Volume 21, Number 1, March 2020 **ISSN 1234-7655**

STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association and Statistics Poland

IN THIS ISSUE:

EDITORIAL OFFICE **ISSN 1234-7655**

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

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

CONTENTS

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1, pp. III–VII

From the Editor

With this issue, the first in the year 2020, we enter the new stage of journal's modernization and its advancement in terms of graphic appearance, including the cover and the main text font. We hope that readers and authors will appreciate these efforts to improve the appearance of the journal, and will enjoy the effects of the changes, which are supposed to work for the increase in its visibility and recognition. The list of almost three dozens of indexation bases, to which the *Statistics in Transition new series* is currently being included, seems to prove the standpoint that high quality papers deserve concern for the form of their presentation.

Altogether, the ten articles included in this issue cover a wide range of theoretical (statistical) and also the application (econometric) types of problems.

It starts with **Rohini Yaday's** and **Rajesh Tailor's** paper ***Estimation of finite population mean using two auxiliary variables under stratified random sampling.*** The problem of an alternative approach to estimate the population mean of the study variable with the help of the auxiliary variable under stratified random sampling is discussed along with specification of the properties of the suggested estimator under large sample approximation. It has been shown that the suggested estimator is more efficient than other considered estimators. To judge the merits of the suggested estimator, an empirical study has been carried out to the support of the present study. For instance, it reveals the conditions under which the suggested estimator has less MSE than the usual combined ratio estimator.

In the paper ***Parametric prediction of finite population total under informative sampling and nonignorable nonresponse,*** **Abdulhakeem Eideh** combines two methodologies used in the model-based survey sampling: the prediction of finite population total (the so-called T), under informative sampling, and full response, and the prediction of T, when the sampling design is noninformative and the nonresponse mechanism is nonignorable. The main problem is how to account for the joint effects of informative sampling designs and of not missing at random response mechanism in statistical models for complex survey data. To this aim the response distribution and relationships between moments of the superpopoulation sample, sample-complement, response and non-response distributions, for the prediction of finite population totals were used. The derived parametric predictors of T, the use the observation for the response set of the study variable or variable of interest, values of auxiliary variables and their population totals, sampling weights, and propensity scores. An interesting

outcome of the present study is that most predictors known from model-based survey sampling, can be derived as a special case from this general theory (see Chambers and Clark, 2012).

In the next paper, ***Horvitz-Thompson estimator based on the auxiliary variable,*** **J. Al-Jararha** and **Mazen Sulaiman** discuss the Horvitz and Thompson (1952) estimator, which is modified and uses the availability of the auxiliary variable. The modified estimators are extended to be employed in stratified sampling designs, along with empirical studies conducted for the comparison purposes. Based on Sugden and Smith (2002) approach, the two exactly unbiased estimators based on their families for estimating perform better than the original estimators, even if the original estimators are asymptotically unbiased or unbiased estimators. Furthermore, the estimators deduced from Horvitz and Thompson (1952) perform better than the deduced estimators from the ratio estimator.

Agnieszka Rossa's and **Agnieszka Palma's** article entitled ***Predicting parity progression ratios for young women by the end of their childbearing life*** starts with justification for choosing the parity progression ratios (PPRs) approach – as playing an important role in fertility – which is used by the authors to analyse future fertility trends. First, they assess age-specific parity progression ratios for women in childbearing ages, and then compare these ratios with ratios at the end of women's reproductive life, and the latter with the completed PPRs. The aim of this study is to adopt a modified Brass method to calculate the projected parity progression ratios using the age-period fertility data sourced from the Human Fertility Database (HFD). The observed and predicted age-specific PPRs are used to examine parity progressions in Poland as a case study.

The paper ***Using the ICAPM to estimate the capital cost of stock portfolios: empirical evidence on the Warsaw stock exchange*** by **Stanisław Urbański, Jacek Leśkow** presents the method for estimating the cost of capital of typical portfolios available on the Warsaw Stock Exchange. The authors introduce the three factor Fama-French model and its two modifications. They also apply the bootstrap method to evaluate the variability of their estimation method. The cost of capital they refer to is related to portfolios of real options linked to projects. The market returns are generated both by stock companies running such projects and by real options modifying selected projects. The estimated cost of capital can serve as a useful indicator for investors and for managers overseeing portfolios of stocks. Also, such an indicator can serve as a general reference while making business decisions. The study demonstrated that the estimated cost of capital assumes highest values for value portfolios and stock companies with high financial indicators and, at the same time, low market prices compared to their book value. By the same token, the estimated cost of capital assumes

low values for growth portfolios and for stock companies characterised by low financial indicators and, at the same time, high market prices compared to their book values.

Dariusz Kotlewski's and **Mirosław Błażej's** article ***Development of KLEMS accounting implemented in Poland*** discusses the main body of the KLEMS growth accounting system recently implemented in Poland. The works on the KLEMS productivity accounting in Poland started in 2013 and focused on areas such as the development of methodology and the availability and assessment of data. These efforts enabled preparing KLEMS data sets pertaining to the Polish economy and proved that unavailable data can be effectively estimated. Additionally, interesting but complex and debatable results were obtained, such as labour hoarding together with remunerations' freezing around the 2009 crisis, accompanied by a natural drop in the capital contribution growth and an increase in the MFP contribution, which most probably indicated effective reorganizations in the economy. In the years 2012–2014, increasing labour and capital contributions did not fully translate into gross value added growth, which led to negative MFP growths, as these are calculated residually. This, however, changed completely in the last two years of the time span covered by the research, namely in 2015–2016. An industry-level analysis became also possible, showing that the Polish economy was developing dynamically and undergoing intensive modernisation, which was obtained, however, with a debatable contribution of the State. To study the debatable features of the Polish economy in a greater detail, a further decomposition of the labour factor growth into four sub-factor contributions instead of two sub-factor contributions was performed. This additional analysis confirmed that labour hoarding phenomenon specific for Poland contributed to a softer impact of the 2007-09 financial crisis on this country's economy.

In the paper, C***hange point detection in CO_2 emission-energy consumption nexus using recursive Bayesian algorithm*****, Adepoju Abosede Adedayo, Awe Olushina Olawale,** focus on the synthesis of conditional dependence structure of recursive Bayesian estimation of dynamic state space models with time-varying parameters using a newly modified recursive Bayesian algorithm. The results of empirical applications to climate data from Nigeria reveal that the relationship between energy consumption and carbon dioxide emission in Nigeria reached the lowest peak in the late 1980s and the highest peak in early 2000. For South Africa, the slope trajectory of the model descended to the lowest in the mid-1990s and attained the highest peak in early 2000. These change-points can be attributed to the economic growth, regime changes, anthropogenic activities, vehicular emissions, population growth and industrial revolution in these countries. These results have implications on climate change prediction and global warming in both countries, and also show that recursive Bayesian dynamic model with time-varying parameters is suitable for statistical inference in climate change and policy analysis. The need for low carbon technologies, which are

capable of plummeting carbon emissions and enhancing sustainable economic growth in South Africa and Nigeria, is hereby recommended and emphasized. This may include policies to enhance efficiency of energy through modification from non-renewable energy to renewable energy, thereby reducing the impending greenhouse effect.

The next paper, by **Grażyna Dehnel** and **Łukasz Wawrowski**, entitled ***Robust estimation of wages in small enterprises: the application to Poland's districts*** is based on the authors' presentation at the **XXVII Conference on „Classification and Data Analysis – Theory and Applications" (Ciechocinek, 10–12 September 2018).** The aim of an empirical study designed to test a small area estimation method was to apply a robust version of the Fay-Herriot model to the estimation of average wages in the small business sector. Unlike the classical Fay-Herriot model, its robust version makes it possible to meet the assumption of normality of random effects under the presence of outliers. Moreover, the use of this version of the Fay-Herriot model helps to improve the precision of estimates, especially in domains where samples are of small sizes. These alternative models are supplied with auxiliary variables. The study seeks to present the characteristics of and differences among small business units cross-classified by selected NACE sections and district units of the provinces of Mazowieckie and Wielkopolskie. It was carried out on the basis of data from a survey conducted by the Statistical Office in Poznan and from administrative registers. It is the first study which attempts to produce estimates of average wages for this sector of the national economy.

The next two papers are included into the research communicates section, which is devoted to new or alternative approaches. The first one, by **Tolga Zaman,** ***New family of exponential estimators for the finite population mean*** proposes a new class of exponential-type estimators in simple random sampling for the estimation of the population mean of the study variable using information of the population proportion possessing certain attributes. Theoretically, mean squared error (MSE) equations of the suggested ratio exponential estimators are obtained and compared with the Naik and Gupta (1996) ratio and product estimators, the ratio and product exponential estimator presented in Singh et al. (2007) and the ratio exponential estimators presented in Zaman and Kadilar (2019a). As a result of these comparisons, it is observed that the proposed estimators always produce more efficient results than the others. In addition, these theoretical results are supported by the application of original data sets.

In the **Piyush Kant Rai's** and **Anu Sirohi's** article ***Alternative approach to moments of order statistics from one-parameter Weibull distribution*****,** the Weibull distribution is used to describe various observed failures of phenomena and widely used in survival analysis and reliability theory. Sometimes it is very difficult to compute moments of such distributions due to various reasons, e.g. analytical issues, multi

parameter cases, etc. This study presents the computation of the moments and the expected value of the product of order statistics in the sample from the one-parameter Weibull distribution. An alternative approach in connection to the survival function is used to obtain these moments and expected values. In addition the characteristic function of the above distribution is also obtained in the form of gamma functions. Further, an illustration is shown to find the first two moments and the expected value of the product of order statistics by using this approach.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1, pp. IX

Submission information for Authors

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1, pp. XI–XII

Editorial Policy

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

***

ABSTRACTING AND INDEXING DATABASES

Statistics in Transition new series is currently covered in:

Databases indexing the journal:

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

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 1–12, DOI 10.21307/stattrans-2020-001
Submitted – 10.01.2018; accepted for publishing – 17.01.2020

Estimation of finite population mean using two auxiliary variables under stratified random sampling

Rohini Yadav[1], **Rajesh Tailor**[2]

ABSTRACT

This paper addresses the problem of an alternative approach to estimating the population mean of the study variable with the help of the auxiliary variable under stratified random sampling. The properties of the suggested estimator have been studied under large sample approximation. It has been demonstrated that the suggested estimator is more efficient than other considered estimators. To judge the merits of the proposed estimator, an empirical study has been carried out to support the present study.

Key words: Study variable, auxiliary variable, stratified random sampling, dual to ratio estimator, bias and mean squared error.

1. Introduction

It is a well-known fact that the supplementary or auxiliary information always increases the precision of the estimators for the population parameters of the study variable. Ratio, product, regression and ratio-cum-product type of estimators are good examples in this context. Cochran (1940) proposed the ratio estimator assuming that the study variable (y) and auxiliary variable (x) are positively correlated, and the population mean of the auxiliary variable is known. However, when the study variable (y) and the auxiliary variable (x) are negatively correlated then the ratio estimator does not perform well. In that situation, the product estimator envisaged by Robson (1957) is appropriate.

Many authors including Murthy (1964), Sisodia and Dwivedi (1981), Upadhyaya and Singh (1999), Singh and Tailor (2003), Singh et al. (2004), Upadhyaya et al. (2011), etc., proposed different ratio type estimators for the population mean $\overline{Y}$ in simple random sampling. Singh and Tailor (2005), Tailor and Sharma (2009), Upadhyaya et al. (2011) and Yadav et al. (2012) proposed different ratio-cum-product estimators of a

[1] Corresponding author. Department of Statistics, Amity Institute of Applied Sciences, Amity University, Noida-201313 (U.P.). E-mail: rohiniyadav.ism@gmail.com.
[2] School of Studies in Statistics, Vikram University, Ujjain-456010 (M.P.). E-mail: tailorraj@gmail.com

finite population mean of the study variable using the values of known parameters of the auxiliary variables in simple random sampling. Srivenkataramana (1980) first proposed dual to ratio estimator, Bandopadhyaya (1980) suggested dual to product estimator for the population mean using transformation on auxiliary variable under simple random sampling.

As we know, the stratified random sampling can provide greater precision than a simple random sampling of the same size and it often requires a smaller sample, which saves money. Due to these shortcomings under simple random sampling, many authors like Sisodia and Dwivedi (1981), Upadhyaya and Singh (1999), Kadilar and Cingi (2003, 2005), Singh et al. (2004), Singh and Vishwakarma (2008, 2010), Koyuncu and Kadilar (2009), Tailor (2009), Tailor el al. (2012), Yadav et al. (2014), Gupta and Shabbir (2015), Tailor et al. (2015) and Mishra et al. (2017) defined ratio estimators and ratio-cum-product estimators under stratified random sampling, which perform better than usual ratio and product estimators in simple random sampling under certain limitations. Motivated by them, an attempt is made to develop an efficient dual to ratio-cum-product estimator of the population mean of the study variable using the knowledge of coefficient of kurtosis of the auxiliary variable under stratified random sampling.

Let the population of size N be equally divided into L strata with N_h elements in the h^{th} stratum such that $N=\sum_{h=1}^{L} N_h$. Let y be the study variable and x and z be two auxiliary variables assuming values y_{hi}, x_{hi} *and* z_{hi} for the i^{th} unit in h^{th} stratum. Let n_h be the size of the sample drawn from h^{th} stratum of size N_h by using simple random sampling without replacement (SRSWOR) such that sample size $n=\sum_{h=1}^{L} n_h$. We define

$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$: h^{th} stratum mean for the study variate y

$\bar{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{hi}$: h^{th} stratum mean for the study variate x

$\bar{Z}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} z_{hi}$: h^{th} stratum mean for the study variate z

$\bar{Y} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} y_{hi} = \frac{1}{N}\sum_{h=1}^{L} N_h\bar{Y}_h = \sum_{h=1}^{L} W_h\bar{Y}_h$: population mean of the study variate y

$\bar{X} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} x_{hi} = \frac{1}{N}\sum_{h=1}^{L} W_h \bar{X}_h$: population mean of the auxiliary variate x

$\bar{Z} = \frac{1}{N}\sum_{h=1}^{L}\sum_{i=1}^{N_h} z_{hi} = \frac{1}{N}\sum_{h=1}^{L} W_h \bar{Z}_h$: population mean of the study variate z

$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$: sample mean of the study variate y for h^{th} stratum

$\bar{x}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} x_{hi}$: sample mean of the auxiliary variate x for h^{th} stratum

$\bar{z}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} z_{hi}$: sample mean of the auxiliary variate z for h^{th} stratum

$W_h = \frac{N_h}{N}$: stratum weight of h^{th} stratum

Hansen et al. (1946) defined the classical combined ratio estimator for the population mean of the study variable y under stratified random sampling as

$$\bar{y}_{RC} = \bar{y}_{st}\left(\frac{\bar{X}}{\bar{x}_{st}}\right) \tag{1.1}$$

Here, it is assumed that the study variable y and auxiliary variable x are positively correlated.

Using the information on two auxiliary variables x and z, Tailor et al. (2012) proposed a ratio-cum-product estimator of the population mean $\bar{Y}$ in stratified random sampling as

$$\hat{\bar{Y}}_{RP}^{ST} = \bar{y}_{st}\left(\frac{\bar{X}}{\bar{x}_{st}}\right)\left(\frac{\bar{z}_{st}}{\bar{Z}}\right) = \sum_{h=1}^{L} W_h \bar{Y}_h \left(\frac{\sum_{h=1}^{L} W_h \bar{X}_h}{\sum_{h=1}^{L} W_h \bar{x}_h}\right)\left(\frac{\sum_{h=1}^{L} W_h \bar{z}_h}{\sum_{h=1}^{L} W_h \bar{Z}_h}\right) \tag{1.2}$$

Tailor et al. (2015) utilized the information of the coefficient of kurtosis of the auxiliary variables x and z, and proposed a ratio-cum-product estimator $\hat{\bar{Y}}_{RP1}^{ST}$ of the population mean $\bar{Y}$ under stratified random sampling as

$$\hat{\bar{Y}}_{RP1}^{ST} = \bar{y}_{st}\left[\frac{\sum_{h=1}^{L} W_h\left\{\bar{X}_h + \beta_{2h}(x)\right\}}{\sum_{h=1}^{L} W_h\left\{\bar{x}_h + \beta_{2h}(x)\right\}}\right]\left[\frac{\sum_{h=1}^{L} W_h\left\{\bar{z}_h + \beta_{2h}(z)\right\}}{\sum_{h=1}^{L} W_h\left\{\bar{Z}_h + \beta_{2h}(z)\right\}}\right] \tag{1.3}$$

where $\beta_{2h}(x)$ *and* $\beta_{2h}(z)$ are the coefficients of kurtosis of the auxiliary variates x and z, respectively in h^{th} stratum.

The mean squared errors (MSE) of the combined ratio estimator $\bar{y}_{RC}$, Tailor et al. (2012) estimator $\hat{\bar{Y}}_{RP}^{ST}$ and Tailor et al. (2015) estimator $\hat{\bar{Y}}_{RP1}^{ST}$, defined in (1.1), (1.2) and (1.3) up to the first order of approximation, are respectively given by

$$MSE\left(\bar{y}_{RC}\right) = \sum_{h=1}^{L} W_h^2 \gamma_h \left(S_{yh}^2 + R_1^2 S_{xh}^2 - 2R_1 S_{yxh}\right) \tag{1.4}$$

$$MSE\left(\hat{\bar{Y}}_{RP}^{ST}\right) = \sum_{h=1}^{L} W_h^2 \gamma_h \left(S_{yh}^2 + R_1^2 S_{xh}^2 + R_2^2 S_{zh}^2 - 2R_1 S_{yxh} + 2R_2 S_{yzh} - 2R_1 R_2 S_{xzh}\right) \tag{1.5}$$

$$MSE\left(\hat{\bar{Y}}_{RP1}^{ST}\right) = \sum_{h=1}^{L} W_h^2 \gamma_h \left[S_{yh}^2 + R_{12}^2 S_{xh}^2 + R_{13}^2 S_{zh}^2 - 2R_{12} S_{yxh} + 2R_{13} S_{yzh} - 2R_{12} R_{13} S_{xzh}\right] \tag{1.6}$$

where $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$, $\bar{x}_{st} = \sum_{h=1}^{L} W_h \bar{x}_h$, $\bar{z}_{st} = \sum_{h=1}^{L} W_h \bar{z}_h$,

$S_{yh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(y_{hi} - \bar{Y}_h\right)^2$, $S_{xh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(x_{hi} - \bar{X}_h\right)^2$,

$S_{zh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(z_{hi} - \bar{Z}_h\right)^2$, $S_{yxh} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(y_{hi} - \bar{Y}_h\right)\left(x_{hi} - \bar{X}_h\right)$,

$S_{yzh} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(y_{hi} - \bar{Y}_h\right)\left(z_{hi} - \bar{Z}_h\right)$, $S_{xzh} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(x_{hi} - \bar{X}_h\right)\left(z_{hi} - \bar{Z}_h\right)$,

$\gamma_h = \left(\frac{1}{n_h} - \frac{1}{N_h}\right)$, $R_1 = \frac{\bar{Y}}{\bar{X}}$, $R_2 = \frac{\bar{Y}}{\bar{Z}}$,

$R_{12} = \frac{\bar{Y}}{\sum_{h=1}^{L} W_h\left\{\bar{X}_h + \beta_{2h}(x)\right\}} = \frac{\bar{Y}}{\bar{X}_{1h}}$ *and* $R_{13} = \frac{\bar{Y}}{\sum_{h=1}^{L} W_h\left\{\bar{Z}_h + \beta_{2h}(z)\right\}} = \frac{\bar{Y}}{\bar{Z}_{1h}}$.

2. The suggested estimator

Motivated by Srivenkataramana (1980) and assuming that the parameters of the auxiliary variables x and z are known, we propose the dual to ratio-cum-product estimator t_{st}^{*} of Tailor et al. (2015) estimator $\hat{\bar{Y}}_{RP1}^{ST}$ of the population mean $\bar{Y}$ of the study variable y, which is defined as

$$t_{st}^{*}=\bar{y}_{st}\left[\frac{\sum_{h=1}^{L}W_h\left\{\bar{x}_h^{*}+\beta_{2h}(x)\right\}}{\sum_{h=1}^{L}W_h\left\{\bar{X}_h+\beta_{2h}(x)\right\}}\right]\left[\frac{\sum_{h=1}^{L}W_h\left\{\bar{Z}_h+\beta_{2h}(z)\right\}}{\sum_{h=1}^{L}W_h\left\{\bar{z}_h^{*}+\beta_{2h}(z)\right\}}\right] \tag{2.1}$$

where $\bar{x}_h^{*}=\left(\frac{N_h\bar{X}_h-n_h\bar{x}_h}{N_h-n_h}\right)$ and $\bar{z}_h^{*}=\left(\frac{N_h\bar{Z}_h-n_h\bar{z}_h}{N_h-n_h}\right)$.

