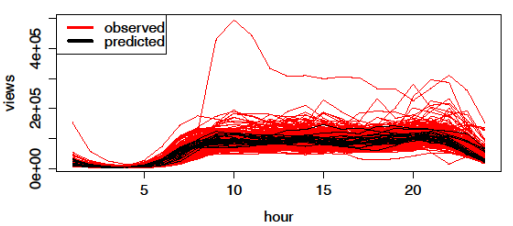

Figure 15. FTS prediction of number of views in the Internet service 1 using moving Frainman & Muniz median

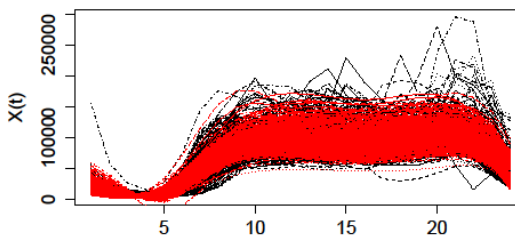

Figure 16. Prediction of number of users in the Internet service 1 using f**ull regression model**

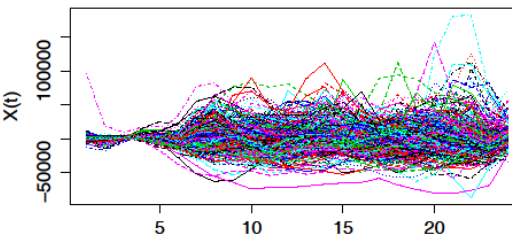

Figure 17. Prediction of number of users in the Internet service 1 using f**ull regression model** – functional residuals

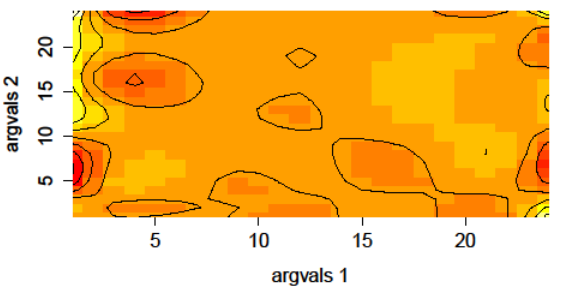

Figure 18. Contour plot: prediction of number of users in the Internet service 1 using full regression model – estimated regression parameters

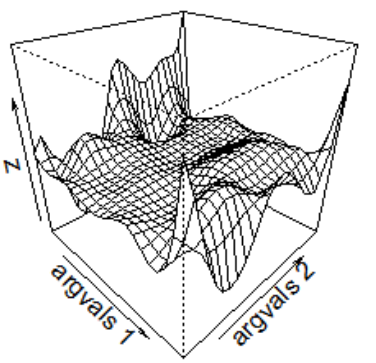

Figure 19. Perspective plot: – prediction of number of users in the Internet service 1 using full regression model – estimated regression parameters

For comparing the methods we divided the data set into two parts of equal sizes. We estimated prediction methods parameters using the first part of the data and tested them using the second part of the data. For testing the methods we used forecast accuracy measures proposed in Didieriksen et al. (2011) defined by formulas (9) and (10). According to our results the Hyndman and Shang method performed best, the moving Frainman and Muniz median performed the second best and the fully linear model was third. Surprisingly, the FAR(1) method as well as the kernel functional regression performed relatively poor in the case of our data set. This finding stays in a contrary to findings of Didieriksen et al. (2011),

where the simulation study was conducted. In the case of our data set, prediction effectiveness of Hyndman and Shang method (100%) in comparison to the moving Frainman and Muniz median and fully linear model was correspondingly as 100% to 91% to 87% in the case of the number of users prediction and as 100% to 99% to 96% in the case of page views prediction. In the case of simulation studies with data simulated from simple nonstationary models (based on models from Didieriksen et al. (2011) for which we changed the mean function and the covariance function) – Frainman and Muniz median performed best.

Additionally, Hyndman and Shang method exhibits the best properties in the context of economic interpretations. The estimated basis functions in a clear way decompose patterns of the Internet behaviour of users. We can easily notice components related to the Internet usage at work as well as the usage at home. The principal component scores time series show importance of the components within the considered period and may be effectively interpreted in a reference to certain political or social events. The eigenvalues corresponding to the eigenfunctions show importance of the particular components for the considered Internet service. We obtained the best predictions using linear exponential smoothing prediction for one-dimensional principal component scores.

In the case of abrupt changes of the data generating mechanism we recommend using moving Frainman and Muniz median which easily adapt the prediction device. It is easy to notice that methods which are based on estimated principal component functions brake down when the covariance operator changes.

Although fully functional model provides complex family of regression diagnostic and goodness of fit measures, its predictive power in the case of our example was below our expectations. Inspection of estimated coefficient function (Fig. 18 – 19) shows relative constant, as to the time arguments t and s, dependency of 24 hour activity of the Internet users.

For all the considered methods, it is possible to calculate the prediction confidence bands. In this context, prediction confidence bands provided by Hyndman and Shang approach based on prediction bands for (uncorrelated) one-dimensional time series prediction seem to be the most informative.

4. Conclusions

The forecasting quality of functional autoregression, fully functional regression and Hyndman & Shang method strongly depend on the stationarity of the underlying functional time series, the choice of a basis system, smoothness of the considered functions, the PCA algorithm used. For the considered empirical example, in the context of prediction as well as explanation of the considered phenomenon Hyndman & Shang method performed best.