Using the transformation on $\bar{x}_h^{*}$ *and* $\bar{z}_h^{*}$ of the auxiliary variables x and z, the suggested estimator t_{st}^{*} in (2.1) can be written as

$$t_{st}^{*}=\left(\sum_{h=1}^{L}W_h\bar{y}_h\right)\left[\frac{\sum_{h=1}^{L}W_h\left\{\left(\frac{N_h\bar{X}_h-n_h\bar{x}_h}{N_h-n_h}\right)+\beta_{2h}(x)\right\}}{\sum_{h=1}^{L}W_h\left\{\bar{X}_h+\beta_{2h}(x)\right\}}\right]\left[\frac{\sum_{h=1}^{L}W_h\left\{\bar{Z}_h+\beta_{2h}(z)\right\}}{\sum_{h=1}^{L}W_h\left\{\left(\frac{N_h\bar{Z}_h-n_h\bar{z}_h}{N_h-n_h}\right)+\beta_{2h}(z)\right\}}\right] \tag{2.2}$$

Let $\bar{y}_h=\bar{Y}(1+e_{0h}),\quad \bar{x}_h=\bar{X}_h(1+e_{1h})\quad$ *and* $\quad\bar{z}_h=\bar{Z}_h(1+e_{2h})$

such that $E(e_{0h})=E(e_{1h})=E(e_{2h})=0$

$$E(e_{0h}^2)=\gamma_h C_{yh}^2,\qquad E(e_{1h}^2)=\gamma_h C_{xh}^2,\qquad E(e_{2h}^2)=\gamma_h C_{zh}^2,$$

$$E(e_{0h}e_{1h})=\gamma_h\rho_{yxh}C_{yh}C_{xh}=\gamma_h\frac{S_{yxh}}{\bar{Y}_h\bar{X}_h},\qquad E(e_{1h}e_{2h})=\gamma_h\rho_{xzh}C_{xh}C_{zh}=\gamma_h\frac{S_{xzh}}{\bar{X}_h\bar{Z}_h}\quad \textit{and}$$

$$E(e_{0h}e_{2h})=\gamma_h\rho_{yzh}C_{yh}C_{zh}=\gamma_h\frac{S_{yzh}}{\bar{Y}_h\bar{Z}_h}$$

Expressing (2.2) in terms of e's, we get

$$= \bar{Y}\left(1+e_0\right)\left(1-e_1\right)\left(1-e_2\right)^{-1}$$

where $e_0 = \dfrac{\sum_{h=1}^{L} W_h \bar{Y}_h e_{0h}}{\sum_{h=1}^{L} W_h \bar{Y}_h} = \dfrac{\sum_{h=1}^{L} W_h \bar{Y}_h e_{0h}}{\bar{Y}}$, $e_1 = \dfrac{\sum_{h=1}^{L} W_h g_h \bar{X}_h e_{1h}}{\sum_{h=1}^{L} W_h \left\{\bar{X}_h + \beta_{2h}(x)\right\}} = \dfrac{\sum_{h=1}^{L} W_h g_h \bar{X}_h e_{1h}}{\bar{X}_{1h}}$,

$e_2 = \dfrac{\sum_{h=1}^{L} W_h g_h \bar{Z}_h e_{2h}}{\sum_{h=1}^{L} W_h \left\{\bar{Z}_h + \beta_{2h}(z)\right\}} = \dfrac{\sum_{h=1}^{L} W_h g_h \bar{Z}_h e_{2h}}{\bar{Z}_{1h}}$ and $g_h = \dfrac{n_n}{N_h - n_n}$

such that

$$E\left(e_0\right) = E\left(e_1\right) = E\left(e_2\right) = 0$$

and

$$E\left(e_0^2\right) = \frac{1}{\bar{Y}^2}\sum_{h=1}^{L} W_h^2 \gamma_h S_{yh}^2, \qquad E\left(e_1^2\right) = \frac{1}{\bar{X}_{1h}^2}\sum_{h=1}^{L} W_h^2 \gamma_h g_h^2 S_{xh}^2,$$

$$E\left(e_2^2\right) = \frac{1}{\bar{Z}_{1h}^2}\sum_{h=1}^{L} W_h^2 \gamma_h g_h^2 S_{zh}^2, \qquad E\left(e_0 e_1\right) = \frac{1}{\bar{Y}\,\bar{X}_{1h}}\sum_{h=1}^{L} W_h^2 \gamma_h g_h S_{yxh},$$

$$E\left(e_1 e_2\right) = \frac{1}{\bar{X}_{1h}\bar{Z}_{1h}}\sum_{h=1}^{L} W_h^2 \gamma_h g_h^2 S_{xzh}, \qquad E\left(e_0 e_2\right) = \frac{1}{\bar{Y}\,\bar{Z}_{1h}}\sum_{h=1}^{L} W_h^2 \gamma_h g_h S_{yzh}$$

To the first degree of approximation, the bias and mean squared error of the suggested estimator t_{st}^* are given by

$$B\left(t_{st}^*\right) = \bar{Y}\sum_{h=1}^{L} W_h^2 \gamma_h g_h \left[\frac{g_h}{\bar{Z}_{1h}}\left(\frac{S_{zh}^2}{\bar{Z}_{1h}} - \frac{S_{xzh}}{\bar{X}_{1h}}\right) + \frac{1}{\bar{Y}}\left(\frac{S_{yzh}}{\bar{Z}_{1h}} - \frac{S_{xyh}}{\bar{X}_{1h}}\right)\right] \tag{2.3}$$

$$MSE\left(t_{st}^*\right) = \sum_{h=1}^{L} W_h^2 \gamma_h \left[S_{yh}^2 + g_h^2 R_{12}^2 S_{xh}^2 + g_h^2 R_{13}^2 S_{zh}^2 - 2g_h R_{12} S_{yxh} + 2g_h R_{13} S_{yzh} - 2g_h^2 R_{12} R_{13} S_{xzh}\right] \tag{2.4}$$

3. Efficiency comparison

Since we know that the variance of the usual unbiased estimator of the study variable y in stratified random sampling is defined as

$$V\left(\bar{y}_{st}\right)=\sum_{h=1}^{L}W_h^2\gamma_h S_{yh}^2 \tag{3.1}$$

From equations (1.4), (1.5), (1.6), (2.4) and (3.1), we have

(i) $MSE\left(t_{st}^{*}\right)<MSE\left(\bar{y}_{st}\right)$ if and only if:

$$\sum_{h=1}^{L}W_h^2\gamma_h g_h^2\left\{R_{12}^2S_{xh}^2+R_{13}^2S_{zh}^2-2R_{12}R_{13}S_{xzh}\right\}-2\sum_{h=1}^{L}W_h^2\gamma_h g_h\left\{R_{12}S_{yxh}-R_{13}S_{yzh}\right\}<0 \tag{3.2}$$

(ii) $MSE\left(t_{st}^{*}\right)<MSE\left(\bar{y}_{RC}\right)$ if and only if:

$$\begin{aligned}\sum_{h=1}^{L}W_h^2\gamma_h g_h^2\left\{R_{12}^2S_{xh}^2+R_{13}^2S_{zh}^2-2R_{12}R_{13}S_{xzh}\right\}-2\sum_{h=1}^{L}W_h^2\gamma_h g_h\left\{R_{12}S_{yxh}-R_{13}S_{yzh}\right\}\\ -\sum_{h=1}^{L}W_h^2\gamma_h R_1\left\{R_1S_{xh}^2-2S_{yxh}\right\}<0\end{aligned} \tag{3.3}$$

(iii) $MSE\left(t_{st}^{*}\right)<MSE\left(\hat{\bar{Y}}_{RP}^{ST}\right)$ if and only if:

$$\begin{aligned}\sum_{h=1}^{L}W_h^2\gamma_h g_h^2\left\{R_{12}^2S_{xh}^2+R_{13}^2S_{zh}^2-2R_{12}R_{13}S_{xzh}\right\}-2\sum_{h=1}^{L}W_h^2\gamma_h g_h\left\{R_{12}S_{yxh}-R_{13}S_{yzh}\right\}\\ -\sum_{h=1}^{L}W_h^2\gamma_h\left\{R_1^2S_{xh}^2+R_2^2S_{zh}^2-2R_1S_{yxh}+2R_2S_{yzh}-2R_1R_2S_{xzh}\right\}<0\end{aligned} \tag{3.4}$$

(iv) $MSE\left(t_{st}^{*}\right)<MSE\left(\hat{\bar{Y}}_{RP1}^{ST}\right)$ if and only if:

$$\begin{aligned}\sum_{h=1}^{L}W_h^2\gamma_h g_h^2\left\{R_{12}^2S_{xh}^2+R_{13}^2S_{zh}^2-2R_{12}R_{13}S_{xzh}\right\}-2\sum_{h=1}^{L}W_h^2\gamma_h g_h\left\{R_{12}S_{yxh}-R_{13}S_{yzh}\right\}\\ -\sum_{h=1}^{L}W_h^2\gamma_h\left\{R_{12}^2S_{xh}^2+R_{13}^2S_{zh}^2-2R_{12}S_{yxh}+2R_{13}S_{yzh}-2R_{12}R_{13}S_{xzh}\right\}<0\end{aligned} \tag{3.5}$$

From equations (3.2), (3.3), (3.4) and (3.5), we obtained the conditions under which the suggested estimator performed better than the usual unbiased estimator, combined ratio estimator $\overline{y}_{RC}$, Tailor et al. (2012) estimator $\hat{\overline{Y}}_{RP}^{ST}$ and Tailor et al. (2015) estimator $\hat{\overline{Y}}_{RP1}^{ST}$.

4. Empirical study

To judge the efficiency of the proposed estimator over the usual unbiased estimator, combined ratio estimator $\overline{y}_{RC}$, Tailor et al. (2012) estimator $\hat{\overline{Y}}_{RP}^{ST}$ and Tailor et al. (2015) estimator $\hat{\overline{Y}}_{RP1}^{ST}$, the following data set is taken. The description of the population is given below:

Population [Source: Murthy (1967), p. 228]

z: Number of workers

y: Output and

x: Fixed capital

	n_1 =2	n_2 =2	N_1 =5	N_2 =5
	$\overline{Z}_1$ =51.80	$\overline{Z}_2$ =60.60	$\overline{X}_1$ =214.4	$\overline{X}_2$ =333.8
N=10	$\overline{Y}_1$ =1925.8	$\overline{Y}_2$ = 3115.6	S_{z_1} =0.75	S_{z_2} =4.84
n=4	S_{x_1} = 74.87	S_{x_2} =66.35	S_{y_1} =615.92	S_{y_2} =340.38
	S_{zx_1} =-38.08	S_{zx_2} = - 287.92	S_{yz_1} = -411.16	S_{yz_2} =-1536.24
	S_{yx_1} =39360.68	S_{yx_2} =22356.50	C_{x_1} =0.35	C_{x_2} =0.20
	C_{z_1} =0.01	C_{z_2} =0.08	$\beta_{21}(x)$ =1.88	$\beta_{22}(x)$ =2.32
	$\beta_{21}(z)$ =1.84		$\beta_{22}(z)$ =1.49	

For the purpose of the efficiency comparison of the proposed estimator, we have computed the percent relative efficiencies (PREs) of the estimators with respect to the usual unbiased estimator $\overline{y}_{st}$ using the formula:

$$PRE(t, \overline{y}_{st}) = \frac{MSE(\overline{y}_{st})}{MSE(t)} \times 100; \qquad \text{where } t = \overline{y}_{st}, \overline{y}_{RC}, \hat{\overline{Y}}_{RP}^{ST}, \hat{\overline{Y}}_{RP1}^{ST} \text{ and } t_{st}^{*}$$

The findings are given in Table 1.

Table 1

Percent relative efficiencies of the estimators $\bar{y}_{st}$, $\bar{y}_{RC}$, $\hat{\bar{Y}}_{RP}^{ST}$, $\hat{\bar{Y}}_{RP1}^{ST}$ and t_{st}^{*} with respect to $\bar{y}_{st}$

Estimators	$\bar{y}_{st}$	$\bar{y}_{RC}$	$\hat{\bar{Y}}_{RP}^{ST}$	$\hat{\bar{Y}}_{RP1}^{ST}$	t_{st}^{*}
Population	100.00	239.8632589	141.9128961	146.8036738	361.4516525

5. SIMULATION STUDY

In the paper, we generated two populations for two auxiliary variables x and z. Population I has equal size stratum and Population II has unequal size stratum. We calculated the variance and MSE's values of $\bar{y}_{st}$, $\bar{y}_{RC}$, $\hat{\bar{Y}}_{RP}^{ST}$, $\hat{\bar{Y}}_{RP1}^{ST}$ and t_{st}^{*} respectively, for different values of the sample size viz. 500, 700, 900; obtained from different stratum using proportional allocation. The variance and MSE's of the estimators are represented in Table 2 and Table 3.

Population I: N **=2500** N_1**=500** N_2**=500** N_3**=500** N_4**=500** N_5**=500**

Table 2

Estimators / n	$\bar{y}_{st}$	$\bar{y}_{RC}$	$\hat{\bar{Y}}_{RP}^{ST}$	$\hat{\bar{Y}}_{RP1}^{ST}$	t_{st}^{*}
500	23.6416	0.004104	0.006670	0.001786	0.001283
700	23.6416	0.001785	0.003674	0.015174	0.000766
900	23.6416	0.001793	0.001898	0.005736	0.001114

Population II: N **=2500** N_1**=500** N_2**=300** N_3**=700** N_4**=600** N_5**=400**

Table 3

Estimators / n	$\bar{y}_{st}$	$\bar{y}_{RC}$	$\hat{\bar{Y}}_{RP}^{ST}$	$\hat{\bar{Y}}_{RP1}^{ST}$	t_{st}^{*}
500	22.9679	0.000955	0.002192	0.002793	0.000432
700	22.9679	0.001202	0.001678	0.011525	0.000514
900	22.9679	0.001018	0.001101	0.0052308	0.000476

From Table 2 and Table 3, we came up with a conclusion that MSE of the proposed estimator is less than all the other considered estimators. So, we can say that the

performance of our proposed estimator is better than the usual unbiased estimator, combined ratio estimator $\overline{y}_{RC}$, Tailor et al. (2012) estimator $\hat{\overline{Y}}_{RP}^{ST}$ and Tailor et al. (2015) estimator $\hat{\overline{Y}}_{RP1}^{ST}$.

5. Conclusion

This paper has suggested a dual to ratio-cum-product estimator to estimate the population mean of the study variable using the knowledge of the population mean as well as the coefficient of kurtosis of two auxiliary variables x and z under stratified random sampling. Its properties have been studied under large sample approximation. Section 3 reveals the conditions under which the suggested estimator has less MSE than the usual combined ratio estimator $\overline{y}_{RC}$, Tailor et al. (2012) estimator $\hat{\overline{Y}}_{RP}^{ST}$ and Tailor et al. (2015) estimator $\hat{\overline{Y}}_{RP1}^{ST}$. This means that the proposed estimator is more efficient than other considered estimators under certain limitations. Table 1 shows that the suggested dual to ratio-cum-product estimator has more percent relative efficiency as compared to the usual combined ratio estimator $\overline{y}_{RC}$, Tailor et al. (2012) estimator $\hat{\overline{Y}}_{RP}^{ST}$ and Tailor et al. (2015) estimator $\hat{\overline{Y}}_{RP1}^{ST}$. In addition, the simulation study has also been carried out to show the efficiency of the suggested estimator, whose results are displayed in Table 2 and Table 3. Therefore, it can be concluded that if information on the coefficient of kurtosis of the auxiliary variables is available for each stratum then the suggested estimator performs well and more efficiently than other considered estimators. Thus, the suggested estimator can be recommended as an alternative use of the estimation of the population mean of the character under study.

Acknowledgements

The authors are very thankful to both learned referees for their suggestions/comments to improve the quality of the paper. We are also very grateful to Ms. Vishwantra Sharma, University of Jammu for her help to conduct the simulation study.

REFERENCES

BANDYOPADHYAY, S., (1980). Improved ratio and product estimators, Sankhya: Indian J. Stat., 42, pp. 45–49.

GUPTA, S., SHABBIR J., (2015). Estimation of Finite Population Mean in Stratified Random Sampling with Two Auxiliary Variables under Double Sampling Design, Communications In Statistics – Theory And Methods 44(13), pp. 2798–2808.

KADILAR, C., CINGI, H., (2003). Ratio estimators in stratified random sampling, Biometrical Journal, 45, pp. 218–225.

KADILAR, C., CINGI, H., (2005). A new estimator in stratified random sampling, Commun. Statist. Theor. Meth. 34, pp. 597–602.

KOYUNCU, N., KADILAR, C., (2009). Ratio and product estimators in stratified random sampling, J. Statist. Plann. Infer. 139, pp. 2552–2558.

MISHRA, M., SINGH, B. P., SINGH, R., (2017). Estimation of population mean using two auxiliary variables in stratified random sampling, Journal of Reliability and Statistical Studies 10(1), pp. 59–68.

MURTHY, M. N., (1967). Sampling theory and methods, Calcutta, India: Statistical Publishing Society, p. 228.

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

SINGH, H. P., TAILOR, R., TAILOR, R., KAKRAN, M., (2004). An improved estimation of population mean using power transformation, Journal Indian Society Agricultural Statistics, 58, pp.223–230.

SINGH, H.P., SINGH. R., ESPEJO, M. R., PINEDA, M. D., (2005). On the efficiency of a dual to ratio-cum-product estimator in sample surveys, Math. Proceed. Royal Irish Academy, 105 A (2), pp. 51–56.

SINGH, H. P., VISHWAKARMA, G. K., (2008). A family of estimators of population mean using auxiliary information in stratified sampling, Commun. Statist. Theor. Meth. 37, pp. 1038–1050.

SINGH, H. P., VISHWAKARMA, G. K., (2010). A general procedure for estimating the population mean in stratified sampling using auxiliary information, Metron, LXVII (1), pp. 47–65.

SISODIA, B. V. S., DWIVEDI, V. K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable, Journal Indian Society Agricultural Statistics, 33, pp. 13–18.

SRIVENKATARAMANA, T., (1980). A dual of ratio estimator in sample surveys, Biometrika, 67, 1, pp. 199–204.

TAILOR, R., (2009). A modified ratio-cum-product estimator of finite population mean in stratified random sampling, Data Science Journal, 8, pp. 182–189 (on line).

TAILOR R., SHARMA, B. K., (2009). A modified ratio-cum-product estimator of finite population mean using known coefficient of variation and coefficient of kurtosis, Statistics in Transition, 10, pp. 15–24.

TAILOR, R., CHOUHAN, S., TAILOR, R., GARG, N., (2012). A ratio-cum-product estimator of population mean in stratified random sampling using two auxiliary variables, Statistica, LXXII, 3, pp. 287–297.

TAILOR, R., LAKHRE, A., TAILOR, R., GARG, N., (2015). An Improved Ratio-Cum-Product Estimator of Population Mean Using Coefficient of Kurtosis of auxiliary variates in Stratified Random Sampling, Journal of Reliability and Statistical Studies, 8, 2, pp. 59–67.

UPADHYAYA, L. N., SINGH, H. P., (1999). Use of transformed auxiliary variable in estimating the finite population mean, Biometrical Journal, 41, pp. 627–636.

UPADHYAYA, L. N., SINGH, H. P., CHATTERJEE, S., YADAV, R., (2011). A Generalized Family of Transformed Ratio-Product Estimators of Finite Population Mean in Sample Surveys. Model Assisted Statistics and Applications, 6, 2, pp. 137–150.

UPADHYAYA, L. N., SINGH, H. P., CHATTERJEE, S., YADAV, R., (2011). Improved Ratio and Product Exponential type Estimators for Finite Population Mean in Sample Surveys, Journal of Statistical Theory and Practice, 5, 2, pp. 285–302.

YADAV, R., UPADHYAYA, L. N., SINGH, H. P., CHATTERJEE, S., (2012). Almost Unbiased Ratio and Product Type Exponential Estimators, Statistics in Transition, 13, 3, pp. 537–550.

YADAV, R., UPADHYAYA, L. N., SINGH, H. P., CHATTERJEE, S., (2014). Improved Ratio and Product Exponential type Estimators for Finite Population Mean in Stratified Random Sampling, Communications in Statistics: Theory & Methods (Taylor & Francis), 43, 15, pp. 3269–3285.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 13–35, DOI 10.21307/stattrans-2020-002
Submitted – 09.07.2019; accepted for publishing – 04.02.2020

Parametric prediction of finite population total under Informative sampling and nonignorable nonresponse

Abdulhakeem Eideh[1]

ABSTRACT

In this paper, we combine two methodologies used in the model-based survey sampling, namely the prediction of the finite population total, named T, under informative sampling and full response, see Sverchkov and Pfeffermann (2004), and the prediction of T with a noninformative sampling design and the nonignorable nonresponse mechanism, see Eideh (2012). The former approach involves the dependence of the first order inclusion probabilities on the study variable, while the latter involves the dependence of the probability of nonresponse on unobserved or missing observations. The main aim of the paper is to consider how to account for the joint effects of informative sampling designs and not-missing-at-random response mechanism in statistical models for complex survey data. For this purpose, theoretically, we use the response distribution and relationships between the moments of the superpopoulation, the sample, sample-complement, response, and nonresponse distributions for the prediction of finite population totals, see Eideh (2016). The derived parametric predictors of T use the observation for the response set of the study variable or variable of interest, values of auxiliary variables and their population totals, sampling weights, and propensity scores. An interesting outcome of the T study is that most predictors known from model-based survey sampling can be derived as a special case from this general theory, see Chambers and Clark (2012).

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

1. Introduction

Data collected by sample surveys are used extensively to make inferences on assumed population models. Often, survey design features (clustering, stratification, unequal probability selection, etc.) are ignored and the sample data are then analysed using classical methods based on simple random sampling. This approach can, however, lead to erroneous inference because of sample selection bias implied by informative sampling - the sample selection probabilities depend on the values of the

[1] Department of Mathematics,Al-Quds University, Abu-Dees Campus, Al-Quds, Palestine.
E-mail: msabdul@staff.alquds.edu.

model outcome variable (or the model outcome variable is correlated with design variables not included in the model). See Pfeffermann et al. (1998) and Eideh and Nathan (2006). In addition to the effect of complex sample design, one of the major problems in the analysis of survey data is that of missing values or nonresponse. Little and Rubin (2002) consider three types of the nonresponse mechanism or the missing data mechanism:

(a) Missing completely at random (MCAR): if the response probability does not depend on the study variable, or the auxiliary population variable, the missing data are MCAR.
(b) Missing at random (MAR) given auxiliary population variable: if the response probability depends on the auxiliary population variable but not on the study variable, the missing data are MAR.
(c) Not missing at random (NMAR): if the response probability depends on the value of a missing study variable, the missing data are NMAR.

So, the cross-classification of the sampling design and the response mechanism is summarized in the following table:

Table 1.

Sampling Design	Response Mechanism		
	MCAR	MAR	NMAR
Informative – I	IMCAR	IMAR	INMAR
Noninformative – N	NMCAR	NMAR	NNMAR

For inference problem, Little (1982) classifies the nonresponse mechanism as ignorable (MAR and MCAR) and nonignorable (NMAR). In this sense, the cross classification of the sampling design and the nonresponse mechanism is:

Table 2.

Sampling Design	Nonresponse Mechanism	
	Ignorable – i	Nonignorable – n
Informative – i	ii	in
Noninformative – n	ni	nn

For more information about prediction modelling approach for potentially nonignorable nonresponse in complex surveys, see Little (1983, 2003).

Pfeffermann and Sikov (2011), and Eideh (2012) consider estimation of superpopulation parameters and prediction of finite population parameters (census parameters) under nonignorable nonresponse via response and nonresponse distributions when the sampling design in noninformative.

Eideh (2016) treated the estimation of finite population mean and superpopulation parameters when the sampling design is informative and the nonresponse mechanism is nonignorable. Sverchkov and Pfeffermann (2004) use of the sample and complement-sample distributions for the semiparametric prediction of finite population totals under single-stage sampling. None of the above studies consider simultaneously the problem of informative sampling and the problem of nonignorable nonresponse in the prediction of finite population total. In this paper, we study, within a modelling framework, the parametric prediction of finite population total, by specifying the probability distribution of the observed measurements under informative sampling and nonignorable nonresponse. This is the most general situation in surveys and other combinations of sampling informativeness and response mechanisms can be considered as special cases.

It should be pointed here that, according to Sarndal (2011), "Nonresponse causes both bias and increased variance. Its square is typically the dominant portion of the Mean Squared Error (MSE). We address primarily surveys on individuals and households with quite large sample sizes, as is typical for Journal of Official Statistics for government surveys; consequently, the variance contribution to MSE is low by comparison. Increased variance due to nonresponse is nevertheless an issue; striking a balance between variance increase and bias reduction is considered, for example, in Little and Vartivarian (2005)." Furthermore, Brick (2013) mentioned that "Model assumptions and adjustments are made in an attempt to compensate for missing data. Because the mechanisms that cause unit nonresponse are almost never adequately reflected in the model assumptions, survey estimates may be biased even after the model based adjustments. Nonresponse also causes a loss in the precision of survey estimates, primarily due to reduced sample size and secondarily as the result of increased variation of the survey weights. However, bias is the dominant component of the nonresponse-related error in the estimates, and nonresponse bias generally does not decrease as the sample size increases. Thus, bias is often the largest component of mean square error of the estimates even for subdomains when the sample size is large". Here, we focus on the bias, variance and MSE.

The paper is structured as follows. In Section 2 we review the definition of sample, sample-complement, response, and nonresponse distributions, and derive new relationships between their moments. In Section 3, we derive a parametric predictions and their biases of finite population total under informative sampling and not missing at random the nonresponse mechanism. Also, we apply the results for three models, namely: common mean population model, simple ratio population model, and simple regression population model. Finally, Section 4 provides the conclusions.

2. Response, nonresponse distributions and relationships between their moments

Let $U = \{1,..., N\}$ denote a finite population consisting of N units. Let y be the study variable of interest and let y_i be the value of y for the ith population unit. A probability sample s is drawn from U according to a specified sampling design. The sample size is denoted by n. Let $\mathbf{x}_i = x_i = (x_{i1},..., x_{ip})'$, $i \in U$ be the values of a vector of auxiliary variables, $x_1,...,x_p$, and $\mathbf{z} = z = \{z_1,...,z_N\}$ be the values of known design variables, used for the sample selection process not included in the model under consideration. In what follows, we consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s) > 0$, and sampling weight $w_i = 1/\pi_i$; $i = 1,..., N$. In practice, the π_i's may depend on the population values $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. We express this dependence by writing: $\pi_i = \Pr(i \in s \mid \mathbf{x}, \mathbf{y}, \mathbf{z})$ for all units $i \in U$. Denote by $\mathbf{I} = (I_1,..., I_N)'$ the N by 1 sample indicator (vector) variable, such that $I_i = 1$ if unit $i \in U$ is selected to the sample and $I_i = 0$ if otherwise, so that $s = \{i \mid i \in U, I_i = 1\}$ and its complement is $\bar{s} = c = \{i \mid i \in U, I_i = 0\}$. We consider the population values $y_1,\ldots,y_N$ as random variables, which are independent realizations from a distribution with probability density functions (pdf) $f_p(y_i \mid \mathbf{x}_i; \theta)$, indexed by a vector of parameters θ.

In addition to the effect of complex sample design, one of the major problems in the analysis of survey data is that of missing values. In recent articles by Eideh (2009), Pfeffermann and Sikov (2011), and Eideh (2012), the authors defined and studied the problem of nonignorable nonresponse using the response and nonresponse distributions where the sampling design is noninformative. Denote by $R = (R_1,..., R_N)'$ the N by 1 response indicator (vector) variable such that $R_i = 1$ if unit $i \in s$ is observed and $R_i = 0$ if otherwise. We assume that these random variables are independent of one another and of the sample selection mechanism (Oh and Scheuren 1983). The response set is defined accordingly as $r = \{i \in s \mid R_i = 1\}$ and the nonresponse set by $\bar{r} = \{i \in s \mid R_i = 0\}$. We assume probability sampling, so that $\pi_i = \Pr(i \in s) > 0$ for all units $i \in U$. Let the response probability $\psi_i = \Pr(i \in r \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) > 0$ for all units $i \in s$ and $\phi_i = 1/\psi_i$ be the response weight

for $i \in s$. Let $O = \{(\mathbf{x}_i, I_i), i \in U\}, \{\pi_i, R_i, i \in s\} \cup \{(y_i, \mathbf{x}_i), i \in r\}$ and $N, n,$ and m be the available information from the sample and response sets.

According to Eideh (2007, 2009), the (marginal) response pdf of y_i is given by:

$$f_r(y_i \mid \mathbf{x}_i, \theta, \eta, \gamma) = \frac{E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma) f_s(y_i \mid \mathbf{x}_i, \theta, \eta)}{E_s(\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma)} \tag{1}$$

where the sample pdf of y_i, see Pfeffermann *et al.* (1998), is:

$$\begin{aligned} f_s(y_i \mid \mathbf{x}_i, \theta, \gamma) &= \frac{\Pr(i \in s \mid \mathbf{x}_i, y_i, \gamma)}{\Pr(i \in s \mid \mathbf{x}_i, \theta, \gamma)} f_p(y_i \mid \mathbf{x}_i, \theta) \\ &= \frac{E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma) f_p(y_i \mid \mathbf{x}_i, \theta)}{E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma)} \end{aligned} \tag{2a}$$

According to Pfeffermann and Sverchkov (1999), we have

$$E_p(y_i \mid \mathbf{x}_i) = \frac{E_s(w_i y_i \mid \mathbf{x}_i)}{E_s(w_i \mid \mathbf{x}_i)} \tag{2b}$$

Combining (1) and (2) we get:

$$f_r(y_i \mid \mathbf{x}_i, \theta, \eta, \gamma) = \frac{E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma)}{E_s(\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma)} \frac{E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma)}{E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma)} f_p(y_i \mid \mathbf{x}_i, \theta)$$

Furthermore, Sverchkov and Pfeffermann (2004) define the sample-complement pdf of y_i as:

$$f_{\bar{s}}(y_i \mid \mathbf{x}_i, \theta, \gamma) = \frac{E_p(1 - \pi_i \mid \mathbf{x}_i, y_i, \gamma) f_p(y_i \mid \mathbf{x}_i, \theta)}{E_p(1 - \pi_i \mid \mathbf{x}_i, \theta, \gamma)} \tag{3a}$$

and

$$E_{\bar{s}}(y_i \mid \mathbf{x}_i) = \frac{E_s\{(w_i - 1) y_i \mid \mathbf{x}_i\}}{E_s\{(w_i - 1) \mid \mathbf{x}_i\}} \tag{3b}$$

According to Eideh (2007, 2009), the (marginal) nonresponse pdf of y_i is given by:

$$f_{\bar{r}}(y_i \mid \mathbf{x}_i, \theta, \eta, \gamma) = \frac{E_s(1-\psi_i \mid \mathbf{x}_i, y_i, \gamma) f_s(y_i \mid \mathbf{x}_i, \theta, \eta)}{E_s(1-\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma)}$$
$$= \frac{E_s(1-\psi_i \mid \mathbf{x}_i, y_i, \gamma)}{E_s(1-\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma)} \frac{E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma)}{E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma)} f_p(y_i \mid \mathbf{x}_i, \theta) \tag{4}$$

Furthermore, for vector of random variables $(y_i, \mathbf{x}_i)$, using Eideh (2007,2009, 2016), we have:

$$E_p(y_i \mid x_i) = \frac{E_r(\phi_i w_i y_i \mid x_i)}{E_r(\phi_i w_i \mid x_i)} \tag{5}$$

$$E_s(y_i \mid x_i) = \frac{E_r(\phi_i y_i \mid x_i)}{E_r(\phi_i \mid x_i)} \tag{6}$$

$$E_{\bar{s}}(y_i \mid x_i) = \frac{E_r\{\phi_i(w_i - 1) y_i \mid x_i\}}{E_r\{\phi_i(w_i - 1) \mid x_i\}} \tag{7}$$

$$E_{\bar{r}}(y_i \mid x_i) = \frac{E_r\{(\phi_i - 1) y_i \mid x_i\}}{E_r\{(\phi_i - 1) \mid x_i\}} \tag{8}$$

Remark 1. The important feature of the formulas (5-8) is that, given $\{x_i, y_i, \emptyset_i, w_i; i \in r\}$, we can identify $E_p(y_i \mid x_i)$, $E_s(y_i \mid x_i)$, $E_{\bar{s}}(y_i \mid x_i)$, $E_r(y_i \mid x_i)$ and $E_{\bar{r}}(y_i \mid x_i)$.

Remark 2. Note that

$$E_r(y_i \mid \mathbf{x}_i) = E_p\left\{ \frac{E_s(\psi_i \mid \mathbf{x}_i, y_i, \gamma)}{E_s(\psi_i \mid \mathbf{x}_i, \theta, \eta, \gamma)} \frac{E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma)}{E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma)} y_i \,\middle|\, \mathbf{x}_i \right\} \neq E_p(y_i \mid \mathbf{x}_i)$$

Thus, estimating $E_p(y_i \mid \mathbf{x}_i)$ is often the main target of inference, which shows that ignoring an informative sampling scheme or NMAR nonresponse and thus estimating implicitly $E_r(y_i \mid \mathbf{x}_i)$ can bias the inference.

Using the equations (1-8), we can prove the following:

$$f_r(y_i \mid x_i) = \frac{E_r(\phi_i w_i \mid x_i)}{E_r(\phi_i w_i \mid x_i, y_i)} f_p(y_i \mid x_i) \tag{9}$$

$$f_{\bar{r}}(y_i \mid x_i) = \frac{E_r\{(\phi_i - 1) \mid x_i, y_i\}}{E_r\{(\phi_i - 1) \mid x_i\}} f_r(y_i \mid x_i) \tag{10}$$

$$f_{\bar{s}}(y_i \mid x_i) = \frac{E_r\{(\phi_i (w_i - 1)) \mid x_i, y_i\}}{E_r\{(\phi_i (w_i - 1)) \mid x_i\}} f_r(y_i \mid x_i) \tag{11}$$

Furthermore, using (1) and (6), we have

$$f_s(y_i \mid \mathbf{x}_i) = \frac{E_r(\phi_i \mid \mathbf{x}_i, y_i) f_r(y_i \mid \mathbf{x}_i)}{E_r(\phi_i \mid \mathbf{x}_i)} \qquad (1^*)$$

Remark 2. Once we identify $f_r(y_i \mid x_i)$, we can completely determine the nonresponse distribution $f_{\bar{r}}(y_i \mid x_i)$ and the nonsampled distribution $f_{\bar{s}}(y_i \mid x_i)$. So, instead of specifying $f_p(y_i \mid x_i)$, we can specify $f_r(y_i \mid x_i)$ based on the study variable and the auxiliary variables for the response set.

Estimation of response probabilities ψ_i for all $i \in s$:

If the nonresponse mechanism is not missing at random, then the classical methods for estimating the response probabilities using auxiliary variables, available for respondents and nonrespondents, is logistic or profit models. If we use the logistic model, then

$$\psi_i = \Pr(R_i = 1 \mid i \in s, x_i) = \frac{\exp(\gamma_0 + \gamma_1 x_i)}{1 + \exp(\gamma_0 + \gamma_1 x_i)}$$

We can fit this model using maximum likelihood approach. Thus, the estimate of ψ_i is:

$$\hat{\psi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}$$

If the nonresponse mechanism is NMAR, then values of y_i for $i \in r$ are available, but for $i \notin r$ are not available, so we cannot fit the following model:

$$\psi_i = \Pr(R_i = 1 \mid i \in s, x_i, y_i) = \frac{\exp(\gamma_0 + \gamma_1 x_i + \gamma_2 y_i)}{1 + \exp\exp(\gamma_0 + \gamma_1 x_i + \gamma_2 y_i)} \tag{12}$$

directly using maximum likelihood method. A recent approach of estimation ψ_i under nonignorable nonresponse is discussed by Sverchkov (2008) and Reddles *et al.* (2016).

3. Parametric prediction of finite population parameter under informative sampling and NMAR nonresponse mechanism

Eideh (2012) uses response and nonresponse distribution to derive new predictors of the finite population total, under common mean or homogeneous population model, simple ratio population model, and simple regression population model. These new predictors take into account the nonignorable nonresponse. In this section we extend the prediction problem when, in addition, the sampling design is informative.

3.1 General theory

Assume single-stage population model. Let

$$T = \sum_{i=1}^{N} y_i = \sum_{i \in s} y_i + \sum_{i \in \bar{s}} y_i = \sum_{i \in r} y_i + \sum_{i \in \bar{r}} y_i + \sum_{i \in \bar{s}} y_i \tag{15}$$

be the finite population total that we want to predict using the data from the response set and possibly values of auxiliary variables. Notice that T can be decomposed into three components: the first component represents the total for observed units in the sample – response set, $\sum_{i \in r} y_i$, the second component represents the total for unobserved units in sample – nonresponse set, $\sum_{i \in \bar{r}} y_i$, and the third component represents the total for non-sample units, $\sum_{i \in \bar{s}} y_i$ - nonsample set.

Let $\hat{T} = \hat{T}(O)$ define the predictor of T based on the available information, from the sample and response set $O = \{(I_i), i \in U\}, \{\pi_i, R_i, i \in s\} \bigcup \{(y_i), i \in r\}$ and

$N, n,$ and m. The mean square error (MSE) of $\hat{T}$ given O with respect to the population pdf is defined by:

$$MSE_p\left(\hat{T}\right) = E_p\left\{\left(\hat{T}-T\right)^2 \mid O\right\} = \left\{\hat{T}-E_p\left(T \mid O\right)\right\}^2 + Var_p\left(T \mid O\right) \tag{16}$$

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

$$\begin{aligned} T^* &= E_p\left(T \mid O\right) = E_p\left\{\left(\sum_{i\in r} y_i + \sum_{i\in \bar{r}} y_i + \sum_{i\in \bar{s}} y_i\right) \middle| O\right\} \\ &= \sum_{i\in r} y_i + \sum_{i\in \bar{r}} E_{\bar{r}}\left(y_i \middle| O\right) + \sum_{i\in \bar{s}} E_{\bar{s}}\left(y_i \middle| O\right) \end{aligned} \tag{17}$$

We know the values $\left\{y_i's, i \in r\right\}$, so the sum $\sum_{i\in r} y_i$ is known, thus we need to predict the total for unobserved units in the sample – nonresponse set, $\sum_{i\in \bar{r}} y_i$, and the total for non-sample units, $\sum_{i\in \bar{s}} y_i$. That is, to predict T we need to predict values for $\left\{y_i's, i \in \bar{r}\right\}$ and $\left\{y_i's, i \in \bar{s}\right\}$. Thus, our aim is to identify an optimal predictor of $\sum_{i\in \bar{s}} y_i$ and $\sum_{i\in \bar{r}} y_i$ based on the given observed data O.

The predictor given in (17) represents the prediction of T for single-stage sampling when the sampling mechanism in informative and missing value mechanism is NMAR. The analysis that follows assumes known model parameters. In practice, the unknown model parameters are replaced under the frequentist approach by sample estimates, yielding the corresponding "empirical predictors." In the present case, maximum likelihood estimation of the model parameters must be based on the response distribution of the observed units in the sample – response set, see Eideh (2016).

Using (7) and (8), equation (17) can be written as:

$$T^* = \sum_{i\in r} y_i + \sum_{i\in \bar{r}} \frac{E_r\left\{\left(\phi_i - 1\right)y_i\right\}}{E_r\left\{\left(\phi_i - 1\right)\right\}} + \sum_{i\in \bar{s}} \frac{E_r\left\{\phi_i\left(w_i - 1\right)y_i\right\}}{E_r\left\{\phi_i\left(w_i - 1\right)\right\}} \tag{18}$$

Hence, T^* can be estimated based only on the data in the response set, $\{y_i, \phi_i, w_i : i \in r\}$. Using method of moments estimates technique, that is, replace the moment under the response distribution by the average over the response set, for example $\hat{E}_r(a_i) = m^{-1}\sum_{i\in r} a_i$, where m is size of the response set r.

From now on, we use the following notations for the predictor of $T = \sum_{i=1}^{N} y_i$:

T_{in}^* - Best linear unbiased predictor of T when the sampling design is informative and the nonresponse mechanism is NMAR (nonignorable).

T_{ii}^* - Best linear unbiased predictor of T when the sampling design is informative and the nonresponse mechanism is ignorable.

T_{nn}^* - Best linear unbiased predictor of T when the sampling design is noninformative and the nonresponse mechanism is nonignorable.

T_{ni}^* - Best linear unbiased predictor of T when the sampling design is noninformative and the nonresponse mechanism is ignorable.

According to (18) and using the method of moments estimator, we can show that the best linear unbiased predictor for T is:

$$\begin{aligned}\hat{T}_{in}^* &= \sum_{i\in r} y_i + (n-m)\frac{\sum_{i\in r}(\phi_i - 1)y_i}{\sum_{i\in r}(\phi_i - 1)} + (N-n)\frac{\sum_{i\in r}\phi_i(w_i - 1)y_i}{\sum_{i\in r}\phi_i(w_i - 1)} \\ &= \sum_{i\in r} w_i^{in} y_i\end{aligned} \tag{19}$$

where

$$w_i^{in} = 1 + (n-m)\frac{(\phi_i - 1)}{\sum_{i\in r}(\phi_i - 1)} + (N-n)\frac{\phi_i(w_i - 1)}{\sum_{i\in r}\phi_i(w_i - 1)} \tag{20}$$

Note that

(a) $\sum_{i\in r}(\phi_i - 1)y_i$ is the Horvitz-Thompson estimator of $\sum_{i\in \bar{r}} y_i$.

(b) $\sum_{i\in r}\phi_i(w_i - 1)y_i$ is the Horvitz-Thompson estimator of $\sum_{i\in \bar{s}} y_i$.

(c) $\dfrac{n-m}{\sum_{i\in r}(\phi_i-1)}$ is the "Hajek type correction" for controlling the variability of the response weights.

(d) $\dfrac{(N-n)}{\sum_{i\in r}\phi_i(w_i-1)}$ is the "Hajek type correction" for controlling the variability of the product of the response weights and the sampling weights.

It is easy to verify tha

a) Under noninformative sampling design and nonignorable nonresponse:

$$w_i^{nn} = 1+(n-m)\frac{(\phi_i-1)}{\sum_{i\in r}(\phi_i-1)}+(N-n)\frac{\phi_i}{\sum_{i\in r}\phi_i}$$

(b) Under noninformative sampling design and ignorable nonresponse:

$$w_i^{ni} = 1+\frac{(n-m)}{m}+\frac{(N-n)}{m}=\frac{N}{m}$$

(c) Under informative sampling design and ignorable nonresponse:

$$w_i^{ii} = 1+\frac{(n-m)}{m}+(N-n)\frac{w_i-1}{\sum_{i\in r}(w_i-1)}$$

According to (1-8), we can write (17) as:

$$\begin{aligned} T_{in}^{*} &= \sum_{i\in r} y_i + \sum_{i\in \bar{r}}\left\{E_s(y_i|O)-\frac{Cov_s\left[(\psi_i,y_i)|O\right]}{1-E_s(\psi_i|O)}\right\}+\sum_{i\in \bar{s}}E_p(y_i|O)-\frac{Cov_p\left[(\pi_i,y_i)|O\right]}{1-E_p(\pi_i|O)} \\ &= \sum_{i\in r} y_i + \sum_{i\in \bar{r}}E_s(y_i|O)+\sum_{i\in \bar{s}}E_p(y_i|O)-\left\{\sum_{i\in \bar{r}}\frac{Cov_s\left[(\psi_i,y_i)|O\right]}{1-E_s(\psi_i|O)}+\sum_{i\in \bar{s}}\frac{Cov_p\left[(\pi_i,y_i)|O\right]}{1-E_p(\pi_i|O)}\right\} \end{aligned} \tag{21}$$

Using (1-8), we can show that the prediction nonresponse bias of T_{in} is:

$$B\left(T_{in}^{*}\right)=E_p\left(T_{in}^{*}-T\right)=-\left\{\sum_{i\in\bar{r}}E_p\left[y_i-E_{\bar{r}}\left(y_i\right)\right]+\sum_{i\in\bar{s}}E_p\left[y_i-E_{\bar{s}}\left(y_i\right)\right]\right\}$$

$$=-\left\{\sum_{i\in\bar{r}}\left[E_p\left(y_i\right)-E_{\bar{r}}\left(y_i\right)\right]+\sum_{i\in\bar{s}}\left[E_p\left(y_i\right)-E_{\bar{s}}\left(y_i\right)\right]\right\}$$

$$=-\left\{\sum_{i\in\bar{r}}\left[\left(E_p\left(y_i\right)-E_s\left(y_i\right)\right)+\frac{Cov_s\left(\psi_i,y_i\right)}{E_s\left(1-\psi_i\right)}\right]+\sum_{i\in\bar{s}}\frac{Cov_p\left(\pi_i,y_i\right)}{1-E_p\left(\pi_i\right)}\right\} \tag{22}$$

Therefore, the predictor T_{in}^{*} in (22) is unbiased T if:

(a) $Cov_s\left(\psi_i,y_i\right)=0$, (or $Cov_r\left(\phi_i,y_i\right)=0$), that is, there is no correlation between the study variable and the response probabilities ψ_i, consequently, the nonresponse mechanism is ignorable, and

(b) $Cov_p\left(\pi_i,y_i\right)=0$, (or $E_r\left(\phi_i\right)E_r\left(\phi_i w_i y_i\right)=E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)$), that is, there is no correlation between the study variable and the first order inclusion probabilities π_i, so the sampling design is noninformative.

If (a) is satisfied then (b) becomes $Cov_r\left(\phi_i w_i,y_i\right)=0$. In other words, if the sampling design is noninformative and the response mechanism is ignorable, so that $E_p\left(y_i\right)=E_s\left(y_i\right)=E_{\bar{s}}\left(y_i\right)=E_{\bar{r}}\left(y_i\right)=E_r\left(y_i\right)$, then T_{in}^{*} in unbiased of T .

Note that the stronger the relationship between the study variable and the response probability, and the study variable and the first order inclusion probabilities, the larger the bias.

According to (10) and (11), we can show that (22) has the following form:

$$B\left(T_{in}^{*}\right)=-\left\{\begin{array}{l}\displaystyle\sum_{i\in\bar{r}}\left\{-\frac{Cov_r\left(\phi_i w_i,y_i\right)}{E_r\left(\phi_i w_i\right)E_r\left\{\left(\phi_i-1\right)\right\}}+\frac{E_r\left(\phi_i w_i y_i\right)E_r\left(\phi_i\right)-E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)}{E_r\left(\phi_i w_i\right)E_r\left\{\left(\phi_i-1\right)\right\}}\right\}\\ \displaystyle-\sum_{i\in\bar{s}}\frac{E_r\left(\phi_i\right)E_r\left(\phi_i w_i y_i\right)-E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)}{E_r\left(\phi_i w_i\right)\left[E_r\left(\phi_i w_i\right)-E_r\left(\phi_i\right)\right]}\end{array}\right\} \tag{23}$$

Hence, the bias $B\left(T_{in}^{*}\right)$ can be estimated based only on the data in the response set, $\left\{y_i, \phi_i, w_i : i \in r\right\}$, using method of moments estimates technique, that is, replace the moment under the response distribution by the average over the response set, for example $\hat{E}_r\left(a_i\right) = m^{-1}\sum_{i\in r} a_i$.

To test the informativeness of the sampling design, see Pfeffermann and Sverchkov (1999) and Eideh and Nathan (2006). Moreover, for testing the ignorability of the nonresponse mechanism, see Eideh (2012).