The moving Frainman and Muniz functional median performed best in the case of simulated processes containing additive outliers. Conceptually simple, the moving functional median seems to be the most promising in the context of

nonstationary functional time series analysis. The nonstationarity issues relate to our current and future studies.

Acknowledgements

The author thanks for financial support from Polish National Science Centre grant UMO-2011/03/B/HS4/01138.

REFERENCES

BOSQ, D., (2000). Linear Processes in Function Spaces. Springer, New-York.

BESSE, P., C., CARDOT, H., STEPHENSON, D., B., (2000). Autoregressive Forecasting of Some Functional Climatic Variations, Scandinavian Journal of Statistics, Vol. 27, No. 4, 637–687.

DIDIERIKSEN, D., KOKOSZKA, P., ZHANG, Xi, (2011). Empirical properties of forecast with the functional autoregressive model, Computational Statistics, DOI 10.1007/s00180-011-0256-2.

FEBRERO-BANDE, M., OVIEDO DE LA FUENTE, M., (2012). Statistical Computing in Functional Data Analysis: The R Package fda.usc, Journal of Statistical Software, 51(4).

FERRATY, F., VIEU, P., (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer-Verlag.

FERRATY, F., (2011). (ed.) Recent Advances in Functional Data Analysis and Related Topic. Physica-Verlag.

FRAINMAN, R., MUNIZ, G., (2001). Trimmed Means for Functional Data. *Test*, 10, 419–440.

GÓRECKI, T., KRZYŚKO, M., (2012). Functional Principal Component Analysis, in: Pociecha J. and Decker R. (Eds), Data Analysis Methods and its Applications, C.H. Beck, Warszawa 2012, 71–87.

HORVATH, L., KOKOSZKA, P., (2012). Inference for Functional Data with Applications, Springer, New York.

HYNDMAN, R. J., KOEHLER, A. B., ORD, J. B., SNYDER, R. D., (2008). Forecasting with exponential smoothing: the state space approach, Springer-Verlag, Berlin.

HYNDEMAN, H. L., SHANG, H. L., (2009). Forecasting Functional Time Series (with discussion) Journal of the Korean Statistical Society 38(3), 199–221.

KOSIOROWSKI, D., (2012). Statistical Depth Functions in Robust Economic Analysis, Publishing House of CUE in Cracow, Cracow.

KOSIOROWSKI, D., (2015). Two Procedures for Robust Monitoring of Probability Distributions of the Economic Data Stream induced by Depth Functions, Operations Research and Decisions, Vol. 25, No. 1 (in press).

KOSIOROWSKI, D., ZAWADZKI, Z., (2014). DepthProc: An R Package for Robust Exploration of Multidimensional Economic Phenomena, arXiv:1408.4542.

KOSIOROWSKI, D., MIELCZAREK, D., RYDLEWSKI, J., SNARSKA, M., (2014a). Sparse Methods for Analysis of Sparse Multivariate Data from Big Economic Databases, Statistics in Transition – new series, Vol. 15, No. 1, 111–133.

KOSIOROWSKI, D., MIELCZAREK, D., RYDLEWSKI, J., SNARSKA, M., (2014b). Applications of the Functional Data Analysis for Extracting Meaningful Information from Families of Yield Curves and Income Distribution Densities, in Knowledge-Economy-Society Contemporary Tools of Organisational Resources Management, ed. P. Lula, Fundation of the CUE, 309–321.

RAMSAY, J. O., HOOKER, G., GRAVES, S., (2009). Functional Data Analysis with R and Matlab, Springer-Verlag, New-York.

SHANG, H. L., (2013). ftsa: An R Package for Analyzing Functional Time Series, The R Journal, Vol. 5/1, 65–72.

ZUO, Y., SERFLING, R., (2000). General notions of statistical depth function. The Annals of Statistics, 28: 461–482.

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 627–636

DURATION-BASED APPROACH TO VAR INDEPENDENCE BACKTESTING

Marta Małecka[1]

ABSTRACT

Dynamic development in the area of value-at-risk (VaR) estimation and growing implementation of VaR-based risk valuation models in investment companies stimulate the need for statistical methods of VaR models evaluation. Following recent changes in Basel Accords, current UE banking supervisory regulations require internal VaR model backtesting, which provides another strong incentive for research on relevant statistical tests. Previous studies have shown that commonly used VaR independence Markov-chain-based testing procedure exhibits low power, which constitutes a particularly serious problem in the case of finite-sample settings. In the paper, as an alternative to the popular Markov test an overview of the group of duration-based VaR backtesting procedures is presented along with exploration of their statistical properties while rejecting a non-realistic assumption of infinite sample size. The Monte Carlo test technique was adopted to provide exact tests, in which asymptotic distributions were replaced with simulated finite sample distributions. A Monte Carlo study, based on the GARCH model, was designed to investigate the size and the power of the tests. Through the comparative analysis we found that, in the light of observed statistical properties, the duration-based approach was superior to the Markov test.

Key words: VaR backtesting, Markov test, Haas test, TUFF test, Weibull test, gamma test, EACD test.
JEL Classification: C22, C52, D53, G11;
AMS Classification: 62M10, 91B84, 62P05.