Particular cases:

Case 1: The sampling design is noninformative and the nonresponse process is nonignorable, so that:

$$E_s\left(y_i|O\right) = E_p\left(y_i|O\right) \text{ and } Cov_p\left[\left(\pi_i, y_i\right)|O\right] = 0$$

Therefore,

$$T_{nn}^{*} = \sum_{i\in r} y_i + \sum_{i\in \bar{r}} E_p\left(y_i|O\right) + \sum_{i\in \bar{s}} E_p\left(y_i|O\right) - \sum_{i\in \bar{r}} \frac{Cov_p\left[\left(\psi_i, y_i\right)|O\right]}{1 - E_p\left(\psi_i|O\right)} \tag{24}$$

$$B\left(T_{nn}^{*}\right) = -\sum_{i\in \bar{r}}\left\{-\frac{Cov_r\left(\phi_i w_i, y_i\right)}{E_r\left(\phi_i w_i\right)E_r\left\{\left(\phi_i - 1\right)\right\}} + \frac{E_r\left(\phi_i w_i y_i\right)E_r\left(\phi_i\right) - E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)}{E_r\left(\phi_i w_i\right)E_r\left\{\left(\phi_i - 1\right)\right\}}\right\} \tag{25}$$

Case 2: The sampling design is noninformative and the nonresponse process is ignorable, so that:

$$E_r\left(y_i|O\right) = E_s\left(y_i|O\right) = E_p\left(y_i|O\right),\ Cov_p\left[\left(\pi_i, y_i\right)|O\right] = 0 \text{ and } Cov_s\left[\left(\psi_i, y_i\right)|O\right] = 0$$

Therefore,

$$T_{ni}^{*} = \sum_{i\in r} y_i + \sum_{i\in \bar{r}} E_p\left(y_i|O\right) + \sum_{i\in \bar{s}} E_p\left(y_i|O\right) \tag{26}$$

$$B\left(T_{ni}^{*}\right) = -\left\{\sum_{i\in \bar{r}}\left[E_p\left(y_i\right) - E_{\bar{r}}\left(y_i\right)\right] + \sum_{i\in \bar{s}}\left[E_p\left(y_i\right) - E_{\bar{s}}\left(y_i\right)\right]\right\} = 0 \tag{27}$$

Case 3: The sampling design is informative and the nonresponse process is ignorable, so that:

$$E_r\left(y_i|O\right)=E_s\left(y_i|O\right) \text{ and } Cov_s\left[\left(\psi_i,y_i\right)|O\right]=0$$

Therefore,

$$T_{ii}^{*}=\sum_{i\in r}y_i+\sum_{i\in\bar{r}}E_s\left(y_i|O\right)+\sum_{i\in\bar{s}}E_p\left(y_i|O\right)-\sum_{i\in\bar{s}}\frac{Cov_p\left[\left(\pi_i,y_i\right)|O\right]}{1-E_p\left(\pi_i|O\right)} \tag{28}$$

$$B\left(T_{ii}^{*}\right)=-\left\{\begin{array}{l}\sum_{i\in\bar{r}}\dfrac{E_r\left(\phi_i\right)E_r\left(\phi_i w_i y_i\right)-E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)}{E_r\left(\phi_i\right)E_r\left(\phi_i w_i\right)}- \\ \sum_{i\in\bar{s}}\dfrac{E_r\left(\phi_i\right)E_r\left(\phi_i w_i y_i\right)-E_r\left(\phi_i y_i\right)E_r\left(\phi_i w_i\right)}{E_r\left(\phi_i w_i\right)\left[E_r\left(\phi_i w_i\right)-E_r\left(\phi_i\right)\right]}\end{array}\right\} \tag{29}$$

3.2. Common mean or homogeneous population model

Chambers and Clark (2012) studied the homogeneous population model under noninformative sampling design. In this section, we will treat in details the homogeneous population model under informative sampling design and nonignorable nonresponse or informative nonresponse mechanism.

Assume that $y_i \underset{p}{\sim} N\left(\mu,\sigma^2\right)$, $i=1,\ldots,N$ are independent normal random variables, with mean $E_p\left(y_i\right)=\mu$ and variance $V_p\left(y_i\right)=\sigma^2$. According to equation (17), the best linear unbiased predictor T_{in} of T requires the computation of $E_{\bar{r}}\left(y_i|O\right)$ and $E_{\bar{s}}\left(y_i|O\right)$, and this based on the specification of $E_p\left(\pi_i|y_i\right)$ and $E_s\left(\psi_i|y_i\right)$. Different models can be considered for $E_p\left(\pi_i|y_i\right)$ and $E_s\left(\psi_i|y_i\right)$, see Eideh (2003, 2012). In this paper, for illustration, we consider the following models:

(a) Exponential inclusion probability model:

$$E_p\left(\pi_i|y_i\right)=\exp\left(\eta y_i\right) \tag{30}$$

(b) Exponential response probability model:

$$E_s\left(\psi_i|y_i\right)=\exp\left(\gamma y_i\right) \tag{31}$$

According to equations (1) and (2), we can show that the sample distribution of y_i is $y_i \underset{s}{\sim} N\left(\mu + \eta\sigma^2, \sigma^2\right)$, $i = 1, \ldots, n$, and the response distribution of y_i is: $y_i \underset{r}{\sim} N\left(\mu + (\eta + \gamma)\sigma^2, \sigma^2\right)$, $i = 1, \ldots, m$. For estimation of μ, σ^2, η and γ; see Eideh (2016).

Computation of $E_{\bar{s}}(y_i)$. After some algebra we can show that

$$E_p(\pi_i) = E_p\left(E_p(\pi_i | y_i)\right) = E_p\left\{\exp(\eta y_i)\right\} = M_p(\eta) = \exp\left(\eta\mu + \frac{\eta^2\sigma^2}{2}\right) \quad (32)$$

$$\begin{aligned} E_p(y_i\pi_i) &= E_p\left(E_p(\pi_i y_i | y_i)\right) = E_p\left(y_i E_p(\pi_i | y_i)\right) \\ &= E_p\left\{y_i \exp(\eta y_i)\right\} = \frac{d}{d\eta} M_p(\eta) \\ &= \left(\mu + \eta\sigma^2\right)\exp\left(\mu\eta + \frac{\eta^2\sigma^2}{2}\right) = E_s(y_i) M_p(\eta) \end{aligned} \quad (33)$$

where $M_p(\eta)$ is the moment generation function of y_i,

$M_p(\eta) = \exp\left(\mu\eta + \frac{\eta^2\sigma^2}{2}\right)$. So that,

$$Cov_p(\pi_i, y_i) = M_p(\eta)\left\{E_s(y_i) - E_p(y_i)\right\} = \left(\eta\sigma^2\right) M_p(\eta) \quad (34)$$

$$\frac{Cov_p(\pi_i, y_i)}{E_p(1 - \pi_i)} = \left(\eta\sigma^2\right)\frac{M_p(\eta)}{1 - M_p(\eta)} \quad (35)$$

Hence, according to (3b), we have:

$$E_{\bar{s}}(y_i) = \mu - \left(\eta\sigma^2\right)\frac{M_p(\eta)}{1 - M_p(\eta)} \quad (36)$$

Computation of $E_{\bar{r}}(y_i)$. Similarly, according to (8), we can show that:

$$E_{\bar{r}}(y_i) = \mu + \eta\sigma^2 - \left(\gamma\sigma^2\right)\frac{M_s(\gamma)}{1 - M_s(\gamma)} \quad (37)$$

where, $M_s(\gamma) = \exp\left(\gamma\left(\mu + \eta\sigma^2\right) + \frac{\gamma^2\sigma^2}{2}\right)$.

Thus, using (17), (36) and (37), the BLUP for T under informative sampling and nonignorable nonresponse is:

$$T_{in,C}^{*} = T_{ni,C}^{*} + \left\{ (n-m)\left(\eta\sigma^2 - (\gamma\sigma^2)\frac{M_s(\gamma)}{1-M_s(\gamma)} \right) + (N-n)\left(-(\eta\sigma^2)\frac{M_p(\eta)}{1-M_p(\eta)} \right) \right\} \tag{38}$$

where

$$T_{ni,C}^{*} = \sum_{i\in r} y_i + (n-m)\mu + (N-n)\mu = \sum_{i\in r} y_i + (N-m)\mu \tag{39}$$

And the nonresponse bias of $T_{in,C}$ is:

$$\begin{aligned} B(T_{in,C}^{*}) &= E_p(T_{in,C}^{*} - T) = -\left\{ \sum_{i\in\bar{r}} E_p[y_i - E_{\bar{r}}(y_i)] + \sum_{i\in\bar{s}} E_p[y_i - E_{\bar{s}}(y_i)] \right\} \\ &= -\left\{ (n-m)\left[-\eta\sigma^2 + (\gamma\sigma^2)\frac{M_s(\gamma)}{1-M_s(\gamma)} \right] + (N-n)\left[(\eta\sigma^2)\frac{M_p(\eta)}{1-M_p(\eta)} \right] \right\} \end{aligned} \tag{40}$$

Particular cases:

Case 1: The sampling design is noninformative, that is ($\eta = 0$) and the nonresponse process is nonignorable:

$$T_{nn,C}^{*} = \sum_{i\in r} y_i + (n-m)\mu + (N-n)\mu + (n-m)\left\{ -(\gamma\sigma^2)\frac{M_p(\gamma)}{1-M_p(\gamma)} \right\} \tag{41}$$

$$B(T_{nn,C}^{*}) = -\left\{ (n-m)(\gamma\sigma^2)\frac{M_p(\gamma)}{1-M_p(\gamma)} \right\} \tag{42}$$

Case 2: The sampling design is noninformative, that is ($\eta = 0$) and the nonresponse process is ignorable, that is ($\gamma = 0$):

$$T_{ni,C}^{*} = \sum_{i\in r} y_i + (n-m)\mu + (N-n)\mu = \sum_{i\in r} y_i + (N-m)\mu \tag{43}$$

$$B(T_{ni,C}^{*}) = 0 \tag{44}$$

Case 3: The sampling design is informative and the nonresponse process is ignorable, that is ($\gamma = 0$):

$$T^*_{ii,C} = \sum_{i\in r} y_i + (n-m)\mu + (N-n)\mu + \left\{(n-m)\eta\sigma^2 + (N-n)\left(-(\eta\sigma^2)\frac{M_p(\eta)}{1-M_p(\eta)}\right)\right\} \tag{45}$$

$$B\left(T^*_{ii,C}\right) = -\left\{(n-m)\left[-\eta\sigma^2\right] + (N-n)\left[(\eta\sigma^2)\frac{M_p(\eta)}{1-M_p(\eta)}\right]\right\} \tag{46}$$

3.3. Simple ratio population model

The simple ratio population model [®] stating that: $y_i | x_i \underset{p}{\sim} N(\beta x_i, \sigma^2 x_i)$, $i = 1,\ldots,N$ are independent normal random variables, with mean $E_p(y_i|x_i) = \beta x_i$ and variance $Var_p(y_i|x_i) = \sigma^2 x_i$.

Under the exponential inclusion probability model:

$$E_p(\pi_i|y_i,x_i) = \exp(\eta_0 x_i + \eta_1 y_i) \tag{47}$$

And the exponential response probability model:

$$E_s(\psi_i|y_i,x_i) = \exp(\gamma_0 x_i + \gamma_1 y_i) \tag{48}$$

Similarly to the previous section, we can show the following: $y_i \mid x_i \underset{s}{\sim} N\left((\eta_1\sigma^2 + \beta)x_i, \sigma^2 x_i\right)$, $i = 1,\ldots,n$ and $y_i \mid x_i \underset{r}{\sim} N\left((\eta_1\sigma^2 + \gamma_1\sigma^2 + \beta)x_i, \sigma^2 x_i\right)$, $i = 1,\ldots,m$.

$$E_{\bar{s}}(y_i) = \beta x_i - (\eta_1\sigma^2 x_i)\frac{\exp(\eta_0 x_i)M_p(\eta_1)}{1-\exp(\eta_0 x_i)M_p(\eta_1)} \tag{49}$$

where

$$M_p(\eta_1) = \exp\left(\eta_1(\beta x_i) + \frac{\sigma^2 x_i \eta_1^2}{2}\right)$$

$$E_{\bar{r}}(y_i) = (\beta + \eta_1\sigma^2)x_i - (\gamma_1\sigma^2 x_i)\frac{\exp(\gamma_0 x_i)M_s(\gamma_1)}{1-\exp(\gamma_0 x_i)M_s(\gamma_1)} \tag{50}$$

where

$$M_s(\gamma_1) = \exp\left(\gamma_1 (\beta + \eta_1 \sigma^2) x_i + \frac{\sigma^2 x_i \eta_1^2}{2} \right)$$

Hence, the BLUP for T under informative sampling and nonignorable nonresponse is:

$$T_{in,R}^{*} = \sum_{i \in r} y_i + (\sum_{i \in U} x_i - \sum_{i \in r} x_i)\beta + \sum_{i \in \bar{r}} \left\{ \eta_1 \sigma^2 x_i - (\gamma_1 \sigma^2 x_i) \frac{\exp(\gamma_0 x_i) M_s(\gamma_1)}{1 - \exp(\gamma_0 x_i) M_s(\gamma_1)} \right\} + \sum_{i \in \bar{s}} \left\{ -(\eta_1 \sigma^2 x_i) \frac{\exp(\eta_0 x_i) M_p(\eta_1)}{1 - \exp(\eta_0 x_i) M_p(\eta_1)} \right\} \tag{51}$$

And the nonresponse bias of $T_{in,R}^{*}$ is:

$$B(T_{in,R}^{*}) = -\left\{ \sum_{i \in \bar{r}} \left[-(\eta_1 \sigma^2 x_i) + (\gamma_1 \sigma^2 x_i) \frac{\exp(\gamma_0 x_i) M_s(\gamma_1)}{1 - \exp(\gamma_0 x_i) M_s(\gamma_1)} \right] + \sum_{i \in \bar{s}} \left[(\eta_1 \sigma^2 x_i) \frac{\exp(\eta_0 x_i) M_p(\eta_1)}{1 - \exp(\eta_0 x_i) M_p(\eta_1)} \right] \right\} \tag{52}$$

Particular cases:

Case 1: The sampling design is noninformative, that is ($\eta = 0$) and the nonresponse process is nonignorable:

$$T_{nn,R}^{*} = \sum_{i \in r} y_i + (\sum_{i \in U} x_i - \sum_{i \in r} x_i)\beta + \sum_{i \in \bar{r}} \left\{ -(\gamma_1 \sigma^2 x_i) \frac{\exp(\gamma_0 x_i) M_p(\gamma_1)}{1 - \exp(\gamma_0 x_i) M_p(\gamma_1)} \right\} \tag{53}$$

$$B(T_{nn,R}^{*}) = -\sum_{i \in \bar{r}} \left[(\gamma_1 \sigma^2 x_i) \frac{\exp(\gamma_0 x_i) M_p(\gamma_1)}{1 - \exp(\gamma_0 x_i) M_p(\gamma_1)} \right] \tag{54}$$

Case 2: The sampling design is noninformative, that is ($\eta = 0$) and the nonresponse process is ignorable, that is ($\gamma = 0$):

$$T_{ni,R}^{*} = \sum_{i\in r} y_i + (\sum_{i\in U} x_i - \sum_{i\in r} x_i)\beta \tag{55}$$

$$B\left(T_{ni,R}\right) = 0 \tag{56}$$

Case 3: The sampling design is informative and the nonresponse process is ignorable, that is ($\gamma = 0$):

$$T_{ii,R}^{*} == \sum_{i\in r} y_i + (\sum_{i\in U} x_i - \sum_{i\in r} x_i)\beta + \sum_{i\in \bar{s}} \left\{ -\left(\eta_1 \sigma^2 x_i\right) \frac{\exp\left(\eta_0 x_i\right) M_p\left(\eta_1\right)}{1 - \exp\left(\eta_0 x_i\right) M_p\left(\eta_1\right)} \right\} \tag{57}$$

$$B\left(T_{ii,R}^{*}\right) = -\left\{ \sum_{i\in \bar{r}} \left[-\left(\eta_1 \sigma^2 x_i\right) \right] + \sum_{i\in \bar{s}} \left[\left(\eta_1 \sigma^2 x_i\right) \frac{\exp\left(\eta_0 x_i\right) M_p\left(\eta_1\right)}{1 - \exp\left(\eta_0 x_i\right) M_p\left(\eta_1\right)} \right] \right\} \tag{58}$$

3.4. Simple regression population model

The simple regression population model (L) stating that: $y_i | x_i \underset{p}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$, $i = 1, \ldots, N$ are independent normal random variables, with mean $E_p\left(y_i | x_i\right) = \beta_0 + \beta_1 x_i$ and variance $Var_p\left(y_i | x_i\right) = \sigma^2$. Assuming the models given in (48) and (49), we can show the following:

$$y_i \mid x_i \underset{s}{\sim} N\left(\beta_0 + \eta_1 \sigma^2 + \beta_1 x_i, \sigma^2\right),\ i = 1, \ldots, n \tag{59}$$

$$y_i \mid x_i \underset{r}{\sim} N\left(\beta_0 + \eta_1 \sigma^2 + \gamma_1 \sigma^2 + \beta_1 x_i, \sigma^2\right),\ i = 1, \ldots, m \tag{60}$$

$$E_{\bar{s}}\left(y_i\right) = \beta_0 + \beta_1 x_i - \left(\eta_1 \sigma^2\right) \frac{\exp\left(\eta_0 x_i\right) M_p\left(\eta_1\right)}{1 - \exp\left(\eta_0 x_i\right) M_p\left(\eta_1\right)} \tag{61}$$

where

$$M_p\left(\eta_1\right) = \exp\left(\eta_1 \left(\beta_0 + \beta_1 x_i\right) + \frac{\eta_1^2 \sigma^2}{2} \right)$$

$$E_{\bar{r}}\left(y_i\right) = \beta_0 + \beta_1 x_i + \eta_1 \sigma^2 - \left(\gamma_1 \sigma^2\right) \frac{\exp\left(\gamma_0 x_i\right) M_s\left(\gamma_1\right)}{1 - \exp\left(\gamma_0 x_i\right) M_s\left(\gamma_1\right)} \tag{62}$$

where

$$M_s(\gamma_1) = \exp\left(\gamma_1\left(\beta_0 + \beta_1 x_i + \eta_1\sigma^2\right) + \frac{\eta_1^2\sigma^2}{2} \right)$$

Therefore, the BLUP for T under informative sampling and nonignorable nonresponse is:

$$T_{in,L}^{*} = \sum_{i\in r} y_i + (N-m)\beta_0 + \beta_1\left(\sum_{i\in U} x_i - \sum_{i\in r} x_i\right) + \sum_{i\in\bar{r}}\left\{\eta_1\sigma^2 - \left(\gamma_1\sigma^2\right)\frac{\exp(\gamma_0 x_i)M_s(\gamma_1)}{1-\exp(\gamma_0 x_i)M_s(\gamma_1)}\right\} + \sum_{i\in\bar{s}}\left\{-\left(\eta_1\sigma^2\right)\frac{\exp(\eta_0 x_i)M_p(\eta_1)}{1-\exp(\eta_0 x_i)M_p(\eta_1)}\right\} \tag{63}$$

And the nonresponse bias of $\hat{T}_{in,L}$ is:

$$B\left(T_{in,L}^{*}\right) = -\left\{\begin{array}{l} \sum_{i\in\bar{r}}\left[-\eta_1\sigma^2 + \left(\gamma_1\sigma^2\right)\frac{\exp(\gamma_0 x_i)M_s(\gamma_1)}{1-\exp(\gamma_0 x_i)M_s(\gamma_1)}\right] + \\ \sum_{i\in\bar{s}}\left[\left(\eta_1\sigma^2\right)\frac{\exp(\eta_0 x_i)M_p(\eta_1)}{1-\exp(\eta_0 x_i)M_p(\eta_1)}\right] \end{array}\right\} \tag{64}$$

Particular cases:

Case 1: The sampling design is noninformative, that is ($\eta_1 = 0$) and the nonresponse process is nonignorable:

$$T_{nn,L}^{*} = \sum_{i\in r} y_i + (N-m)\beta_0 + \beta_1\left(\sum_{i\in U} x_i - \sum_{i\in r} x_i\right) + \sum_{i\in\bar{r}}\left\{-\left(\gamma_1\sigma^2\right)\frac{\exp(\gamma_0 x_i)M_p(\gamma_1)}{1-\exp(\gamma_0 x_i)M_p(\gamma_1)}\right\} \tag{65}$$

$$B\left(T_{nn,L}^{*}\right) = -\sum_{i\in\bar{r}}\left[\left(\gamma_1\sigma^2\right)\frac{\exp(\gamma_0 x_i)M_p(\gamma_1)}{1-\exp(\gamma_0 x_i)M_p(\gamma_1)}\right] \tag{66}$$

Case 2: The sampling design is noninformative, that is ($\eta_1 = 0$) and the nonresponse process is ignorable, that is ($\gamma_1 = 0$):

$$T_{ni,L}^{*} = \sum_{i\in r} y_i + (N-m)\beta_0 + \beta_1\left(\sum_{i\in U} x_i - \sum_{i\in r} x_i\right) \tag{67}$$
$$B\left(T_{ni,L}^{*}\right) = 0$$

Case 3: The sampling design is informative and the nonresponse process is ignorable, that is ($\gamma_1 = 0$):

$$T_{ii,L}^{*} = \sum_{i\in r} y_i + (N-m)\beta_0 + \beta_1\left(\sum_{i\in U} x_i - \sum_{i\in r} x_i\right) + \eta_1\sigma^2(n-m) + \sum_{i\in \bar{s}}\left\{-(\eta_1\sigma^2)\frac{\exp(\eta_0 x_i)M_p(\eta_1)}{1-\exp(\eta_0 x_i)M_p(\eta_1)}\right\} \tag{68}$$

$$B\left(T_{ii,L}^{*}\right) = -\left\{\sum_{i\in \bar{r}} -\eta_1\sigma^2 + \sum_{i\in \bar{s}}\left[(\eta_1\sigma^2)\frac{\exp(\eta_0 x_i)M_p(\eta_1)}{1-\exp(\eta_0 x_i)M_p(\eta_1)}\right]\right\} \tag{69}$$

Remark: Empirical BLUP of T . In practice the parameters are unknown, so in order to obtain the empirical best unbiased predictors, we replace the unknown parameters by their estimates, see Eideh (2016).

4. Conclusions

In this paper we combine two methodologies used in the model-based survey sampling: the prediction of finite population total T, under informative sampling, and full response, and the prediction of T when the sampling design is noninformative and the nonresponse mechanism is nonignorable. One incorporates the dependence of the first order inclusion probabilities on the study variable, while the other incorporates the dependence of the probability of nonresponse on unobserved or missing observations. For this aim, we use the response distribution and relationships between moments of the superpopulation, sample, sample-complement, response, and non-response distributions, for the prediction of finite population totals. Common best linear unbiased predictors derived under model-based survey sampling are shown to be special cases of the present theory. The general theory was applied for a homogeneous

population model, simple ratio population model, simple linear regression population model, and multiple regression population model. The paper is purely mathematical and focuses on the role of informativeness of the sampling design and informativeness of nonresponse in adjusting various predictors for bias reduction. Further experimentation (simulation and real data problem) with this kind of predictors is therefore highly recommended. We hope that the new mathematical results obtained will encourage further theoretical, empirical and practical research in these directions.

Acknowledgements

The research was partially supported by a grant from DAAD (Deutscher Akademischer Austauschdienst German Academic Exchange Service) - Research Stays for University Academics and Scientists, 2018. Also, the author would like to thank Professor Timo Schmid for hosting the author during his visit to Free University Berlin.

REFERENCES

NONRESPONSE and WEIGHTING ADJUSTMENTS: A Critical Review, Journal of Official Statistics 29, pp. 329–353.

CHAMBERS, R. L., CLARK, R. G., (2012). An Introduction to Model-Based Survey Sampling with Applications, Oxford Statistical Science Series.

EIDEH, A. H., (2016). Estimation of Finite Population Mean and Superpopulation Parameters when the Sampling Design is Informative and Nonresponse Mechanism is Nonignorable. Pakistan Journal of Statistics and Operation Research (PJSOR), Pak.j.stat.oper.res. Vol. XII No. 3, 2016, pp. 467–489.

EIDEH A. H., (2012). Estimation and Prediction under Nonignorable Nonresponse via Response and Nonresponse Distributions, Journal of Indian Society of Agriculture Statistics., 66(3) 2012, pp 359–380.

EIDEH A. H., (2009). On the use of the sample distribution and sample likelihood for inference under informative probability sampling. DIRASAT (Natural Science), Volume 36 (2009), Number 1, pp18–29.

EIDEH, A. H., (2007). Method of Moments Estimators of Finite Population Parameters in Case of Full Response and Nonresponse, Contributed Paper for the 56th Biennial Session of the International Statistical Institute, August 22–29, 2007, Lisboa, Portugal, ISI 2007 Book of Abstracts, p 430.

EIDEH, A. H., NATHAN, G., (2006). Fitting time series models for longitudinal survey data under informative sampling, Journal of Statistical Planning and Inference 136, 9, pp 3052–306, [Corrigendum, 137 (2007), p 628].

LITTLE, R. J. A., (1983). Superpopulation models for nonresponse. In Incomplete Data in Sample Surveys, Vol. 2, Theory and bibliographies, part VI, pp. 337–413. New York, Academic Press.

LITTLE, R. J. A., (2003). Bayesian methods for unit and item nonresponse, In Analysis of survey data, Chambers R.L. and Skinner C.L., pp. 289–306, Wiley, New York.

LITTLE, R. J. A., RUBIN, D. B., (2002). Statistical analysis with missing data. New York: Wiley.

LITTLE, R.J.A., VARTIVARIAN, S., (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means?, Survey Methodology, Vol. 31, No. 2, pp. 161–168.

PFEFFERMANN, D., KRIEGER, A. M., RINOTT, Y., (1998). Parametric distributions of complex survey data under informative probability sampling'. Statistica Sinica, 8, pp 1087–1114.

PFEFFERMANN, D., SVERCHKOV, M., (1999). Parametric and semi-parametric estimation of regression models fitted to survey data, Sankhya, 61, B, pp 166–186.

PFEFFERMANN, D., SIKOV, A., (2011). Imputation and Estimation under Nonignorable Nonresponse in Household Surveys with Missing Covariate Information, Journal of Official Statistics, Vol. 27, No. 2, 2011, pp. 181–209.

RIDDLES, M. K., KIM, J. K., IM, J., (2016). A propensity-score-adjustment method for nonignorable nonresponse, J. Surv. Stat. Methodol., 4, pp. 215–245.

SÄRNDAL, C .E., (2011). The 2010 Morris Hansen lecture dealing with survey nonresponse in data collection, in estimation, Journal of Official Statistics, Vol. 27, No. 1, 2011, pp. 1–21.

SVERCHKOV, M., PFEFFERMANN, D., (2004). Prediction of finite population totals based on the sample distribution, Survey Methodology, 30, pp 79–92.

SVERCHKOV, M., (2008), A New Approach to Estimation of Response Probabilities When Missing Data Are Not Missing at Random, Proceedings of the Survey Research Methods Section, pp. 867–874.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1, pp. 37–54, DOI 10.21307/stattrans-2020-003
Submitted – 20.05.2019; accepted for publishing – 10.12.2019

Horvitz-Thompson estimator based on theauxiliary variable

J. Al-Jararha[1], **Mazen Sulaiman**[2]

ABSTRACT

In this paper, the Horvitz and Thompson (1952) estimator will be modified; so that, the modified estimators will use the availability of the auxiliary variable. Furthermore, the modified estimators are extended to be used in stratified sampling designs. Empirical studies are given for comparison purposes.

Key words: Horvitz-Thompson Estimator, Stratified Sampling Designs, Dual Calibration, GREG Type Estimator.

1. Introduction

Consider the finite population U of N units indexed by the set $\{1,2,\cdots,N\}$. For the ith unit, let y_i be the value of the interest variable Y, and x_i be the value of the auxiliary variable X. The values of X are known for all the units in the population and correlated with the study variable Y. Without loss of generality, we can assume that $x_i > 0$ for $i = 1, 2, \ldots, N$. Based on a probability sampling design $p(.)$, draw a random sample s from U. The first order inclusion probability π_i is defined by $\pi_i = \sum_{s \ni i} p(s)$, and the second inclusion probability π_{ij} is defined by $\pi_{ij} = \sum_{s \ni i,\, j} p(s)$, for $i \neq j$, and $\pi_{ij} = \pi_i$ when $i = j$. The probability sampling design $p(.)$ is assumed to be a measurable design. The population total for the auxiliary variable X is $t_x = \sum_{i \in U} x_i$.

Horvitz and Thompson (1952) proposed the following estimator

$$\begin{aligned} \hat{t}_{y\pi} &= \sum_{i \in U} \frac{y_i}{\pi_i} I_{\{i \in s\}} \\ &= \sum_{i \in s} d_i y_i \end{aligned} \tag{1}$$

to estimate the finite population total $t_y = \sum_{i \in U} y_i$, where $d_i = 1/\pi_i$ are the sampling design weights and $I_{\{i \in s\}}$ is one if $i \in s$ and zero otherwise. The $\hat{t}_{y\pi}$ is exactly an unbiased estimator for t_y.

Remark 1.1 *The availability and the calibration on the auxiliary variables can be used to increase the precision of estimators. However, the Horvitz and Thompson (1952) estimator does not use the availability of the auxiliary variables. Therefore, the Horvitz and Thompson (1952) estimator will be modified, so that the modified estimators will use the availability of the auxiliary variable.*

[1]Department of Statistics, Yarmouk University, Irbid, Jordan. E-mail: jehad@yu.edu.jo. ORCID: https://orcid.org/0000-0001-8233-9849.
[2]Department of Statistics, Yarmouk University, Irbid, Jordan . E-mail: jomaa_mazen@yahoo.com

Deville and Särndal proposed the following estimator

$$\hat{t}_{y.ds} = \sum_{i \in U} w_i y_i I_{\{i \in s\}} = \sum_{i \in s} w_i y_i, \tag{2}$$

for estimating t_y, where w_i, $i \in s$, are the new sampling design weights that calibrated the sampling design weights d_i defined by Eq.(1) based on the calibration on the known population total for the auxiliary variable X and the chi-square distance. The calibrated weights w_i are obtained by minimizing the chi-square distance, subject to the side condition. As a result of this, the calibrated weights w_i are given by

$$w_i = d_i + \frac{t_x - \hat{t}_{x\pi}}{\sum_{i \in s} d_i q_i x_i^2} d_i q_i x_i, \tag{3}$$

Therefore, Eq. (2) is reduced to

$$\hat{t}_{y.ds} = \hat{t}_{y\pi} + \hat{\beta}_{ds} (t_x - \hat{t}_{x\pi}) \tag{4}$$

which is a GREG type estimator, where q_i's are known positive weights unrelated to d_i, $\hat{\beta}_{ds} = \frac{\sum_{i \in s} d_i q_i x_i y_i}{\sum_{i \in s} d_i q_i x_i^2}$, and $\hat{t}_{x\pi}$ is the Horvitz and Thompson (1952) estimator of t_x.

Stearns and Singh (2008) summarized the developments by several researchers on the GREG estimators and used the calibration idea to propose three new estimators of the variance of the GREG estimators.

Singh (2013) estimated t_y based on the dual calibration approach and his approach is summarized by the following.

Let

$$\hat{t}_{sin} = \sum_{i \in s} \omega_i x_i \tag{5}$$

subject to

$$\sum_{i \in s} d_i = \sum_{i \in s} \omega_i \tag{6}$$

and a new constraint α defined by

$$\alpha = \frac{1}{2} \sum_{i \in s} \frac{(\omega_i - d_i)^2}{d_i q_i} \tag{7}$$

As a result of this, the proposed estimator is

$$\hat{t}_{y.sin} = \hat{t}_{y\pi} + \hat{\beta}_{sin} (t_x - \hat{t}_{x\pi}) \tag{8}$$

to estimate the finite population total t_y; $\hat{t}_{y.sin}$ is a GREG type estimator, where

$$\hat{\beta}_{sin} = \frac{S_{xy}}{S_{xx}}, \tag{9}$$

where

$$S_{xy} = \sum_{i \in s} d_i q_i \left(y_i - \frac{\sum_{i \in s} d_i q_i y_i}{\sum_{i \in s} d_i q_i} \right) \left(x_i - \frac{\sum_{i \in s} d_i q_i x_i}{\sum_{i \in s} d_i q_i} \right) \tag{10}$$

and

$$S_{xx} = \sum_{i \in s} d_i q_i \left(x_i - \sum_{i \in s} d_i q_i x_i / \sum_{i \in s} d_i q_i \right)^2 \tag{11}$$

Two concerns about Eq.(8) are raised by Singh (2013), Remark 1 and Remark 2. Al-Yaseen (2014) showed that the estimator given by Eq.(8) can be obtained theoretically, which clarifies the first concern mentioned in Remark 1. Al-Jararha (2015) made an attempt to suggest a way to use the dual calibration of the design weights in the case of multi-auxiliary variables; in other words, an attempt to give an answer to the second concern in Remark 2.

Sugden and Smith (2002) defined the term strictly linear estimator and proposed two exactly unbiased estimators for the general linear estimates. The possibility of construction an exactly unbiased estimator from a general linear estimator, the constructed unbiased estimator is called a strictly linear estimate. Consider the general linear estimates of t_y, defined by Godambe (1955), to be of the form

$$\hat{t}_y = \sum_{i \in s} b_{si} y_i. \tag{12}$$

The exactly unbiased estimators, based on the Sugden and Smith (2002) approach, from $\hat{t}_y$ are defined by

$$\hat{t}_{y(1)} = \hat{t}_y - \sum_{i \in s} (B_i - 1) y_i / \pi_i \tag{13}$$

and

$$\hat{t}_{y(2)} = \sum_{i \in s} b_{si} y_i / B_i \tag{14}$$

for estimating the finite population total t_y, where

$$B_i = \sum_{s \ni i} p(s) b_{si}. \tag{15}$$

Recently, different authors have adopted the calibration technique to modify the original weights in stratified sampling designs. In the case of stratified sampling designs, Nidhi, Sisodia, Singh and Singh (2017) proposed a class of calibration estimators for estimating

the population mean. Based on the availability of two auxiliary variables in the study and in the case of stratified sampling designs, Ozgul (2018) proposed a calibration estimator for estimating the population mean.

The Horvitz and Thompson (1952) estimator is well known in survey sampling for estimating the finite population total t_y. However, this estimator does not use the availability of the auxiliary variable. In order to improve the precision of this estimator, an attempt to generalize this estimator will be given, so that the modified Horvitz and Thompson (1952) estimators will use the availability of the auxiliary variable. Furthermore, our approach can be applied in the case of stratified sampling designs.

2. Proposed Approach

Based on the dual calibration approach, the estimator

$$\hat{t}_{y.new} = \sum_{i \in S} \omega_i y_i \tag{16}$$

is proposed to estimate the finite population total t_y, by modifying the constraint α of the Singh (2013) approach. In other words, redefine α as

$$\alpha = \frac{1}{2} \sum_{i \in s} \frac{(\omega_i - d_i)^2}{d_i q_i} + \frac{1}{2} \phi^2 \sum_{i \in s} \frac{\omega_i^2}{d_i q_i}, \tag{17}$$

where ϕ is a positive quantity.

The problem now is to minimize

$$\hat{t}_x = \sum_{i \in s} \omega_i x_i \tag{18}$$

with respect to ω_i subject to

$$\sum_{i \in s} \omega_i = \sum_{i \in s} d_i \tag{19}$$

and a new constraint α defined by Eq.(17).

The Lagrange function is defined by

$$l = \sum_{i \in s} \omega_i x_i - \lambda_1 \left(\sum_{i \in s} \omega_i - \sum_{i \in s} d_i \right) - \lambda_2 \left(\frac{1}{2} \sum_{i \in s} \frac{(\omega_i - d_i)^2}{d_i q_i} + \frac{1}{2} \phi^2 \sum_{i \in s} \frac{\omega_i^2}{d_i q_i} - \alpha \right) \tag{20}$$

where λ_1 and λ_2 are the Lagrange multipliers.

Differentiating the right hand side of Eq.(20) with respect to ω_i, equate to zero, and

solving for ω_i, we have

$$\omega_i = \frac{1}{1+\phi^2}\left(d_i + \frac{d_i q_i}{\lambda_2}(x_i - \lambda_1)\right) \tag{21}$$

Summing both sides of Eq.(21) over all possible sampled values and using Eq.(19), we have

$$\lambda_1 = \frac{1}{\sum_{i\in s} d_i q_i}\left(\sum_{i\in s} d_i q_i x_i - \phi^2 \lambda_2 \sum_{i\in s} di\right) \tag{22}$$

Now, substituting Eq.(21) into Eq.(17), we have

$$2\alpha\left(1+\phi^2\right)\lambda_2^2 = \phi^2\lambda_2^2 \sum_{i\in s}\frac{d_i}{q_i} + \sum_{i\in s} d_i q_i x_i^2 - 2\lambda_1 \sum_{i\in s} d_i q_i x_i + \lambda_1^2 \sum_{i\in s} d_i q_i \tag{23}$$

Substituting Eq.(22) into Eq.(23), we have

$$\lambda_2 = \pm\frac{1}{c}\sqrt{\sum_{i\in s} d_i q_i \left(x_i - \sum_{i\in s} d_i q_i x_i / \sum_{i\in s} d_i q_i\right)^2} \tag{24}$$

where

$$c = \sqrt{2\alpha\left(1+\phi^2\right) - \phi^2 \sum_{i\in s} d_i/q_i - \phi^4 \left(\sum_{i\in s} d_i\right)^2 / \sum_{i\in s} d_i q_i}. \tag{25}$$

Ignore the negative sign, where the sign is to be determined by the choice of the sign of c. Substituting Eq.(24) into Eq.(22) and using the result in Eq.(21), multiplying ω_i by y_i and summing over $i \in s$ we have

$$\hat{t}_{y.new} = \frac{1}{1+\phi^2}\left(\sum_{i\in s} d_i y_i + \phi^2 \left(\sum_{i\in s} d_i / \sum_{i\in s} d_i q_i\right) \sum_{i\in s} d_i q_i y_i + \delta c\right) \tag{26}$$

where

$$\delta \quad = \quad S_{xy}/\sqrt{S_{xx}}, \tag{27}$$

where c, S_{xy}, and S_{xy} are given by Eq.(25), Eq.(10), and Eq.(11) respectively. With the same reasons adopted by Singh (2013), the best choice of c is

$$c = \frac{t_x - \hat{t}_{x\pi}}{\sqrt{S_{xx}}} \sim N(0,1);$$

therefore,

$$\hat{t}_{y.new} = \lambda \hat{t}_{y.sin} + (1-\lambda)\widetilde{t}_{y\pi}, \tag{28}$$

where $\lambda = 1/\left(1+\phi^2\right)$, $\hat{t}_{y.sin}$ is defined by Eq.(8), and

$$\widetilde{t}_{y\pi} = \frac{\hat{t}_{1\pi}}{\hat{t}_{q\pi}} \hat{t}_{qy\pi}. \tag{29}$$

Furthermore, $\hat{t}_{1\pi} = \sum_{i\in s}(1/\pi_i)$, $\hat{t}_{q\pi} = \sum_{i\in s}(q_i/\pi_i)$, and $\hat{t}_{qy\pi} = \sum_{i\in s}(q_i y_i/\pi_i)$ be the Horvitz and Thompson (1952) estimators for N, t_q, and t_{qy}, respectively.

Remark 2.1 *Since $\lambda \in (0,1)$, Eq.(28) is a convex transformation between $\hat{t}_{y.sin}$ and $\widetilde{t}_{y\pi}$, defined by Eq.(8) and Eq.(29) respectively. At the same time, as $\phi^2 \to \infty \Rightarrow \lambda \to 0 \Rightarrow \hat{t}_{y.new} \to \widetilde{t}_{y\pi}$; moreover, as $\phi^2 \to 0 \Rightarrow \lambda \to 1 \Rightarrow \hat{t}_{y.new} \to \hat{t}_{y.sin}$.*

The performance of $\hat{t}_{y.new}$ will be discussed through simulations from real data set. We will compare $\hat{t}_{y.new}$, $\hat{t}_{y.sin}$, and $\widetilde{t}_{y\pi}$. Consider the FEV data set which was used by Singh (2013) and downloaded from http://www.amstat.org/publications/jse/datasets/fev.dat.txt. Let Y be the Forced expiratory volume, $t_y = 1724$; and the auxiliary variable X be the Children height in inches, $t_x = 39988$. Our aim is to estimate t_y by using $\hat{t}_{y.new}$, $\hat{t}_{y.sin}$, and $\widetilde{t}_{y\pi}$. To achieve our aim, simulate $\upsilon = 3000$ independent random samples from the FEV data set by using procedure `surveyselect` of SAS Institute, under SRSWR design. For $q_i = x_i$ and based on the random samples, estimate t_y by $\hat{t}_{y.new}$, $\hat{t}_{y.sin}$, and $\widetilde{t}_{y\pi}$. Furthermore, compute the empirical mean (Em.Mean), relative bias (RB), and empirical relative mean squares error (REMSE) of the estimators $\hat{t}_{y.new}$, $\hat{t}_{y.sin}$, and $\widetilde{t}_{y\pi}$; where

$$\text{EM.Mean}\left(\hat{t}_y^*\right) = \frac{1}{\upsilon}\sum_{i=1}^{\upsilon}\left(\hat{t}_y^*\right)_i \tag{30}$$

$$\text{RB}\left(\hat{t}_y^*\right) = \frac{\text{EM.Mean}\left(\hat{t}_y^*\right) - t_y}{t_y} \times 100\% \tag{31}$$

$$\text{REMSE}\left(\hat{t}_y^*\right) = \frac{\sum_{i=1}^{\upsilon}\left(\hat{t}_y^* - t_y\right)^2}{\sum_{i=1}^{\upsilon}\left(\hat{t}_{y.new} - t_y\right)^2}, \tag{32}$$

where $\text{EM.Mean}\left(\hat{t}_y^*\right)$, $\text{RB}\left(\hat{t}_y^*\right)$, and $\text{REMSE}\left(\hat{t}_y^*\right)$ are the empirical mean, relative bias, and relative mean squares error of the estimator $\hat{t}_y^*$. For $n = 25, 35, 45, 55, 65,$ and 75. The results are summarized in Table (1).

From Table (1), in the sense of REMSE, the estimator $\hat{t}_{y.sin}$ performs better than $\hat{t}_{y.new}$ and $\widetilde{t}_{y\pi}$ for all values of n and for the different values of $\lambda = 0, 0.25, 0.5, 0.75,$ and 1. However, $\text{REMSE}\left(\widetilde{t}_{y\pi}\right)$ varies from 1 to 6.74; at the same time, $\text{REMSE}\left(\widetilde{t}_{y\pi}\right) = 6.74$ is attainable for large $n = 75$. From this point, concentrations will be focused on the performance of $\widetilde{t}_{y\pi}$ in order to improve the performance of $\hat{t}_{y.new}$. The remaining of this article will be focused on the improvement of $\widetilde{t}_{y\pi}$.

Remark 2.2 *The Horvitz and Thompson (1952) estimator defined by Eq.(1) is a special case from Eq.(29), namely when $q_i = 1$ (or a positive constant). Hence, $\widetilde{t}_{y\pi}$ is modified $\hat{t}_{y\pi}$ for estimating the finite population total t_y. Further, $\widetilde{t}_{y\pi}$ uses the availability of the auxiliary variable through q_i's.*

To the first order and by using Taylor expansion, expanding the right hand side of Eq.(29), we have

$$\widetilde{t}_{y\pi} \simeq \frac{t_1}{t_q} t_{qy} + \frac{t_{qy}}{t_q}\left(\hat{t}_{1\pi} - t_1\right) + \frac{t_1}{t_q}\left(\hat{t}_{qy\pi} - t_{qy}\right) - \frac{t_1 t_{qy}}{t_q^2}\left(\hat{t}_{q\pi} - t_q\right) \tag{33}$$

Therefore, the bias of $\widetilde{t}_{y\pi}$ is given by

$$Bias\left(\widetilde{t}_{y\pi}\right) = t_y - \frac{t_1}{t_q} t_{qy}. \tag{34}$$

Remark 2.3 *It is clear from Eq.(34) that $\widetilde{t}_{y\pi}$ is a biased estimator for estimating the finite population total t_y. However, $\widetilde{t}_{y\pi}$ is a strictly linear estimator; therefore, we can deduce two exactly unbiased estimators from $\widetilde{t}_{y\pi}$ based on Sugden and Smith (2002).*

From Eq.(29), rewrite $\widetilde{t}_{y\pi}$ as

$$\widetilde{t}_{y\pi} = \sum_{i \in s} b_{si} y_i, \tag{35}$$

where

$$b_{si} = \frac{q_i/\pi_i}{\sum_{i\in s}\left(q_i/\pi_i\right)/\sum_{i\in s}\left(1/\pi_i\right)}. \tag{36}$$

From Eq.(15), recall the definition of B_i,

$$\begin{aligned} B_i &= \sum_{s \ni i} p(s)\, b_{si} \\ &= \frac{q_i}{\pi_i} \sum_{s \ni i}\left[p(s)\,\frac{\sum_{i\in s}\left(1/\pi_i\right)}{\sum_{i\in s}\left(q_i/\pi_i\right)}\right] \end{aligned} \tag{37}$$

Based on Sugden and Smith (2002) approach, the two exactly unbiased estimators deduced from $\widetilde{t}_{y\pi}$ for t_y are

$$\widetilde{t}_{y\pi(1)} = \widetilde{t}_{y\pi} - \sum_{i\in s}\left(B_i - 1\right) y_i/\pi_i \tag{38}$$

and

$$\widetilde{t}_{y\pi(2)} = \sum_{i\in s} \frac{b_{si}}{B_i} y_i \tag{39}$$

where b_{si} and B_i are defined by Eq.(36) and Eq.(37) respectively.