1. Introduction

In the context of business practice, value-at-risk (VaR) measure is by far the most popular approach to market risk valuation. Its increasing range of applications constantly boosts scientific discussion on various aspects of VaR. There is a parallel discussion in literature on VaR estimation methods and statistical evaluation of VaR models. Commonly used, Markov-chain-based test

[1] University of Lodz, Department of Statistical Methods. E-mail: marta.malecka@uni.lodz.pl.

(Christoffersen, 1998), aimed at evaluating independence in VaR forecasts has been shown to exhibit unsatisfactory power (Lopez, 1999). For practical significance of the independence property, there has been a constant development in statistical testing procedures aimed at detecting various forms of dependence in VaR violation series. As an alternative to testing the number of exceptions and working on Markov property assumption, it was proposed to adopt a duration approach, which is based on transformation of the failure process into the duration series.

The duration-based approach was primarily motivated by the concept of the time-until-first-failure test, in which the reverse of no-hit period is treated as an estimate of the success probability in the Bernoulli model (Kupiec, 1995). Both this test and its generalization in the form of the time-between-failures test (Haas, 2001) were based on the Bernoulli process assumption. Another line of research explored the properties of the memory-free exponential distribution and included the regression-based exponential autoregressive conditional duration test (EACD test, Engle and Russel, 1998). Further approach utilizing the memory-free property was based on testing the assumption of the exponential distribution against the alternative of a wider class of probability distributions (Christoffersen and Pelletier, 2004).

The aim of this paper was to provide a revision of independence VaR tests based on durations between VaR exceptions and to present a comparative analysis of their statistical properties. We compared duration-based approach to the broadly used Markov independence test. Asymptotic probability distributions of the considered tests are known, however when the number of VaR violations is small, which is common in practice, there may be substantial differences between them and their finite sample analogues. Therefore statistical properties of all tests were evaluated with the use of Monte Carlo tests technique, which allowed us to obtain the null distribution of tests statistics in finite sample setting. Such a technique has a great advantage of providing exact tests based on any statistics whose finite sample distribution is intractable but can be simulated (Dufour, 2006). Power properties of the tests were assessed in the simulation study in which GARCH-process assumption was adopted to stay in line with widely recognized facts about financial time series observed in daily intervals.

Section 2 of this paper introduces the methodological framework for duration-based testing. Section 3, dedicated to the simulation study, outlines the Monte Carlo tests procedure, provides details of the Monte Carlo study and contains simulation results. The final section summarizes and concludes the article.

2. Duration-based VaR tests

VaR evaluation framework is based on the stochastic process of VaR failures:

$$I_{t+1} = \begin{cases} 1, & r_{t+1} < VaR_t(p) \\ 0, & r_{t+1} \geq VaR_t(p) \end{cases}, \tag{1}$$

where $p-$ tolerance level, r_t- value of the rate of return at time t, $VaR_t(p)-$ value of the VaR forecast from moment t. Independence tests, based on the VaR failure process, use various forms of the alternative hypothesis. The alternative of the two-state Markov chain was proposed to test for serial correlation (Christoffersen, 1998). The null hypothesis in Christoffersen's Markov test, formulated in terms of conditional probabilities of a single-step transition in the $\{I_t\}$ process, $H_0 : \pi_{01} = \pi_{11}$, is verified by the statistic

$$LR_{ind} = -2\log \frac{\hat{\pi}_1^{t_1}(1-\hat{\pi}_1)^{t_0}}{\hat{\pi}_{01}^{t_{01}}(1-\hat{\pi}_{01})^{t_{00}}\hat{\pi}_{11}^{t_{11}}(1-\hat{\pi}_{11})^{t_{10}}} \sim_{as} \chi^2_{(1)} \tag{2}$$

where $\hat{\pi}_1 = \frac{t_1}{t_0 + t_1}$, t_0- number of non-exceptions, t_1- number of exceptions, $\pi_{ij}-$ probability of transition from the state i to the state j, $\hat{\pi}_{01} = \frac{t_{01}}{t_0}$, $\hat{\pi}_{11} = \frac{t_{11}}{t_1}$, $t_{ij}-$ number of transitions form the state i to the state j. State 0 in the above notation is interpreted as non-exception, while 1 represents VaR failure. Under the null, the probability of an exception at time t does not depend on the state of the process at time $t-1$, which means that null hypothesis is equivalent to the iid Bernoulli series.

By contrast to testing the parameter restriction in the assumed Markov chain, duration-based tests use a transformation of the underlying $\{I_t\}$ process into a duration series $\{V_i\}$ defined as:

$$V_i = t_i - t_{i-1}, \tag{3}$$

where t_i denotes the day of the violation number i. The *TUFF* test (time-until-first-failure test), based on the assumption that the $\{I_t\}$ series is drawn from the Bernoulli process, investigates the time of no-hit sequence until the first VaR violation. The reverse of this time constitutes the estimate of the probability of success in the assumed Bernoulli model. The *TUFF* test generalization to the time-between-failures test (Hass, 2001), which requires all durations between violations, examines time-changeability of the Bernoulli process parameter.