Remark 2.4 *Eq.(35) shows that $\widetilde{t}_{y\pi}$ is a general linear estimator of t_y. Furthermore, $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ are two exactly unbiased estimators for t_y deduced from $\widetilde{t}_{y\pi}$; therefore, $\widetilde{t}_{y\pi}$ is a strictly linear estimator based on the Sugden and Smith (2002) definition. Hence, the estimators $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ are generalization of the Horvitz and Thompson (1952) estimator and use the availability of the auxiliary variable.*

Since $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ are exactly unbiased estimators for t_y, the infinite number of exactly unbiased estimators is defined by

$$\widetilde{t}_{y\pi} = \omega\widetilde{t}_{y\pi(1)} + (1-\omega)\widetilde{t}_{y\pi(2)}, \quad \text{for} \quad 0 \leq \omega \leq 1. \tag{40}$$

Remark 2.5 *The estimator $\widetilde{t}_{y\pi}$ is a convex transformation and an unbiased estimator for estimating the population total t_y.*

2.1. Modified Horvitz-Thompson and Stratified Sampling Designs

The finite population U of size N is divided into L non-overlapping strata $U_1, U_2, \ldots, U_L$; $U = \bigcup_{h=1}^{L} U_h$. The population total for the h^{th} stratum is $t_{yh} = \sum_{i \in U_h} y_i$. Furthermore, the h^{th} stratum is of size N_h and $N = \sum_{h=1}^{L} N_h$. The population total t_y is redefined as

$$t_y = \sum_{h=1}^{L} t_{yh}. \tag{41}$$

For the h^{th} stratum and based on a measurable sampling design $p_h(.)$, draw a random sample s_h of size n_h from U_h. Assume $\bar{x}_h = \sum_{i \in U_h} x_i/N_h$ is known for $h = 1, 2, \ldots, L$. Apply $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ to the h^{th} stratum. In other words, estimate t_{yh} by

$$\widetilde{t}_{y\pi(1).h} \quad = \quad \widetilde{t}_{y\pi.st} - \sum_{i \in s_h} (B_i - 1)\, y_i/\pi_i, \tag{42}$$

or by

$$\widetilde{t}_{y\pi(2).h} \quad = \quad \sum_{i \in s_h} \frac{b_{s_{hi}}}{B_i} y_i. \tag{43}$$

In this case, $\widetilde{t}_{y\pi(1).h}$ and $\widetilde{t}_{y\pi(2).h}$ are exactly two unbiased estimators for t_{yh}, where

$$\widetilde{t}_{y\pi.st} = \sum_{h=1}^{L} \frac{\hat{t}_{1\pi.h}}{\hat{t}_{q\pi.h}} \hat{t}_{qy\pi.h}; \tag{44}$$

$\hat{t}_{1\pi.h} = \sum_{i \in s_h} (1/\pi_i)$, $\hat{t}_{q\pi.h} = \sum_{i \in s_h} (q_i/\pi_i)$, and $\hat{t}_{qy\pi.h} = \sum_{i \in s_h} (q_i y_i/\pi_i)$ be the Horvitz and Thompson (1952) estimators for N_h, $t_{q.h}$, and $t_{qy.h}$, respectively.

From Eq.(41), estimate t_y by

$$\widetilde{t}_{y\pi(1).st} \quad = \quad \sum_{h=1}^{L} \widetilde{t}_{y\pi(1).h}, \tag{45}$$

or by

$$\widetilde{t}_{yh\pi(2).st} = \sum_{h=1}^{L} \widetilde{t}_{y\pi(2).h}, \tag{46}$$

where $\widetilde{t}_{y\pi(1).h}$ and $\widetilde{t}_{y\pi(2).h}$ are defined in Eq.(42) and Eq.(43), respectively.

Remark 2.6 *The two estimators $\widetilde{t}_{y\pi(1).h}$ and $\widetilde{t}_{y\pi(2).h}$ are two exactly unbiased estimators for t_{yh}. Based on this idea, the two estimators $\widetilde{t}_{y\pi(1).st}$ and $\widetilde{t}_{y\pi(2).st}$ are exactly unbiased estimators for t_y; therefore, the accumulation of bias across strata is avoided.*

2.2. Special Cases

The exactly unbiased estimators $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ are given by Eq.(38) and Eq.(39) respectively, deduced from the modified HT estimator $\widetilde{t}_{y\pi}$, depending on the weight q_i. Therefore, $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ can use the availability of the auxiliary variables through q_i. In this section, different special cases are considered.

As we mentioned earlier, $\widetilde{t}_{y\pi}$ reduces to $\hat{t}_{y\pi}$, the ordinary Horvitz and Thompson (1952) estimator, when q_i's are one or positive constant. Furthermore, from Eq.(36), $b_{si} = 1/\pi_i$ and from Eq.(37), $B_i = 1$. Therefore,

$$\widetilde{t}_{y\pi(1)} = \widetilde{t}_{y\pi(2)} = \hat{t}_{y\pi}, \tag{47}$$

i.e. $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ are identical; in other words, $\widetilde{t}_{y\pi}$ is exactly an unbiased estimator for t_y. In this case, the Sugden and Smith (2002) approach gives exactly one unbiased estimator for estimating t_y.

Draw a random sample s of size n from the population U of size N by using the simple random sample without replacement (SRSWR) design. Under SRSWR design, $p(s) = 1/\begin{pmatrix} N \\ n \end{pmatrix}$ and $\pi_i = n/N$. Consider the following two cases:

a. $q_i = \pi_i$.

In this case, $b_{si} = \frac{N}{n}$ and $B_i = 1$. Therefore,

$$\hat{t}_{y\pi} = \widetilde{t}_{y\pi(1)} = \widetilde{t}_{y\pi(2)} = N\bar{y}_s, \tag{48}$$

which is well-known estimator for estimating t_y, where $\bar{y}_s = \sum_{i=1}^{n} y_i/n$. In this case, the two exactly unbiased estimators based on Sugden and Smith (2002) are reduced to one unbiased estimator, i.e. the Sugden and Smith (2002) approach produces exactly only one unbiased estimator.

b. $q_i = x_i, x_i > 0$.

In this case, $b_{si} = Nx_i / \sum_{j\in s} x_j$ and $B_i = Nx_i p(s) \sum_{s\ni i} \left(\sum_{j\in s} x_j\right)^{-1}$. Therefore,

$$\widetilde{t}_{y\pi(1)} = N\left[\frac{\sum_{i\in s} x_i y_i}{\sum_{i\in s} x_i} + \bar{y}_s - \frac{1}{\binom{N-1}{n-1}} \sum_{i\in s}\left\{\sum_{s\ni i}\left(\sum_{j\in s} x_j\right)^{-1}\right\} x_i y_i\right], \tag{49}$$

and

$$\widetilde{t}_{y\pi(2)} = \binom{N}{n} \sum_{i\in s}\left[\frac{y_i}{\sum_{s\ni i}\left(\sum_{j\in s} x_j\right)^{-1}}\right] / \sum_{j\in s} x_j \tag{50}$$

$$= \hat{T}_{R(2)}, \tag{51}$$

where $\hat{T}_{R(2)}$ is an estimate of t_y defined by Sugden and Smith (2002), Eq.(4.5).

3. Empirical Studies

Sugden and Smith (2002) considered the ratio estimator

$$\hat{T}_R = t_x \frac{\hat{t}_{y\pi}}{\hat{t}_{x\pi}} \tag{52}$$

as a general linear estimator for the population total t_y. $\hat{T}_R$ is asymptotically an unbiased estimator of t_y. Since $\hat{T}_R$ produces two exactly unbiased estimators of t_y, $\hat{T}_R$ is a strictly linear estimator for t_y. Under SRSWR, $B_{Ri} = t_x \sum_{s\ni i}\left(\sum_{j\in s} x_j\right)^{-1} / \binom{N}{n}$. In this case, the exactly unbiased estimators are

$$\hat{T}_{R(1)} = \hat{T}_R - \frac{N}{n}\sum_{i\in s}(B_{Ri} - 1) y_i, \tag{53}$$

and $\widetilde{T}_{R(2)}$, defined by Eq.(51).

Assume all the values of the auxiliary variable are available in the study; under SRSWR design, the estimators $\hat{t}_{y\pi}$, $\widetilde{t}_{y\pi(1)}$, $\widetilde{t}_{y\pi(2)}$, $\hat{T}_{R(2)}$, $\hat{T}_R$, and $\hat{T}_{R(1)}$ defined by Eq.(48), (49), (50), (51), (52), (53) respectively, will be used in the empirical studies.

Consider the data set given by Example(4.9), Page 139, Lohr (2010). In this example, X is the photo counts of dead trees and Y is the field counts of dead trees; $N = 25$, $t_x = 265$, and $t_y = 289$. From this data set, under SRSWR, draw all random samples of sizes $n = 2, 3, 4$. The computations are implemented by using a SAS program written under the iml procedure. The number of all random samples is $m = 300, 2300, 12650$ for $n = 2, 3, 4$ respectively. The relative efficiency of the ratio family is defined by $MSE\left(\hat{T}_{R(i)}\right)/MSE\left(\hat{T}_R\right)$ and relative efficiency of the Horvitz and Thompson (1952) family is defined by $MSE\left(\widetilde{t}_{y\pi(i)}\right)/MSE\left(\hat{t}_{y\pi}\right)$ for $i = 1, 2$. The results are given in Table(2).

In the case of a stratification, consider the data set cars93 from Scheaffer, Menden-

hall and Ott (2006). The data set cars93 consists of different variables; for our study, let $X := MPGCITY$, $Y := MPGHIGH$, and the stratifications based on the variable "typecode". The cars93 data set is summarized by the following table.

h^{th} stratum	1	2	3	4	5	6	total
N_h	20	16	22	11	14	9	$N = 92$
t_{xh}	598	363	430	202	305	153	$t_x = 2051$
t_{yh}	712	478	588	294	403	197	$t_y = 2672$

For the h^{th} stratum, $h = 1, \ldots, 6$, the results are given in Tables (3),...,(8) respectively. Based on the stratified sampling design, the population total t_y is estimated by using the estimators $\hat{t}_{y\pi}$, $\widetilde{t}_{y\pi(1)}$, $\widetilde{t}_{y\pi(2)}$, $\hat{T}_{R(2)}$, $\hat{T}_R$, and $\hat{T}_{R(1)}$; for $n = 12,\ 18,\ 24$. The results are given in Table (9). At the same time, Table (9) is computed from Tables (3),...,(8).

4. Concluding Remarks

In this paper, the Horvitz and Thompson (1952) estimator is modified so that the modified estimators can use the availability of the auxiliary variable in the study. Based on the Sugden and Smith (2002) approach, two exactly unbiased estimators for estimating the population total t_y are deduced from the modified estimator. Furthermore, the exactly two unbiased estimators can be used in stratified sampling designs.

From Table(2), the deduced estimators $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ are exactly unbiased estimators for estimating t_y and perform better than the original Horvitz and Thompson (1952) estimator $\hat{t}_{y\pi}$, in the sense of relative efficiency. Moreover, Table(2) supports the same conclusion mentioned by Sugden and Smith (2002), i.e. the estimators $\hat{T}_{R(1)}$ and $\hat{T}_{R(2)}$ are exactly unbiased estimators and perform better than the original ratio estimator $\hat{T}_R$, in the sense of relative efficiency.

Based on the Sugden and Smith (2002) approach, the two exactly unbiased estimators based on their families for estimating t_y perform better than the original estimators even if the original estimators are asymptotically unbiased or unbiased estimators. Furthermore, the estimators deduced from Horvitz and Thompson (1952) perform better than the deduced estimators from the ratio estimator. Small sample sizes are usually selected in the case of stratified sampling design; moreover, the deduced estimators can be applied to every stratum and aggregated together to estimate the population total.

For $h = 1, \ldots, 6$ the results are given by Tables (3),...,(8), respectively. Table (9) is computed from Tables (3),...,(8), and shows that $\widetilde{t}_{y\pi(1)}$, $\widetilde{t}_{y\pi(2)}$, $\hat{T}_{R(1)}$, and $\hat{T}_{R(2)}$ are exactly unbiased estimators for t_y. Furthermore, the bias of the ratio estimator $\hat{T}_R$, is negligible (asymptotically unbiased) and performs better than $\hat{t}_{y\pi}$ (exactly unbiased) in the sense of relative efficiency. $\hat{T}_{R(1)}$ and $\hat{T}_{R(2)}$ estimators perform better than the ratio estimator $\hat{T}_R$ for all $n = 12, 18, 24$. At the same time, the relative efficiency of $\hat{T}_{R(1)}$ and $\hat{T}_{R(2)}$ are approximately the same for $n = 12, 18, 24$. In the case of the Horvtiz-Thompson family, the deduced estimators $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ perform significantly better than the original estimator $\hat{t}_{y\pi}$, in the sense of relative efficiency. Furthermore, the estimators $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ deliver approximately the same performance, for all $n = 12, 18, 24$.

From Eq.(51), we have $\widetilde{t}_{y\pi(2)} = \hat{T}_{R(2)}$; therefore, the ratio family and the Horvitz-Thompson family can be compared. Tables (2), (3),..., (9) show that

$$\frac{MSE\left(\widetilde{t}_{y\pi(1)}\right)}{MSE\left(\widetilde{t}_{y\pi(2)}\right)} \cong \frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_{R(2)}\right)},$$

for all values of n. Therefore, the deduced estimators $\widetilde{t}_{y\pi(1)}$ and $\widetilde{t}_{y\pi(2)}$ from the Horvitz-Thompson family and $\hat{T}_{R(1)}$ and $\hat{T}_{R(2)}$ from the ratio family perform better than the original families even though the original families are unbiased or asymptotically unbiased estimators.

Acknowledgement
The authors are grateful to the referees for their valuable comments and suggestions. Also, our thanks are extended to the editorial office of Statistics in Transition new series for their cooperation.

		$\hat{t}_{y.sin}$	$\widetilde{t}_{y\pi}$	$\hat{t}_{y.new}$			$\hat{t}_{y.sin}$	$\widetilde{t}_{y\pi}$	$\hat{t}_{y.new}$
$n=25$	Em.Mean	1720.34	1770.59	1770.59	$n=55$	Em.Mean	1722.70	1768.77	1768.77
$\lambda \to 0$	RB	-0.24	2.68	2.68	$\lambda \to 0$	RB	-0.10	2.57	2.57
	REMSE	0.21	1.00	1.00		REMSE	0.17	1.00	1.00
$n=25$	Em.Mean	1720.34	1770.59	1758.03	$n=55$	Em.Mean	1722.70	1768.77	1757.25
$\lambda = 0.25$	RB	-0.24	2.68	1.95	$\lambda = 0.25$	RB	-0.10	2.57	1.90
	REMSE	0.32	1.54	1.00		REMSE	0.27	1.58	1.00
$n=25$	Em.Mean	1720.34	1770.59	1745.47	$n=55$	Em.Mean	1722.70	1768.77	1745.73
$\lambda = 0.50$	RB	-0.24	2.68	1.22	$\lambda = 0.50$	RB	-0.10	2.57	1.23
	REMSE	0.52	2.51	1.00		REMSE	0.45	2.68	1.00
$n=25$	Em.Mean	1720.34	1770.59	1732.90	$n=55$	Em.Mean	1722.70	1768.77	1734.21
$\lambda = 0.75$	RB	-0.24	2.68	0.49	$\lambda = 0.75$	RB	-0.10	2.57	0.57
	REMSE	0.83	3.98	1.00		REMSE	0.78	4.58	1.00
$n=25$	Em.Mean	1720.34	1770.59	1720.34	$n=55$	Em.Mean	1722.70	1768.77	1722.70
$\lambda \to 1$	RB	-0.24	2.68	-0.24	$\lambda \to 1$	RB	-0.10	2.57	-0.10
	REMSE	1.00	4.81	1.00		REMSE	1.00	5.89	1.00
$n=35$	Em.Mean	1722.08	1770.28	1770.28	$n=65$	Em.Mean	1722.93	1768.89	1768.89
$\lambda \to 0$	RB	-0.14	2.66	2.66	$\lambda \to 0$	RB	-0.09	2.58	2.58
	REMSE	0.19	1.00	1.00		REMSE	0.16	1.00	1.00
$n=35$	Em.Mean	1722.08	1770.28	1758.23	$n=65$	Em.Mean	1722.93	1768.89	1757.40
$\lambda = 0.25$	RB	-0.14	2.66	1.96	$\lambda = 0.25$	RB	-0.09	2.58	1.91
	REMSE	0.30	1.55	1.00		REMSE	0.26	1.57	1.00
$n=35$	Em.Mean	1722.08	1770.28	1746.18	$n=65$	Em.Mean	1722.93	1768.89	1745.91
$\lambda = 0.50$	RB	-0.14	2.66	1.26	$\lambda = 0.50$	RB	-0.09	2.58	1.24
	REMSE	0.49	2.56	1.00		REMSE	0.44	2.67	1.00
$n=35$	Em.Mean	1722.08	1770.28	1734.13	$n=65$	Em.Mean	1722.93	1768.89	1734.42
$\lambda = 0.75$	RB	-0.14	2.66	0.56	$\lambda = 0.75$	RB	-0.09	2.58	0.58
	REMSE	0.80	4.17	1.00		REMSE	0.76	4.59	1.00
$n=35$	Em.Mean	1722.08	1770.28	1722.08	$n=65$	Em.Mean	1722.93	1768.89	1722.93
$\lambda \to 1$	RB	-0.14	2.66	-0.14	$\lambda \to 1$	RB	-0.09	2.58	-0.09
	REMSE	1.00	5.20	1.00		REMSE	1.00	6.07	1.00
$n=45$	Em.Mean	1721.84	1769.24	1769.24	$n=75$	Em.Mean	1723.25	1770.39	1770.39
$\lambda \to 0$	RB	-0.15	2.60	2.60	$\lambda \to 0$	RB	-0.07	2.66	2.66
	REMSE	0.19	1.00	1.00		REMSE	0.15	1.00	1.00
$n=45$	Em.Mean	1721.84	1769.24	1757.39	$n=75$	Em.Mean	1723.25	1770.39	1758.61
$\lambda = 0.25$	RB	-0.15	2.60	1.91	$\lambda = 0.25$	RB	-0.07	2.66	1.98
	REMSE	0.29	1.55	1.00		REMSE	0.25	1.59	1.00
$n=45$	Em.Mean	1721.84	1769.24	1745.54	$n=75$	Em.Mean	1723.25	1770.39	1746.82
$\lambda = 0.50$	RB	-0.15	2.60	1.22	$\lambda = 0.50$	RB	-0.07	2.66	1.30
	REMSE	0.48	2.57	1.00		REMSE	0.42	2.73	1.00
$n=45$	Em.Mean	1721.84	1769.24	1733.69	$n=75$	Em.Mean	1723.25	1770.39	1735.04
$\lambda = 0.75$	RB	-0.15	2.60	0.54	$\lambda = 0.75$	RB	-0.07	2.66	0.61
	REMSE	0.78	4.22	1.00		REMSE	0.74	4.81	1.00
$n=45$	Em.Mean	1721.84	1769.24	1721.84	$n=75$	Em.Mean	1723.25	1770.39	1723.25
$\lambda \to 1$	RB	-0.15	2.60	-0.15	$\lambda \to 1$	RB	-0.07	2.66	-0.07
	REMSE	1.00	5.39	1.00		REMSE	1.00	6.47	1.00

Table 1: Computations are based on Eq.(28). The REMSE's are computed by using Eq. (32) for $\hat{t}_y^* = \hat{t}_{y.sin}, \widetilde{t}_{y\pi}$, and $\hat{t}_{y.new}$.

Estimator	$n=2$		$n=3$		$n=4$	
	t_y	S_y^2	t_y	S_y^2	t_y	S_y^2
$\hat{T}_R$	294.3058	2545.752	292.291	1549.1841	291.322	1081.5572
	(5.3059)	(**2573.904**)	(3.292)	(**1560.0213**)	(2.322)	(**1086.9488**)
$\hat{T}_{R(1)}$	289	1684.9349	289	1111.1305	289	834.9348
$\hat{T}_{R(2)}$	289	1650.5674	289	1087.7975	289	821.8695
$\hat{t}_{y\pi}$	289	2613.375	289	1666.5	289	1193.0625
$\tilde{t}_{y\pi(1)}$	289	1689.2317	289	1115.3438	289	845.6788
$\tilde{t}_{y\pi(2)}$	289	1650.5674	289	1087.7975	289	821.8695
$\hat{T}_R$ Family	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.6546$ $\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.6413$ $\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 1.0208$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.7123$ $\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.6973$ $\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 1.0215$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.7682$ $\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.7561$ $\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 1.0159$	
Horvtiz-Thopson family	$\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.6464$ $\frac{MSE(\tilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.6316$ $\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\tilde{t}_{y\pi(2)})} = 1.0234$		$\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.6693$ $\frac{MSE(\tilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.6527$ $\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\tilde{t}_{y\pi(2)})} = 1.0253$		$\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.7088$ $\frac{MSE(\tilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.6889$ $\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\tilde{t}_{y\pi(2)})} = 1.029$	

Table 2: Empirical results based on real data set. For the estimator $\hat{T}_R$: the number between brackets under the mean is the bias and the bold one under the variance is the MSE of $\hat{T}_R$.

Estimator	$n_h=2$		$n_h=3$		$n_h=4$	
	t_{yh}	S_{yh}^2	t_{yh}	S_{yh}^2	t_{yh}	S_{yh}^2
$\hat{T}_R$	715.7886	1533.6014	714.4398	975.7520	713.7416	691.1576
	(3.7886)	(**1547.9549**)	(2.4398)	(**981.7046**)	(1.741556)	(**694.1906**)
$\hat{T}_{R(1)}$	712	1363.2611	712	522.6314	712	358.48553
$\hat{T}_{R(2)}$	712	1405.6602	712	523.4136	712	353.7434
$\hat{t}_{y\pi}$	712	5900.2105	712	3714.9474	712	2622.3158
$\tilde{t}_{y\pi(1)}$	712	1483.1003	712	578.5657	712	391.3314
$\tilde{t}_{y\pi(2)}$	712	1405.6602	712	523.4136	712	353.7434
$\hat{T}_R$ Family	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.8807$ $\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.9081$ $\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9698$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.5324$ $\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.5332$ $\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9985$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.5164$ $\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.5096$ $\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 1.0134$	
Hort-Thom family	$\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.2514$ $\frac{MSE(\tilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.2382$ $\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\tilde{t}_{y\pi(2)})} = 1.0551$		$\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.1557$ $\frac{MSE(\tilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.1409$ $\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\tilde{t}_{y\pi(2)})} = 1.10537$		$\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.1492$ $\frac{MSE(\tilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.1349$ $\frac{MSE(\tilde{t}_{y\pi(1)})}{MSE(\tilde{t}_{y\pi(2)})} = 1.1063$	

Table 3: Empirical results from cars93 for Stratum (1): When typecode=1

Estimator	$n_h = 2$		$n_h = 3$		$n_h = 4$	
	t_{yh}	S^2_{yh}	t_{yh}	S^2_{yh}	t_{yh}	S^2_{yh}
$\hat{T}_R$	478.0681	306.3392	478.0383	188.7059	478.0252	130.3291
	(0.0681)	(**306.3439**)	(0.0383)	(**188.7074**)	(0.0252)	(**130.3297**)
$\hat{T}_{R(1)}$	478	443.4048	478	219.6018	478	139.9649
$\hat{T}_{R(2)}$	478	444.9974	478	220.1128	478	140.1862
$\hat{t}_{y\pi}$	478	968.8000	478	599.7333	478	415.2000
$\widetilde{t}_{y\pi(1)}$	478	446.9986	478	221.5335	478	141.1973
$\widetilde{t}_{y\pi(2)}$	478	444.9974	478	220.1128	478	140.1862
$\hat{T}_R$ Family	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 1.4474$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 1.1637$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 1.0739$	
	$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 1.4526$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 1.1664$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 1.0756$	
	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9964$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9977$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9984$	
Hort-Thom family	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.4614$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.3694$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.3401$	
	$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.4593$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.3671$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.3376$	
	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0045$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0065$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0072$	

Table 4: Empirical results from cars93 for Stratum (2): When typecode=2

Estimator	$n_h = 2$		$n_h = 3$		$n_h = 4$	
	t_{yh}	S^2_{yh}	t_{yh}	S^2_{yh}	t_{yh}	S^2_{yh}
$\hat{T}_R$	588.4592	449.2140	588.2849	277.1965	588.2004	194.4815
	(0.4592)	(**449.4248**)	(0.2849)	(**277.2777**)	(0.2004)	(**194.5216**)
$\hat{T}_{R(1)}$	588	525.4486	588	260.7300	588	174.2959
$\hat{T}_{R(2)}$	588	527.9447	588	261.3186	588	174.4998
$\hat{t}_{y\pi}$	588	1386.6667	588	878.2222	588	624.0000
$\widetilde{t}_{y\pi(1)}$	588	532.0924	588	264.2138	588	176.5988
$\widetilde{t}_{y\pi(2)}$	588	527.9447	588	261.3186	588	174.4998
$\hat{T}_R$ Family	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 1.1692$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.9403$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.8960$	
	$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 1.1747$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.9424$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.8971$	
	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9953$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9978$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9988$	
Hort-Thom family	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.3837$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.3009$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.2830$	
	$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.3807$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.2976$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.2797$	
	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0079$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0111$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0120$	

Table 5: Empirical results from cars93 for Stratum (3): When typecode=3

Estimator	$n_h = 2$		$n_h = 3$		$n_h = 4$	
	t_{yh}	s^2_{yh}	t_{yh}	s^2_{yh}	t_{yh}	s^2_{yh}
$\hat{T}_R$	294.4412	99.6516	294.2596	58.9851	294.1697	38.6799
	(0.4412)	(**99.8463**)	(0.2596)	(**59.0524**)	(0.1697)	(**38.7087**)
$\hat{T}_{R(1)}$	294	32.2542	294	28.9585	294	24.42012
$\hat{T}_{R(2)}$	294	31.9504	294	28.8365	294	24.3733
$\hat{t}_{y\pi}$	294	80.1000	294	47.4667	294	31.1500
$\tilde{t}_{y\pi(1)}$	294	32.0873	294	29.0126	294	24.5711
$\tilde{t}_{y\pi(2)}$	294	31.9504	294	28.8365	294	24.3733
$\hat{T}_R$ Family	$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_R\right)} = 0.3230$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_R\right)} = 0.4904$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_R\right)} = 0.6309$	
	$\frac{MSE\left(\hat{T}_{R(2)}\right)}{MSE\left(\hat{T}_R\right)} = 0.3199$		$\frac{MSE\left(\hat{T}_{R(2)}\right)}{MSE\left(\hat{T}_R\right)} = 0.4883$		$\frac{MSE\left(\hat{T}_{R(2)}\right)}{MSE\left(\hat{T}_R\right)} = 0.6297$	
	$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_{R(2)}\right)} = 1.0095$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_{R(2)}\right)} = 1.0042$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_{R(2)}\right)} = 1.0019$	
Hort-Thom	$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.4006$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.6112$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.7888$	
family	$\frac{MSE\left(\tilde{t}_{y\pi(2)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.3989$		$\frac{MSE\left(\tilde{t}_{y\pi(2)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.6075$		$\frac{MSE\left(\tilde{t}_{y\pi(2)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.7825$	
	$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\tilde{t}_{y\pi(2)}\right)} = 1.0043$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\tilde{t}_{y\pi(2)}\right)} = 1.0061$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\tilde{t}_{y\pi(2)}\right)} = 1.0081$	

Table 6: Empirical results from cars93 for Stratum (4): When typecode=4

Estimator	$n_h = 2$		$n_h = 3$		$n_h = 4$	
	t_{yh}	s^2_{yh}	t_{yh}	s^2_{yh}	t_{yh}	s^2_{yh}
$\hat{T}_R$	404.9682	464.2094	404.2003	279.8831	403.8170	189.1534
	(1.9682)	(**468.0833**)	(1.2003)	(**281.3238**)	(0.8170)	(**189.8209**)
$\hat{T}_{R(1)}$	403	219.7774	403	117.0185	403	96.4114
$\hat{T}_{R(2)}$	403	221.1506	403	114.8667	403	94.8695
$\hat{t}_{y\pi}$	403	1113.6923	403	680.5897	403	464.0385
$\tilde{t}_{y\pi(1)}$	403	234.9529	403	123.6658	403	100.9741
$\tilde{t}_{y\pi(2)}$	403	221.1506	403	114.8667	403	94.8695
$\hat{T}_R$ Family	$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_R\right)} = 0.4695$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_R\right)} = 0.4160$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_R\right)} = 0.5079$	
	$\frac{MSE\left(\hat{T}_{R(2)}\right)}{MSE\left(\hat{T}_R\right)} = 0.4725$		$\frac{MSE\left(\hat{T}_{R(2)}\right)}{MSE\left(\hat{T}_R\right)} = 0.4083$		$\frac{MSE\left(\hat{T}_{R(2)}\right)}{MSE\left(\hat{T}_R\right)} = 0.4998$	
	$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_{R(2)}\right)} = 0.9938$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_{R(2)}\right)} = 1.0187$		$\frac{MSE\left(\hat{T}_{R(1)}\right)}{MSE\left(\hat{T}_{R(2)}\right)} = 1.0163$	
Hort-Thom	$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.2110$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.1817$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.2176$	
family	$\frac{MSE\left(\tilde{t}_{y\pi(2)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.1986$		$\frac{MSE\left(\tilde{t}_{y\pi(2)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.1688$		$\frac{MSE\left(\tilde{t}_{y\pi(2)}\right)}{MSE\left(\hat{t}_{y\pi}\right)} = 0.2044$	
	$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\tilde{t}_{y\pi(2)}\right)} = 1.0624$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\tilde{t}_{y\pi(2)}\right)} = 1.0766$		$\frac{MSE\left(\tilde{t}_{y\pi(1)}\right)}{MSE\left(\tilde{t}_{y\pi(2)}\right)} = 1.0644$	

Table 7: Empirical results from cars93 for Stratum (5): When typecode=5

Estimator	$n_h = 2$		$n_h = 3$		$n_h = 4$	
	t_{yh}	S^2_{yh}	t_{yh}	S^2_{yh}	t_{yh}	S^2_{yh}
$\hat{T}_R$	197.0919	23.2645	197.0515	13.2744	197.0319	8.2891
	(0.0919)	(**23.2730**)	(0.0515)	(**13.2770**)	(0.0319)	(**8.2901**)
$\hat{T}_{R(1)}$	197	22.6543	197	11.1726	197	7.1120
$\hat{T}_{R(2)}$	197	22.6484	197	11.1820	197	7.1201
$\hat{t}_{y\pi}$	197	66.5000	197	38.0000	197	23.75
$\widetilde{t}_{y\pi(1)}$	197	22.6953	197	11.2614	197	7.2046
$\widetilde{t}_{y\pi(2)}$	197	22.6484	197	11.1820	197	7.1201
$\hat{T}_R$ Family	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.9734$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.8415$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.8579$	
	$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.9732$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.8422$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.8589$	
	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 1.0003$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9992$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9989$	
Hort-Thom family	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.3413$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.2964$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.3034$	
	$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.3406$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.2943$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.2998$	
	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0021$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0071$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0119$	

Table 8: Empirical results from cars93 for Stratum (6): When typecode=6

Estimator	$n = 12$		$n = 18$		$n = 24$	
	t_y	S^2_y	t_y	S^2_y	t_y	S^2_y
$\hat{T}_R$	2678.8172	453844.48	2676.2744	178628.06	2674.9857	88260.276
	(6.8171677)	(**453863.13**)	(4.2743444)	(**178635.6**)	(2.9856758)	(**88264.047**)
$\hat{T}_{R(1)}$	2672	431418.52	2672	117843.17	2672	56576.944
$\hat{T}_{R(2)}$	2672	439877.98	2672	117935.4	2672	56173.604
$\hat{t}_{y\pi}$	2672	1575220	2672	621987.81	2672	308599.03
$\widetilde{t}_{y\pi(1)}$	2672	456121.5	2672	125146.15	2672	59655.399
$\widetilde{t}_{y\pi(2)}$	2672	439877.98	2672	117935.4	2672	56173.604
$\hat{T}_R$ Family	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.9506$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.6597$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_R)} = 0.6410$	
	$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.9692$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.6602$		$\frac{MSE(\hat{T}_{R(2)})}{MSE(\hat{T}_R)} = 0.6364$	
	$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9808$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 0.9992$		$\frac{MSE(\hat{T}_{R(1)})}{MSE(\hat{T}_{R(2)})} = 1.0072$	
Hort-Thom family	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.2896$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.2012$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\hat{t}_{y\pi})} = 0.1933$	
	$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.2793$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.1896$		$\frac{MSE(\widetilde{t}_{y\pi(2)})}{MSE(\hat{t}_{y\pi})} = 0.1820$	
	$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0369$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0611$		$\frac{MSE(\widetilde{t}_{y\pi(1)})}{MSE(\widetilde{t}_{y\pi(2)})} = 1.0620$	

Table 9: Empirical results from cars93 based on stratified sampling designs

REFERENCES

AL-JARARHA, J., (2015). A Dual Problem of Calibration of Design Weights Based on Multi-Auxiliary Variables, Communications for Statistical Applications and Methods, 22(2), pp. 137–146.

AL-YASEEN, A., (2014). Penalized Chi-Square Distance and the Dual Calibration for Estimating the Finite Population Total, Master Thesis. Statistics Deperatment. Yarmouk University, Jordan.

DEVILLE, J.-C., SÄRNDAL, C.-E., (1992). Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, 87, pp. 376–382.

GODAMBE, V. P., (1955). A Unified Theory of Sampling from Finite Populations, J. Roy. Statist. Soc., B17, pp. 269–278.

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

LOHR, S. L. (2010). Sampling: Design and Analysis (2nd ed.), Boston: Brooks/Cole, Cengage Learning.

NIDHI, B. V. S., SISODIA, S. SINGH, SINGH S. K., (2017). Calibration approach estimation of the mean in stratified sampling and stratified double sampling, Communications in Statistics - Theory and Methods, 46(10), pp. 4932-4942.

OZGUL, N., (2018). New calibration estimator based on two auxiliary variables in stratified sampling. Communications in Statistics - Theory and Methods, doi = 10.1080/03610926.2018.1433852, pp. 1–12.

SCHEAFFER, R. L., MENDENHALL, W., OTT, R. L. (2006). Elementary Survey Sampling (6th ed.), Belmont, CA: Duxbury.

SINGH, S., (2013). A Dual Problem of Calibration of Design Weights, Statistics: A Journal of Theoretical and Applied Statistics, 47(3), pp. 566–574.

STEARNS, M., S. SINGH, (2008). On the estimation of the general parameter, Computational Statistics Data Analysis, 52, pp. 4253–4271.

SUGDEN, R., SMITH T., (2002). Exact linear unbiased estimation in survey sampling, Journal of Statistical Planning and Inference, 102 (1), pp. 25–38.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1, pp. 55–71, DOI 10.21307/stattrans-2020-004
Submitted – 31.07.2019; accepted for publishing – 09.01.2020

Predicting parity progression ratios for young women by the end of their childbearing life

Agnieszka Rossa[1], **Agnieszka Palma**[2]

ABSTRACT

Parity progression ratios (PPR's) have been extensively described in literature on demography and have played an important role in fertility, unlike the idea of calculating projected parity progression ratios proposed by Brass (1985). However, we decided to use this method in our paper to analyse future fertility trends, firstly by assessing age-specific parity progression ratios for women in childbearing ages, and then by comparing these ratios with ratios at the end of women's reproductive life, as well as by comparing the latter with the completed PPR's. More specifically, the aim of this study is to adopt a modified Brass method to calculate the projected parity progression ratios using the age-period fertility data sourced from the Human Fertility Database (HFD). We progress to use the observed and predicted age-specific PPR's to examine parity progressions in Poland as a case study.

Key words: fertility rates, parity, projected parity progression ratios.

1. Introduction

The long-term decline in cohort fertility rates across developed countries has been widely studied and documented (Frejka and Calot 2001, Kohler et al. 2002, Billiari and Kohler 2004, Frejka 2008, Myrskyla et al. 2013, Sobotka 2013).

Empirical findings and evidence from the EU countries over last decades fit to the dominant demographic theories such as demographic transition and second demographic transition postulating that as societies progress, fertility tends to decrease. The period total fertility rate TFR declined considerably between 1980 and 2003 in most of the EU countries reaching the level below 1.30 between 2000 and 2003. According to Kohler et al. (2002) such low levels of TFR are termed "lowest-low" fertility. During the 1990s there were several lowest-low fertility countries in Southern, Central, South-Eastern and Eastern Europe, e.g. in Bulgaria, Czechia, Greece, Spain, Italy, Latvia, Slovenia, Slovakia. In several European countries fertility started to increase gradually around 2005, and decrease again with the financial crisis in 2008. More recently, according to the annual Eurostat reports, the period EU-wide total fertility rate attained 1.59 live births per woman in 2017, ranging from 1.26 in Malta to 1.90 in France. Moreover, almost half of children born in the EU in 2017 were first-born children.

[1]Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Lodz. E-mail: agnieszka.rossa@uni.lodz.pl. ORCID: https://orcid.org/0000-0002-0444-4181

[2]Institute of Statistics and Demography, Faculty of Economics and Sociology, University of Lodz. E-mail: agnieszka.palma@uni.lodz.pl

Several social and economic factors can serve as a response to observed fertility patterns and spatial differences, e.g. economic uncertainty, recession, increased incentives to invest in higher education and labour market experience, new lifestyle opportunities, reproductive behaviour, contraceptive use, abortion availability, or even late home-leaving by young adults, which is strongly correlated with high costs of formation of separate households.

Studying parity progression ratios can deliver more interesting details for understanding fertility changes and differences in parity distributions (Henry 1980, Paradysz 1995, Preston et al. 2001). Parity at a given point in time is defined as the number of children ever born by a women and the Parity Progression Ratio (PPR) of an i-th order reflects the proportion of women with i children who continue to an $i+1$-th live birth during their reproductive life. Thus, parity progression ratios allow to assess how frequently women are moving from the lower to higher parity.

Changes in particular PPR's may provide insight into processes of fertility with respect to the propensity of women to have children. Frejka (2008) found that decreasing PPR's to first and second births played a key role in fertility declines among European women born after 1955. In the Central and Eastern Europe fertility decline was driven primarily by falling PPR's to second births. Kohler et al. (2002) and Billiari and Kohler (2004) suggested that a pattern of the lowest-low fertility in Europe during the 1990s is characterized by a delay of childbearing especially for first births as well as by low progression probability after the first child but not by low probability of the first childbearing.

Usually, parity projection ratios are calculated for cohorts of women who have finished their reproductive life. A particular PPR of an i-th order is then defined as the ratio of the number of women at parity $i+1$ or more to the number of women at parity i or more and is treated as a fixed and completed cohort measure. PPR's for younger women are also calculated but are considered as uncompleted age-specific parity measures since women in reproductive ages move to higher parities and the distribution of their parities is changing. Brass (1985) proposed a methods which enables one to use parity data on younger women to calculate the so-called Projected Parity Progression Ratios (PPPR's) considered as completed progression ratios expected to be achieved by younger women by the end of their reproductive life. The method is based on the assumption that the current age pattern of specific fertility remains constant at the level observed at a given point in time.

In the paper a modified formula of the projected parity progression ratio is applied to investigate changes in the parity distribution in Poland as a case study. Some *ex post* comparisons are also conducted, i.e. between the observed and predicted PPR's for women in various age groups and between the latter and the completed PPR's observed for women attaining age 49 in a particular calendar year. Findings formulated from the comparisons allow one to assess the prediction accuracy of the modified Brass method as well as to make a contribution to explaining the future change in parity distribution over ten-year time horizon. The procedure is illustrated in details on the age-period fertility data of Poland sourced from the Human Fertility Database.

2. Notation and assumptions

In our analysis we use the following notation adopted to the population of women and to live births in a particular year:

N – total exposure-to-risk,

B – the total number of births,

N_x – exposure-to-risk in age interval $[x, x+1)$ for women attaining age x,

B_x – the number of births delivered by women aged x (in completed years),

$N_x(i)$ – exposure-to-risk in age interval $[x, x+1)$ for women attaining age x and of parity i,

$B_x(i)$ – the number of births delivered by women aged x and of parity i,

$N(i)$ – total exposure-to-risk for women of an i-th parity,

$B(i)$ – the total number of births to women of an i-th parity.

The following relations hold

$$N_x = \sum_{i=0}^{\pi} N_x(i), \quad N(i) = \sum_{x=\alpha}^{\beta} N_x(i), \quad N = \sum_{i=1}^{\pi} N(i) = \sum_{x=\alpha}^{\beta} N_x, \tag{1}$$

and

$$B_x = \sum_{i=0}^{\pi} B_x(i), \quad B(i) = \sum_{x=\alpha}^{\beta} B_x(i), \quad B = \sum_{i=1}^{\pi} B(i) = \sum_{x=\alpha}^{\beta} B_x, \tag{2}$$

where π is the highest parity in the data set, and α, β define the limits of the reproductive age range $[\alpha, \beta+1)$. Further, we will assume $\alpha = 15$ and $\beta = 49$.

We will also assume that in the given calendar year women had at most one birth, i.e. there are neither multiple deliveries nor multiple confinements, and that age-specific fertility rates for the reference period will continue to characterize future fertility patterns.

Note, that numbers of women P_x, P_{x+1} attaining respective ages x and $x+1$ during the reference year are closely related to exposure-to-risk N_x. Let us assume that the birthdays of females are distributed uniformly within the calendar year. Then each of the P_x females

contributes on average $\frac{1}{2}$ of person-years to exposure N_x. Similarly, each of the P_{x+1} females contributes on average $\frac{1}{2}$ of person-years to N_x. On the other hand, assuming uniform distribution of deaths within the year, each of the D_x^L deaths (i.e. deaths in the lower triangle according to the Lexis diagram) among P_x females reduces exposure-to-risk by $\frac{1}{3}$ of person-years, on average, while each of the D_x^U deaths (i.e. deaths in the upper Lexis triangle) contributes an average $\frac{1}{3}$ of person-years to exposure N_x. Thus, N_x can be written as

$$N_x \approx \frac{1}{2}\left(P_x + P_{x+1}\right) + \frac{1}{3}\left(D_x^U - D_x^L\right). \tag{3}$$

Assuming $D_x^U \approx D_x^L$, expression (3) comes down to

$$N_x \approx \frac{1}{2}\left(P_x + P_{x+1}\right). \tag{4}$$

Analogous approximate equality refers to $N_x(i)$, i.e.

$$N_x(i) \approx \frac{1}{2}\left(P_x(i) + P_{x+1}(i)\right). \tag{5}$$

Further, exposure-to-risk N_x will be treated as an approximate average number of women aged x (in completed years) and similarly $N_x(i)$ – as an approximate average number of women aged x and of parity i.

3. Period Specific Fertility Rates and Average Parity

The Age-Specific Fertility Rate (ASFR) and the Age-Specific i-th Order Fertility Rate (ASOFR) for women aged x are defined as follows

$$ASFR_x = \frac{B_x}{N_x} \qquad \text{and} \qquad ASOFR_x(i) = \frac{B_x(i)}{N_x}. \tag{6}$$

Note that $ASOFR_x(i)$ cannot be termed "order-specific rate" as the denominator N_x is unidentified by parity. Observe also that

$$ASFR_x = \sum_{i=1}^{\pi} ASOFR_x(i). \tag{7}$$

Average parity P in a population is calculated by dividing the total number of children ever born by the number of women N, i.e.

$$P = \frac{1}{N} \cdot \sum_{j=0}^{\pi} j \cdot N(j) = \frac{1 \cdot N(1)}{N} + \frac{2 \cdot N(2)}{N} + \frac{3 \cdot N(3)}{N} + \ldots + \frac{\pi \cdot N(\pi)}{N}. \tag{8}$$

4. Cumulated Age-Specific *i*-th Order Fertility Rates

In this section we will employ age-specific and age-specific *i*-th order fertility rates given in (6) for one-year age bands $[x, x+1)$ to define total and total order fertility rates as well as cumulated age-order fertility rates.

Using the α, β as the limits for the summation, the Total Fertility Rate (TFR) and the Total *i*-th Order Fertility Rate (TOFR) are defined as

$$TFR = \sum_{x=\alpha}^{\beta} ASFR_x \qquad \text{and} \qquad TOFR(i) = \sum_{x=\alpha}^{\beta} ASOFR_x(i). \tag{9}$$

Then the Cumulated Age-Specific *i*-th Order Fertility Rate is determined by summing age-specific *i*-th order specific fertility rates up to the desired age x. Thus, we have

$$TOFR_y(i) = \sum_{x=\alpha}^{y} ASOFR_x(i). \tag{10}$$

It follows from (9) and (10) that $TOFR_y(i) = TOFR(i)$ for $y = \beta$.

5. Conventional and Projected Parity Progression Ratios

The concept of a parity progression ratio was introduced by Henry in 1953 as a useful measure of fertility. Later, many researchers have proposed methods to evaluate parity progression ratios (PPR's) or the projected parity progression ratios (PPPR's) (Srinivasan 1968, Feeney 1983, Yadava and Bhattacharya 1985, Brass 1985, Feeney and Jingyuan 1987, Yadava et al. 1992, Islam and Yadava 1997, Bhardwaj et al. 2010, Yadava and Kumar 2011).

The conventional PPR of an *i*-th order is the proportion of women who progress from *i*-th to $i+1$-th parity. In other words, it is the chance that a female after giving birth to *i*-th child will ever deliver another child. Projected parity progression ratios indicate the possible future evolution of parity progression for younger women, taking into account both current fertility and the women's childbearing history.

Parity progression ratios can be calculated on a cohort or period basis depending on the data available. For cohorts, they are usually calculated for women who have completed their childbearing, e.g. for women aged 49. Cohort ratios are often calculated from the census data whereas period ratios use probabilities of giving birth in a defined reference period. In our analysis we use the age-period fertility data sourced from the Human Fertility Database to calculate PPR's and generalized PPPR's for the female population in Poland (see Section 6).