The Haas test statistic, being a natural generalization of the *TUFF* statistic, takes the following form:

$$LR_{ind,H} = -2\ln\left[\frac{\alpha(1-\alpha)^{V_1-1}}{p_1(1-p_1)^{V_1-1}}\right] + \sum_{i=2}^{N} -2\ln\left[\frac{\alpha(1-\alpha)^{V_i-1}}{p_i(1-p_i)^{V_i-1}}\right], \tag{4}$$

where $p_i = \frac{1}{V_i}$, V_1 – time until first failure, V_i – time between $(i-1)^{\text{th}}$ and i^{th} violation.

An alternative approach to duration testing is to utilize the exponential distribution as the only memory-free random distribution. The null hypothesis of the exponential distribution may be tested against the alternative distribution that allows dependence in the duration series. Similarly to the Markov and Haas test, the proposed exponential distribution tests are based on the LR framework (Domański et al., 2014), where the null model is nested in the alternative hypothesis. Therefore, the alternative family of distributions, in each variant of the test, involves the exponential distribution as a special case.

The alternative distributions that nest the null hypothesis of the exponential distribution, proposed in the literature, involve Weibull and gamma distributions. In the case of the Weibull distribution, the pdf takes the form:

$$f_W(v_i) = a^b b v_i^{b-1} e^{-(av_i)^b} \tag{5}$$

and includes the exponential distribution as a special case for $b=1$. Therefore, the null hypothesis takes the form H0:b=1 and the Weibull test requires fitting the unrestricted Weibull model and its restricted version for b=1.

Similarly for $b=1$ the exponential distribution is nested in the gamma distribution, given by the pdf:

$$f_\Gamma(v_i) = \frac{a^b v_i^{b-1} e^{-av_i}}{\Gamma(b)}. \tag{6}$$

As above, in an unrestricted case, it is necessary to maximize the gamma log likelihood function with respect to parameters a and b (Christoffersen and Pelletier, 2004).

The above tests, based on a distribution of durations between VaR violations, do not take any account of the ordering of VaR failures, which is considered in the exponential autoregressive conditional duration (EACD) procedure (Engle and Russel, 1998). The EACD test verifies the independence of VaR failures utilizing the regression of the durations on their past values:

$$E_{i-1}(V_i) = a + bV_{i-1}. \tag{7}$$

The exponential distribution assumption is also adopted, which gives the conditional pdf function of the duration V_i of the form:

$$f_{EACD}(v_i) = \frac{1}{a + bv_{i-1}} e^{-\frac{v_i}{a+bv_{i-1}}}, \tag{8}$$

which, for $b = 0$, nests the null model with the exponential distribution.

The above tests require computation of the log likelihood function for the unrestricted and restricted case, which, if we take account of possible presence of censored durations at the beginning and at the end of the series, takes the following form:

$$\begin{aligned} \ln L(V, \Theta) &= C_1 \ln S(V_1) + (1 - C_1) \ln f(V_1) + \sum_{i=2}^{N-1} \ln f(V_i) + \\ &+ C_N \ln S(V_N) + (1 - C_N) \ln f(V_N) \end{aligned} \tag{9}$$

where C_i takes the value of 1 if the duration V_i is censored and 0 otherwise, S is the survival function of the variable V_i and N is the number of VaR failures (Christoffersen and Pelletier, 2004).

3. Size and power properties

With regard to practical implementation of the considered tests, which normally involves finite sample setting, we used a Monte Carlo (MC) tests technique. Such a technique provides exact tests based on any statistic whose finite sample distribution can be simulated (Dufour, 2006). Following MC tests procedure, we generated $M = 9999$ realizations of the test statistic S_i from the null model and replaced the theoretical distribution of the test statistic F by its sample analogue based on $S_1, ..., S_M$. To generate the $\{I_t\}$ series under the null, we used the Bernoulli distribution with the probability of success p, equal to the assumed level of VaR failure tolerance. Having calculated the survival function:

$$\hat{G}_M(x) = \frac{1}{M} \sum_{i=1}^{M} 1(S_i \geq x) \tag{10}$$

we computed the empirical quantiles of the test statistic distribution. For the test statistic S_0, the corresponding Monte Carlo p-value was obtained according to the formula:

$$\hat{p}_M(S_0) = \frac{M\hat{G}_M(S_0) + 1}{M + 1}. \tag{11}$$

The simulated distributions showed that all tests tend to be oversized in finite samples and they do not converge to the nominal test size of 5% with lengthening time series [Tab. 1]. The differences between the empirical and theoretical distribution quantiles were confirmed by the graphical comparison of the simulated and theoretical densities (Fig. 1-5). The Haas and EACD test statistics exhibited the largest discrepancies in the shape of the simulated and theoretical probability density function, which indicated that practical application of these tests shouldn't be based on the asymptotic distributions. The empirical distribution of the Haas test was moved to the right off the theoretical shape, hence theoretical quantiles tended to be too small, translating into increased rejection rates. In the case of the EACD test the empirical distribution lied to the left of the theoretical curve, which gave undersized rejection rates.

Table 1. Empirical size of the duration-based tests compared to Markov test

	Series length					
	250	**500**	**750**	**1000**	**1250**	**1500**
Markov test	0.078	0.084	0.117	0.130	0.125	0.111
Haas test	0.112	0.135	0.183	0.195	0.223	0.199
Weibull test	0.076	0.073	0.084	0.088	0.108	0.114
Gamma test	0.071	0.080	0.090	0.109	0.128	0.157
EACD test	0.008	0.007	0.009	0.011	0.012	0.015

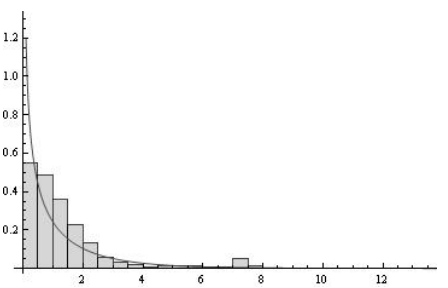

Figure 1. Simulated and theoretical pdf of the Markov test statistics for sample size 250

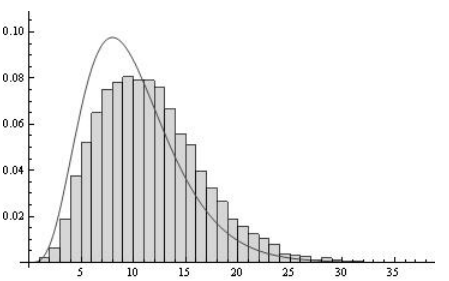

Figure 2. Simulated and theoretical pdf of the Haas test statistics for sample size 250

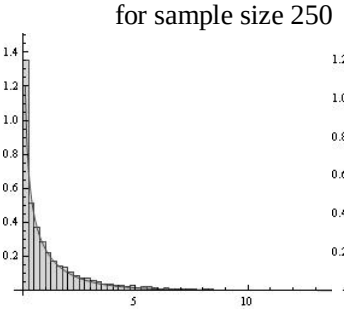

Figure 3. Simulated and theoretical pdf of the Weibull test statistics for sample size 250

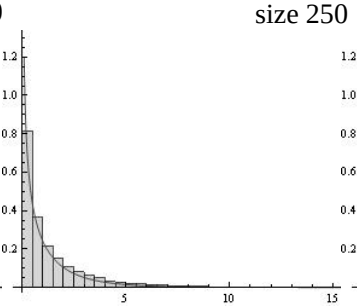

Figure 4. Simulated and theoretical pdf of the gamma test statistics for sample size 250

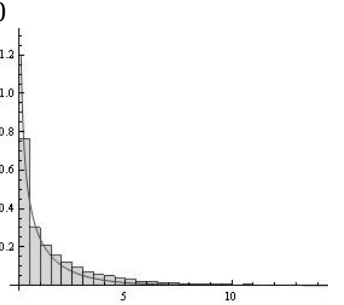

Figure 5. Simulated and theoretical pdf of the EACD test statistics for sample size 250

For the power comparison, we utilized the Monte Carlo simulation technique. The alternative model was obtained by generating return process from the GARCH-normal model with variance equation of the form $h_t = \omega + \alpha\varepsilon_{t-1}^2 + \beta h_{t-1}$ and computing VaR estimates from the incorrect homoscedastic model. The strength of the correlation in VaR failure series was assessed by the correlation coefficient of the squared returns ρ, whose value was set to 0.1, 0.3 and 0.5 in subsequent variants of the simulation experiment. The parameter values in the return data generating process: $\omega = 0.000001$, $\beta = 0.85$ were chosen so as to stay in line with real financial process parameter estimates for daily data on stock markets (Małecka, 2011). The value of α parameter ensured the required level of ρ (Fiszeder, 2009). VaR forecasts were set to the level of the 0.05 quantile of the unconditional distribution of the return process, which guaranteed the appropriate overall failure rate.

Having obtained the VaR violation series and the resulting duration series, we could compute the test statistics and use the Monte Carlo tests technique to evaluate corresponding p-values. Rejection rates under alternative hypothesis were calculated over 10000 Monte Carlo trials. The study was repeated for sample sizes 250, 500,..., 1500.

In the simulation study, we rejected cases for which the test was not feasible, which constituted a nontrivial sample selection rule. This was particularly frequent for small samples when no VaR failures or a very small number of VaR failures occur. Therefore we reported effective power rates, which correspond to multiplying raw power by the sample selection frequency.

Table 2. Empirical effective power of the duration-based tests compared to Markov test

Test	ρ	Series length					
		250	**500**	**750**	**1000**	**1250**	**1500**
Markov test	0.1	0,082	0,157	0,175	0,250	0,266	0,296
	0.3	0,199	0,423	0,579	0,683	0,749	0,824
	0.5	0,203	0,492	0,611	0,720	0,798	0,857
Haas test	0.1	0,353	0,490	0,622	0,717	0,750	0,843
	0.3	0,390	0,594	0,747	0,834	0,904	0,945
	0.5	0,473	0,662	0,790	0,889	0,924	0,959
Weibull test	0.1	0,064	0,134	0,199	0,318	0,318	0,436
	0.3	0,303	0,679	0,851	0,932	0,968	0,974
	0.5	0,381	0,731	0,878	0,938	0,971	0,985
Gamma test	0.1	0,026	0,051	0,079	0,136	0,144	0,197
	0.3	0,104	0,499	0,745	0,884	0,945	0,967
	0.5	0,120	0,546	0,815	0,915	0,961	0,972
EACD test	0.1	0,135	0,227	0,253	0,279	0,302	0,318
	0.3	0,177	0,358	0,512	0,567	0,621	0,667
	0.5	0,239	0,408	0,503	0,590	0,625	0,676

The comparative analysis of the empirical power (Tab. 2) indicated superiority of the duration-based approach to the Markov test. Finite sample rejection rates showed that for all sample sizes the Haas test exhibited the highest power. It was superior to other tests both for the shortest examined series of 250 observations, which is the minimal series length required for the VaR backtesting by the banking supervision in EU countries, and for the longest series. In all experimental variants, the observed power of the Haas test exceeded 30%. In the case of longest series the empirical power was over 90%. However, in the light of large discrepancy between the empirical and theoretical null distribution, the Haas test application should be limited to the analysis carried out with the use of the Monte Carlo test technique, which guarantees the exact test size.

Comparison of the two procedures based on testing the memory-free property against the Weibull or gamma alternative showed that for small sample sizes the Weibull test outperformed the gamma test. Apart from Haas test, the Weibull approach was another procedure superior to the Markov test. For smallest examined sample size, the observed power of this test was over 30% in two out of three experiment variants. For the largest samples the power estimates reached the levels of over 90%.

From all the considered procedures, including Markov test, the EACD test exhibited the lowest empirical power. This test was also outperformed by all other tests in terms of the test size.

4. Summary and conclusions

The paper explored the family of tests based on durations between subsequent VaR failures and provided insight into statistical properties of duration-based tests in comparison to commonly used Christoffersen's Markov test of 1998. Within the duration-based framework we presented the 1995 Kupiec concept of the time-until-first-failure test and its generalization by Haas – the time-between-failures test of 2001, which are based on the Bernoulli process assumption. Further line of enquiry was the regression-based approach by Engle and Russel of 1998, which utilized the concept of testing the properties of the exponential distribution. Finally we investigated procedures, proposed by Christoffersen and Pelletier in 2004, based on the assumption of the memory-free exponential distribution tested against the alternative involving a wider class of probability distributions. Statistical properties of all tests were evaluated with the use of the Monte Carlo tests technique, which allowed us to obtain the null distribution of tests statistics in finite sample setting. Power properties of the tests were assessed in the simulation study based on the GARCH-normal assumption.

The comparative analysis indicated superiority of the duration-based approach to the Markov test. Finite sample rejection rates were the highest for the Haas test. On the other hand, the Haas test statistic exhibited the largest discrepancy in the shape of the empirical and theoretical probability density function, which

indicated that asymptotic critical values for small samples can be misleading. Rejection rates for Weibull tests were higher than for the gamma procedure, also based on checking the memory-free property, and this test was the second duration-based procedure superior to the Markov test. The EACD test was outperformed by all other procedures in terms of both test size and power.

Acknowledgments

The research was supported by the Polish National Science Centre grant DEC-2013/11/N/HS4/03354.

REFERENCES

BERKOWITZ, J., CHRISTOFFERSEN, P., PELLETIER, D., (2011). Evaluating Value-at-Risk Models with Desk-Level Data. Management Science 12(57), 2213–2227.

CHRISTOFFERSEN, P., (1998). Evaluating Interval Forecasts. International Economic Review 39, 841–862.

CHRISTOFFERSEN, P., PELLETIER, D., (2004). Backtesting Value-at-Risk: A Duration-Based Approach. Journal of Financial Econometrics 1(2), 84–108.

DOMANSKI, CZ., PEKASIEWICZ, D., BASZCZYNSKA, A., WITASZCZYK, A., (2014). Testy statystyczne w procesie podejmowania decyzji. Wyd. UŁ, Łódź.

DUFOUR, J. M., (2006). Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics. Journal of Econometrics 133(2), 443–477.

ENGLE, R. F., RUSSEL, J. R., (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. Econometrica 66(5), 1127–62.

FISZEDER, P., (2009). Modele klasy GARCH w empirycznych badaniach finansowych. Wydawnictwo naukowe uniwersytetu Mikołaja Kopernika, Toruń.

HAAS, M., (2001). New methods in backtesting. Mimeo. Financial Engineering Research Center Caesar, Friedensplatz, Bonn.

KUPIEC, P., (1995). Techniques for Verifying the Accuracy of Risk Measurement Models. Journal of Derivatives 2, 174–184.

LOPEZ, J., (1999). Methods for Evaluating Value-at-Risk Estimates. FRBSF Economic Review 2, 3–17.

MAŁECKA, M., (2011). Prognozowanie zmienności indeksów giełdowych przy wykorzystaniu modelu klasy GARCH. Ekonomista 6, 843–860.

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 637—639

REPORT

The XXXIII International Conference on Multivariate Statistical Analysis, 17–19 November 2014, Łódź, Poland

The 33rd edition of the International Conference on Multivariate Statistical Analysis was held in Lodz, Poland, on November 17−19, 2014. The MSA 2014 conference was organized by the Department of Statistical Methods of the University of Lodz, the Institute of Statistics and Demography of the University of Lodz, the Polish Statistical Association and the Committee on Statistics and Econometrics of Polish Academy of Sciences. The Organizing Committee included: Professor Czesław Domański (Chairman) and Marta Małecka and Elżbieta Zalewska, (scientific secretaries) from the Department of Statistical Methods of the University of Lodz.

The Mayor of the City of Lodz, Hanna Zdanowska took the honorary patronage of the Multivariate Statistical Analysis MSA 2014 conference. Its organization was financially supported by the National Bank of Poland, the Polish Academy of Sciences, the Lodz City Council and StatSoft Polska Sp. z o.o.

The 2014 edition, as all previous Multivariate Statistical Analysis conferences, aimed at creating the opportunity for scientists and practitioners of statistics to present and discuss the latest theoretical achievements in the field of the multivariate statistical analysis, its practical aspects and applications. A number of presented and discussed statistical issues were based on questions identified during previous MSA conferences. The scientific programme covered various statistical problems, including multivariate estimation methods, statistical tests, non-parametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods in finance and economy, especially methods used in capital market and risk management. The topic range also included design of experiments and survey sampling methodology, mainly for the social science purposes. The conference was attended by 87 participants from the main academic centres in Poland (Bialystok, Katowice, Krakow, Olsztyn, Opole, Poznan, Rzeszow, Szczecin, Torun, Warszawa, Wroclaw) and from abroad (Finland, Ukraine, Lithuania). The list of participants included scientists, academic tutors as well as representatives of the National Bank of Poland, local statistical offices and business. In 17 sessions 67 papers were presented, including 3 invited lectures.

The conference was opened by Professor Czesław Domański. The subsequent speakers on the conference opening were Professor Antoni Różalski, a representative of the Rector of the University of Lodz, Professor Włodzimierz Nykiel and Professor Paweł Starosta, the Dean of the Faculty of Economics and Sociology of the University of Lodz.

After the opening ceremony all participants had the opportunity to attend the invited lecture by Professor Józef Pociecha, Professor Barbara Pawełek, Mateusz Baryła and Sabina Augustyn (Cracow University of Economics) *Crucial Problems Of Corporate Bankruptcy Modelling And Prediction.* The second invited lecture was presented by Professor Tadeusz Trzaskalik (University of Economics in Katowice) at the opening of the second day of the conference and was dedicated to *Modelling And Synthesis Of Preferences In Discrete Multicriteria Decision Problems.* At the conference closing the participants attended the invited lecture by Tomasz Górecki, Professor Mirosław Krzyśko and Waldemar Wołyński (Adam Mickiewicz University) entitled *Multivariate Functional Regression Analysis With Application To Classification Problems.*

Among regular conference sessions, the historical plenary session was held, chaired by Professor Mirosław Krzyśko, and dedicated to eminent Polish scientists. Professor Czesław Domański (University of Lodz) recalled the work and profile of outstanding Polish statisticians: Władysław Grabski and Oskar Lange. Professors Tadeusz Bednarski (University of Wroclaw), Józef Dziechciarz (Wroclaw University of Economics) and Paweł Starosta (University of Lodz) shared their memories of Jerzy Spława-Neyman, Zdzisław Hellwig and Władysław Welfe, respectively. Professor Iwona Markowicz (Szczecin University) dedicated her presentation to Mirosława Gazińska. Grzegorz Wyłupek (University of Wroclaw) outlined the life and scientific achievements of Józef Łukaszewicz and Piotr Nowak (University of Wroclaw) presented the profile of and Jarosław Bartoszewicz.

Other sessions were chaired respectively by:

SESSION II	Professor Tadeusz Bednarski (University of Wroclaw)
SESSION III	A Professor Grażyna Trzpiot (University of Economics in Katowice)
SESSION III	B Professor Andrzej Dudek (Wroclaw University of Economics)
SESSION IV A	Professor Daniel Kosiorowski (Cracow University of Economics)
SESSION IV B	Professor Józef Pociecha (Cracow University of Economics)
SESSION V A	Professor Janusz Wywiał (University of Economics in Katowice)
SESSION VI	A Professor Jerzy Korzeniewski (University of Lodz)
SESSION VI	B Professor Tadeusz Trzaskalik (University of Economics in Katowice)

SESSION VII A	Professor Andrzej Sokołowski (Cracow University of Economics)
SESSION VII B	Professor Marek Walesiak (Wroclaw University of Economics)
SESSION VIII A	Professor Grzegorz Kończak (University of Economics in Katowice)
SESSION VIII B	Professor Józef Dziechciarz (Wroclaw University of Economics)
SESSION IX A	Professor Grażyna Dehnel (Poznan University of Economics)
SESSION IX B	Professor Małgorzata Markowska (Wroclaw University of Economics)
SESSION X	Professor Bronisław Ceranka (Poznan University of Life Sciences)

The MSA 2014 conference was closed by the Chairman of the Organizing Committee, Professor Czesław Domański, who summarized the Conference as very effective and added that all discussions and doubts should become inspirations and strong motivations to further work for both scientists and practitioners. Finally, he thanked all the guests, conference partners and sponsors.

The next edition of Multivariate Statistical Analysis Conference MSA 2015 is planned on November $16^{th} - 18^{th}$, 2015, and will be held in Lodz, Poland. The Chairman of the Organizing Committee, Professor Czesław Domański informed all that this will be the 34^{th} edition of the conference and kindly invited all interested scientists, researchers and students to take part in it.

Prepared by:

Marta Małecka
Department of Statistical Methods, University of Lodz

Elżbieta Zalewska
Department of Statistical Methods, University of Lodz

STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 641–644

ABOUT THE AUTHORS

Agunloye Oluokun Kasali is a Lecturer at the Department of Mathematics, Obafemi Awolowo University, Ile-Ife, Nigeria, where he teaches Statistics. His main research area is Time Series Analysis.

Ahmad Zahoor obtained a Master of Science degree in Statistics from University of Agricultural, Faisalabad. In 2008 he received his PhD degree in Statistics (in the area of Survey Sampling) under the supervision of Prof. Muhammad Hanif from NCBA&E Lahore Pakistan and Postdoc Fellowship with Prof. Danny Pfeffermann from University of Southampton, UK, in 2012. He has been involved in teaching and research in different universities since 2004. Currently he is doing another PhD from University of Southampton, UK, under the supervision of Prof. Danny Pfeffermann and Prof. Li Chun Zhang. His areas of research are estimation in two-phase sampling, complex survey modelling, and modelling progressive data. Since 2008 he has published 33 articles and two books.

Arnab Raghunath is a Professor of Statistics at the Department of Statistics, University of Botswana, Gaborone, Botswana. He is a Sampling theory Expert.

Białek Jacek is an Assistant Professor in the Department of Statistical Methods in the Institute of Statistics and Demography at the University of Lodz. He graduated with a Master of Science degree in Applied Mathematics from the Technical University of Lodz (Department of Physics, Informatics and Applied Mathematics, 2003). In 2007 he defended his PhD thesis in economics at the Faculty of Economy and Sociology of the University of Lodz. He is interested in the development of statistical methods and, in particular, the price index theory.

Kosiorowski Daniel is an Associate Professor at the Department of Statistics in Cracow University of Economics, Cracow, Poland. His main research areas involve theory and applications of robust and nonparametric statistics and in particular theory and economic applications of the statistical depth functions.

Małecka Marta has graduated from the University of Lodz, Poland, where she obtained Bachelor's and Master's degrees in three study programmes run at two faculties: International Relations, Informatics and Econometrics at the Faculty of Economics and Sociology and Mathematics at the Faculty of Mathematics and Informatics. In 2014 she obtained a PhD degree in Economics at the University of Lodz. Her main fields of interest include probability theory, statistical tests, extreme value theory, market risk management and risk measures. She has published 20 research papers in both national and international journals, including JCR list and Web of Science indexed papers. Since 2014 she has been running the National Science Centre project „Hypothesis Testing in Market Risk Evaluation".

Rezankova Hana is a Professor and a vice-head of the Department of Statistics and Probability at the University of Economics, Prague, Czech Republic. She received her Master's and Ph.D. degrees from the University of Economics, Prague. From 2008, she is a full Professor of statistics. Her main interests are multivariate statistical methods, mainly cluster analysis and categorical data analysis. She is the president of the Czech Statistical Society and a member of the Czech Statistical Council of the Czech Statistical Office.

Sahid Ummara received her MPhil degree in Statistics from Quaid-i-Azam University Islamabad. She has been teaching in University of Gujrat as a lecturer since 2007. Her area of research is estimation in successive sampling. She has published 5 articles so far.

Shangodoyin Dahud Kehinde is a Professor of Statistics at the Department of Statistics, University of Botswana, Gaborone, Botswana. He is a Time Series Expert and the Vice President of Botswana Statistics Association.

Sulc Zdenek is a Ph.D. student at the Department of Statistics and Probability at the University of Economics, Prague, Czech Republic. He received his Master's degree in Statistics and Insurance from the same university in 2012. His interests are categorical data analysis and classification methods.

Turczak Anna works at The West Pomeranian Business School in Szczecin. She obtained a Master's degree in economics in 1999. In 2004 she defended her PhD thesis in economics at the Faculty of Economics and Management at the

University of Szczecin. She is particularly interested in issues relating to rational management of company's resources and the use of quantitative methods in economics. She investigates the usefulness of mathematical, statistical, econometrical and operations research methods in commerce finance, capital market, logistics and especially production management. She is a lecturer of International Business Studies. She has a very rich experience in business practice. She has been – among other positions – Marketing Department Manager, Sales Department Manager, and then Planning and Development Plenipotentiary of the President of the Board at Meat Plant AGRYF Inc. ANIMEX Group.

Wywiał Janusz L. is a Professor of statistics at the University of Economics in Katowice. His primary research interests include survey sampling methods, testing statistical hypotheses, applications of statistical methods in financial auditing.

Zalewska Marta is an Assistant Professor at the Department of Prevention of Environmental Hazards and Allergology of the Medical University of Warsaw. Her area of expertise is practical statistics in various problems of scientific research, in particular applications of statistics to medicine, sport and public health. She works together with medical doctors, biologists, epidemiologists performing statistical data analyses. She has authored about 20 research papers.

Zieliński Wojciech is a Full Professor in Department of Econometrics and Statistics of the University of Life Sciences in Warsaw. His main research area is classical inference and its applications, especially problems of interval estimation and testing statistical hypotheses as well as robust inference. He has published more than 80 articles and textbooks.

Zubair Rahma obtained her Master of Science degree in statistics from University of Gujrat in 2012. Her thesis title was "A General Class of Estimators Using Mixture of Auxiliary Variables in Two-Phase Sampling". She is now teaching in a public college at Sialkot, Pakistan.

Zwiech Patrycja works at the Faculty of Economics and Management at the University of Szczecin. She holds PhD in economics since 2006. She is particularly interested in issues relating to labour markets, social inequalities and

discrimination, the cohesion policy, human resources management. She is the author or co-author of more than 100 publications. She was on scientific placement at the London Metropolitan University, Working Lives Research Institute and at the Universite Jean Moulin Lyon 3.