5.1. Parity Progression Ratio

Let us consider the number $W(i)$ of women in a population having attained parity i or higher. Note that

$$W(i) = \sum_{j=i}^{\pi} N(j) = N(i) + N(i+1) + \cdots + N(\pi). \tag{11}$$

It is clear that $W(0)$ is equal to the total number of women in the population

$$W(0) = \sum_{j=0}^{\pi} N(j) = N. \tag{12}$$

The proportion $M(i)$ of women ever-attaining parity i, i.e. the share of women who have at least i children, can be expressed as

$$M(i) = \frac{W(i)}{N} = \frac{1}{N} \cdot \sum_{j=i}^{\pi} N(j). \tag{13}$$

The corresponding proportion at parity zero or higher is $M(0) = N/N = 1$.

By analogy, the number $W_x(i)$ and the proportion $M_x(i)$ of women aged x and having attained parity i or higher are as follows

$$W_x(i) = \sum_{j=i}^{\pi} N_x(j), \tag{14}$$

$$M_x(i) = \frac{W_x(i)}{N_x} = \frac{1}{N_x} \cdot \sum_{j=i}^{\pi} N_x(j). \tag{15}$$

Then, the associated parity progression ratios $PPR(i)$ and $PPR_x(i)$ can be expressed as

$$PPR(i) = \frac{W(i+1)}{W(i)} = \frac{W(i+1)/N}{W(i)/N} = \frac{M(i+1)}{M(i)}, \tag{16}$$

$$PPR_x(i) = \frac{W_x(i+1)}{W_x(i)} = \frac{W_x(i+1)/N_x}{W_x(i)/N_x} = \frac{M_x(i+1)}{M_x(i)}. \tag{17}$$

It is worth noting that ratios $PPR_x(i)$ calculated for younger women should be treated as uncompleted age-specific PPR's. In such cases it is also reasonable to calculate projected parity progression ratios $PPPR_x(i)$ in order to estimate completed parity progressions for younger women by the end of their reproductive life.

5.2. Projected Parity Progression Ratio and its generalization

According to the Brass concept (see, e.g. Moultrie et al. 2013, p. 74) the difference between the Total Order Fertility Rate and the Cumulated Age-Specific i-th Order Fertility

Rate,

$$TOFR(i) - TOFR_x(i), \quad i = 1,2,\ldots,\pi, \tag{18}$$

can be treated as an estimate of an additional proportion of women aged x expected to achieve parity i by the end of their childbearing years. This interpretation is admissible under the assumption that the current fertility pattern will remain constant until the end of women's reproductive life and that in the given year every women had at most one birth.

Let us consider a modified version of formula (18) by substituting $TOFR_y(i)$ for $TOFR(i)$ in (18), where $y > x$. Under assumptions as stated above, the difference of the form

$$TOFR_y(i) - TOFR_x(i), \quad i = 1,2,\ldots,\pi, \quad y > x, \tag{19}$$

can be treated as an estimate of an additional proportion of women aged x expected to achieve parity i at age y. Note that formula (19) reduces to (18) when $y = \beta$.

Then, the proportions of women aged x projected to achieve at least parity i at age $y > x$ can be defined as

$$M^*_{x,y}(i) = M_x(i) + TOFR_y(i) - TOFR_x(i), \qquad M^*_{x,y}(0) = M_x(0) = 1. \tag{20}$$

For $y = \beta$ we will write $M^*_x(i)$ instead of $M^*_{x,y}(i)$. Thus, in this case we have

$$M^*_x(i) = M_x(i) + TOFR(i) - TOFR_x(i), \tag{21}$$

Generalized projected parity progression ratios for women aged x will be considered as progression ratios expected to be reached after $y - x$ years. They will be expressed as

$$PPPR_{x,y}(i) = \frac{M^*_{x,y}(i+1)}{M^*_{x,y}(i)}. \tag{22}$$

6. Parity distribution in Poland – analysis based on projected parity progression ratios

To examine the measures of fertility presented in previous sections we applied the most recent fertility data on the distribution of the female population in Poland tabulated by one-year age groups and by parity as well as the distribution of births attained to this population and tabulated by birth order and mothers' age (in completed years).

6.1. The input data

The input data contained in Tables 1 and 2 were sourced from the Human Fertility Database. The body of Table 1 shows age-parity exposure of Polish female population in the last available year 2016, whereas Table 2 displays counts of live births in Poland in the same year tabulated by birth order and mothers' age.

Table 1: Age-parity female exposure $N_x(i)$ of the 2016 female population in Poland (average number of females by age and parity)

	$N_x(i)$					
Age x	$i=0$	$i=1$	$i=2$	$i=3$	$i\geq 4$	total
-15	699085.42	165.06	5.45	0	0	699255.93
16	184257.97	625.43	15.58	0	0	184898.98
17	185483.01	1798.32	58.44	1.98	0	187341.75
18	189041.11	4135.90	222.94	5.45	0.49	193405.89
19	194699.88	8154.46	656.40	37.69	1.99	203550.42
20	194719.18	13174.49	1460.10	110.39	6.94	209471.10
21	196404.47	19087.15	2868.78	285.58	27.31	218673.29
22	201459.53	26170.88	4972.21	591.05	73.55	233267.22
23	198931.98	32937.34	7624.92	1034.25	169.31	240697.80
24	197317.30	40905.71	11273.04	1640.15	301.11	251437.31
25	194496.67	50425.01	16150.27	2433.19	503.30	264008.44
26	181595.98	59855.72	22111.05	3501.29	801.03	267865.07
27	167695.53	69161.72	29309.41	4701.49	1159.01	272027.16
28	157336.79	78246.95	38409.45	6152.04	1573.96	281719.19
29	144603.24	84933.58	47992.62	7766.47	1965.86	287261.77
30	137485.75	92972.07	60900.12	10248.90	2636.65	304243.49
31	131856.06	98881.18	75666.38	13357.25	3552.52	323313.39
32	122603.24	97951.45	86973.82	16151.36	4282.29	327962.16
33	113050.33	94418.04	96023.29	18921.63	5066.54	327479.83
34	101354.89	87457.72	99899.98	20814.14	5747.44	315274.17
35	92601.81	83963.21	105242.26	23310.58	6632.15	311750.01
36	85900.89	82755.66	110598.42	26078.02	7683.17	313016.16
37	77920.47	80183.70	112355.60	28034.34	8720.38	307214.49
38	70508.67	78048.62	113088.62	30099.28	9948.63	301693.82
39	65645.77	77014.28	114896.35	32326.78	11178.48	301061.66
40	61221.21	74909.35	114266.77	33816.61	12380.23	296594.17
41	56286.57	70929.06	110469.02	34585.13	13553.56	285823.34
42	52374.76	66343.63	106479.98	35118.83	14795.63	275112.83
43	49362.05	61801.96	102184.14	35642.20	15832.16	264822.51
44	45290.59	57896.05	98669.16	36265.96	16929.75	255051.51
45	41136.72	54218.25	95804.85	36888.67	18122.17	246170.66
46	36632.17	51109.92	93088.02	37827.05	19325.51	237982.67
47	33993.21	48309.58	90251.72	39001.04	20315.62	231871.17
48	32899.15	45636.39	88909.58	39772.78	21526.43	228744.33
49	31780.03	43773.84	89020.08	40891.76	23095.80	228561.51

Source: HUMAN FERTILITY DATABASE. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria), available at www.humanfertility.org (data downloaded on [06/06/2019]).

Table 2: Number of births $B_x(i)$ by mothers' age x and birth order i, Poland, 2016

	$B_x(i)$				
Age x	$i=1$	$i=2$	$i=3$	$i\geq 4$	total
-15	250	5	0	0	255
16	752	14	0	0	766
17	1591	91	3	0	1685
18	3007	263	5	0	3275
19	4642	611	60	5	5318
20	5770	1148	108	10	7036
21	6718	1836	246	32	8832
22	7556	2637	378	65	10636
23	8879	3554	572	123	13128
24	10543	4519	876	205	16143
25	12817	6025	1058	272	20172
26	14392	7331	1313	406	23442
27	14981	8903	1662	511	26058
28	14894	10711	2037	606	28248
29	13735	12074	2384	692	28885
30	12589	13594	2844	879	29906
31	10394	13742	3220	965	28322
32	8476	13249	3568	1008	26301
33	6558	11758	3662	1203	23181
34	4862	9443	3273	1126	18704
35	3823	7630	3114	1236	15804
36	2877	5968	2938	1261	13045
37	2121	4397	2505	1153	10177
38	1547	2981	2043	1085	7655
39	1138	2130	1525	998	5791
40	744	1369	1046	812	3971
41	515	839	714	615	2683
42	275	439	494	439	1647
43	127	235	239	300	901
44	68	98	109	184	460
45	28	45	49	85	207
46	10	15	20	30	75
47	8	12	7	20	47
48	5	2	0	3	10
49	1	4	1	2	8

Source: As in table 1.

6.2. Results

A. As a first step of the analysis, the number $W_x(i)$ and the proportion $M_x(i)$ of women aged x and having attained parity i or higher are calculated. For this purpose, we use formulas (14) and (15). Next, parity progression ratios $PPR_x(i)$ from (17) are computed. Table 3 reveals results concerning $M_x(i)$ and $PPR_x(i)$.

Table 3: Proportions $M_x(i)$ of women aged x attaining parity i and parity progression ratios $PPR_x(i)$

	$M_x(i)$					$PPR_x(i)$			
Age x	$i=0$	$i=1$	$i=2$	$i=3$	$i\geq 4$	$i=0$	$i=1$	$i=2$	$i=3$
-15	1	0.0002	0.0000	0.0000	0.0000	0.0002	0.0000		
16	1	0.0035	0.0001	0.0000	0.0000	0.0035	0.0286	0.0000	
17	1	0.0099	0.0003	0.0000	0.0000	0.0099	0.0303	0.0000	
18	1	0.0226	0.0012	0.0000	0.0000	0.0226	0.0531	0.0000	
19	1	0.0435	0.0034	0.0002	0.0000	0.0435	0.0782	0.0588	0.0000
20	1	0.0704	0.0075	0.0006	0.0000	0.0704	0.1065	0.0800	0.0000
21	1	0.1018	0.0145	0.0014	0.0001	0.1018	0.1424	0.0966	0.0714
22	1	0.1364	0.0242	0.0028	0.0003	0.1364	0.1774	0.1157	0.1071
23	1	0.1735	0.0367	0.0050	0.0007	0.1735	0.2115	0.1362	0.1400
24	1	0.2152	0.0526	0.0077	0.0012	0.2152	0.2444	0.1464	0.1558
25	1	0.2633	0.0723	0.0111	0.0019	0.2633	0.2746	0.1535	0.1712
26	1	0.3221	0.0986	0.0161	0.0030	0.3221	0.3061	0.1633	0.1863
27	1	0.3835	0.1293	0.0215	0.0043	0.3835	0.3372	0.1663	0.2000
28	1	0.4415	0.1638	0.0274	0.0056	0.4415	0.3710	0.1673	0.2044
29	1	0.4966	0.2009	0.0339	0.0068	0.4966	0.4046	0.1687	0.2006
30	1	0.5481	0.2425	0.0424	0.0087	0.5481	0.4424	0.1748	0.2052
31	1	0.5922	0.2863	0.0523	0.0110	0.5922	0.4835	0.1827	0.2103
32	1	0.6262	0.3275	0.0623	0.0131	0.6262	0.5230	0.1902	0.2103
33	1	0.6548	0.3665	0.0733	0.0155	0.6548	0.5597	0.2000	0.2115
34	1	0.6785	0.4011	0.0842	0.0182	0.6785	0.5912	0.2099	0.2162
35	1	0.7030	0.4336	0.0960	0.0213	0.7030	0.6168	0.2214	0.2219
36	1	0.7256	0.4612	0.1079	0.0245	0.7256	0.6356	0.2340	0.2271
37	1	0.7464	0.4854	0.1196	0.0284	0.7464	0.6503	0.2464	0.2375
38	1	0.7663	0.5076	0.1327	0.0330	0.7663	0.6624	0.2614	0.2487
39	1	0.7820	0.5261	0.1445	0.0371	0.7820	0.6728	0.2747	0.2567
40	1	0.7936	0.5410	0.1558	0.0417	0.7936	0.6817	0.2880	0.2677
41	1	0.8031	0.5549	0.1684	0.0474	0.8031	0.6909	0.3035	0.2815
42	1	0.8096	0.5685	0.1814	0.0538	0.8096	0.7022	0.3191	0.2966
43	1	0.8136	0.5802	0.1944	0.0598	0.8136	0.7131	0.3351	0.3076
44	1	0.8224	0.5954	0.2086	0.0664	0.8224	0.7240	0.3504	0.3183
45	1	0.8329	0.6126	0.2235	0.0736	0.8329	0.7355	0.3648	0.3293
46	1	0.8461	0.6313	0.2402	0.0812	0.8461	0.7461	0.3805	0.3381
47	1	0.8534	0.6450	0.2558	0.0876	0.8534	0.7558	0.3966	0.3425
48	1	0.8562	0.6567	0.2680	0.0941	0.8562	0.7670	0.4081	0.3511
49	1	0.8705	0.6945	0.3057	0.1142	0.8705	0.7978	0.4402	0.3735

Source: Authors' own calculations.

Interpretation of the data in Table 1 is rather straightforward. For instance, while 26.33% women aged 25 have had at least one birth ($M_{25}(1) = 0.2633$), only 7.23% have had two or more births ($M_{25}(2) = 0.0723$). On the other hand, the parity progression ratios suggest that 15.35% women aged 25 who had two children went on to have a third ($PPR_{25}(2) = 0.1535$).

B. In the second step the Cumulated Age-Specific i-th Order Fertility Rates and the Total Order Fertility Rates, i.e. $TOFR_x(i)$, $TOFR(i)$ from (10) and (9), respectively, are derived. Based on differences $TOFR(i) - TOFR_x(i)$, additional proportions of women aged x expected to attain parity i by the end of their childbearing years are found. Results are given in Table 4.

Table 4: Cumulated Age-Specific i-th Order Fertility Rates $TOFR_x(i)$ and differences $TOFR(i) - TOFR_x(i)$

	$TOFR_x(i)$				$TOFR(i) - TOFR_x(i)$			
Age x	$i=1$	$i=2$	$i=3$	$i \geq 4$	$i=1$	$i=2$	$i=3$	$i \geq 4$
-15	0.0004	0.0000	0.0000	0.0000	0.6568	0.5048	0.1406	0.0350
16	0.0044	0.0001	0.0000	0.0000	0.6527	0.5047	0.1406	0.0350
17	0.0129	0.0006	0.0000	0.0000	0.6442	0.5042	0.1406	0.0350
18	0.0285	0.0019	0.0000	0.0000	0.6287	0.5029	0.1405	0.0350
19	0.0513	0.0049	0.0003	0.0000	0.6059	0.4999	0.1403	0.0350
20	0.0788	0.0104	0.0009	0.0001	0.5783	0.4944	0.1397	0.0349
21	0.1095	0.0188	0.0020	0.0002	0.5476	0.486	0.1386	0.0348
22	0.1419	0.0301	0.0036	0.0005	0.5152	0.4747	0.1370	0.0345
23	0.1788	0.0449	0.0060	0.0009	0.4783	0.4599	0.1346	0.0341
24	0.2207	0.0628	0.0095	0.0015	0.4364	0.4420	0.1311	0.0335
25	0.2693	0.0857	0.0135	0.0023	0.3878	0.4191	0.1271	0.0327
26	0.3230	0.1130	0.0184	0.0034	0.3341	0.3918	0.1222	0.0316
27	0.3781	0.1458	0.0245	0.0047	0.2790	0.3590	0.1161	0.0303
28	0.4310	0.1838	0.0317	0.0062	0.2262	0.3210	0.1089	0.0288
29	0.4788	0.2258	0.0400	0.0079	0.1784	0.2790	0.1006	0.0271
30	0.5202	0.2705	0.0494	0.0098	0.1370	0.2343	0.0912	0.0252
31	0.5523	0.3130	0.0593	0.0118	0.1048	0.1918	0.0813	0.0232
32	0.5781	0.3534	0.0702	0.0139	0.0790	0.1514	0.0704	0.0211
33	0.5982	0.3893	0.0814	0.0163	0.0590	0.1155	0.0592	0.0187
34	0.6136	0.4193	0.0918	0.0186	0.0435	0.0856	0.0488	0.0164
35	0.6259	0.4437	0.1017	0.0212	0.0313	0.0611	0.0388	0.0138
36	0.6351	0.4628	0.1111	0.0238	0.0221	0.0420	0.0295	0.0112
37	0.6420	0.4771	0.1193	0.0262	0.0152	0.0277	0.0213	0.0088
38	0.6471	0.4870	0.1261	0.0283	0.0101	0.0178	0.0145	0.0067
39	0.6509	0.4941	0.1311	0.0303	0.0063	0.0107	0.0095	0.0047
40	0.6534	0.4987	0.1346	0.0318	0.0038	0.0061	0.0059	0.0032
41	0.6552	0.5016	0.1371	0.0329	0.0020	0.0032	0.0034	0.0021
42	0.6562	0.5032	0.1389	0.0337	0.0010	0.0016	0.0016	0.0012
43	0.6567	0.5041	0.1398	0.0343	0.0005	0.0007	0.0007	0.0006
44	0.6569	0.5045	0.1403	0.0347	0.0002	0.0003	0.0003	0.0003
45	0.6570	0.5047	0.1405	0.0349	0.0001	0.0001	0.0001	0.0001
46	0.6571	0.5047	0.1406	0.0349	0.0001	0.0001	0.0000	0.0000
47	0.6571	0.5048	0.1406	0.0350	0.0000	0.0000	0.0000	0.0000
48	0.6571	0.5048	0.1406	0.0350	0.0000	0.0000	0.0000	0.0000
49	0.6571	0.5048	0.1406	0.0350	0.0000	0.0000	0.0000	0.0000

Source: Authors' own calculations.

For example, the cumulated age-order fertility rate up to the age of 25 for parity $i = 2$ would be

$$TOFR_{25}(2) = 0.0857.$$

The Total Order Fertility Rate for the same parity, $TOFR(i)$, is equal to the cumulated age-order fertility rate up to the end of women's reproductive life, i.e. up to the age of 49. Thus, we have

$$TOFR(2) = TOFR_{49}(2) = 0.5048.$$

It implies that that difference

$$TOFR(2) - TOFR_{25}(2) = 0.4191$$

estimates the additional proportion of women aged 25 expected to achieve parity 2 by the end of their childbearing years. In other words, it is the anticipated future increment of proportion of women at parity 2.

This interpretation is valid under the assumptions that women had at most one birth in the given year and that current fertility will remain constant until the end of women's reproductive life.

C. Next, we derive projected proportions of women aged x who will attain at least parity i by the end of their childbearing years using formula (21). Thus, projected proportions $M_x^*(i)$ are calculated by adding the future order increments (18) to $M_x(i)$. Finally, the Projected Parity Progression Ratios between for parity i are computed using formula (22), i.e. as ratios of proportions $M_x^*(i)$ of women expected to attain each successive parity at any given age (Table 5).

For instance, the proportions of women aged 25 projected to achieve at least parity 2 by the end of their childbearing life equals

$$M_{25}^*(2) = M_{25}(2) + TOFR(2) - TOFR_{25}(2) =$$

$$= 0.0723 + 0.4191 = 0.4914.$$

It follows that the proportion of women aged 25 with one child who are projected to have at least two children is 75.47% ($PPPR_{25}(1) = 0.7547$), whereas the proportion of women with two births who are projected to have at least three children is 28.12% ($PPPR_{25}(2) = 0.2812$).

Table 5: Projected proportions $M_x^*(i)$ of women expected to attain at least parity i and projected parity progression ratios $PPPR_x(i)$

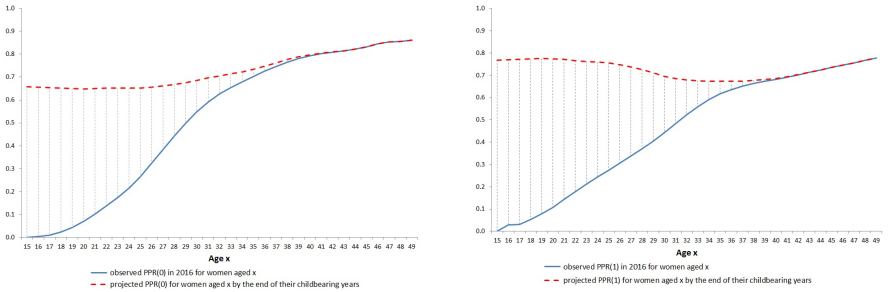
	$M_x^*(i)$				$PPPR_x(i)$			
Age x	$i=1$	$i=2$	$i=3$	$i=4$	$i=0$	$i=1$	$i=2$	$i=3$
-15	0.6570	0.5048	0.1406	0.0350	0.6570	0.7683	0.2785	0.2489
16	0.6562	0.5048	0.1406	0.0350	0.6562	0.7693	0.2785	0.2489
17	0.6541	0.5045	0.1406	0.0350	0.6541	0.7713	0.2787	0.2489
18	0.6513	0.5041	0.1405	0.0350	0.6513	0.7740	0.2787	0.2491
19	0.6494	0.5033	0.1405	0.0350	0.6494	0.7750	0.2792	0.2491
20	0.6487	0.5019	0.1403	0.0349	0.6487	0.7737	0.2795	0.2488
21	0.6494	0.5005	0.1400	0.0349	0.6494	0.7707	0.2797	0.2493
22	0.6516	0.4989	0.1398	0.0348	0.6516	0.7657	0.2802	0.2489
23	0.6518	0.4966	0.1396	0.0348	0.6518	0.7619	0.2811	0.2493
24	0.6516	0.4946	0.1388	0.0347	0.6516	0.7591	0.2806	0.2500
25	0.6511	0.4914	0.1382	0.0346	0.6511	0.7547	0.2812	0.2504
26	0.6562	0.4904	0.1383	0.0346	0.6562	0.7473	0.2820	0.2502
27	0.6625	0.4883	0.1376	0.0346	0.6625	0.7371	0.2818	0.2515
28	0.6677	0.4848	0.1363	0.0344	0.6677	0.7261	0.2811	0.2524
29	0.6750	0.4799	0.1345	0.0339	0.6750	0.7110	0.2803	0.2520
30	0.6851	0.4768	0.1336	0.0339	0.6851	0.6960	0.2802	0.2537
31	0.6970	0.4781	0.1336	0.0342	0.6970	0.6859	0.2794	0.2560
32	0.7052	0.4789	0.1327	0.0342	0.7052	0.6791	0.2771	0.2577
33	0.7138	0.4820	0.1325	0.0342	0.7138	0.6753	0.2749	0.2581
34	0.7220	0.4867	0.1330	0.0346	0.7220	0.6741	0.2733	0.2602
35	0.7343	0.4947	0.1348	0.0351	0.7343	0.6737	0.2725	0.2604
36	0.7477	0.5032	0.1374	0.0357	0.7477	0.6730	0.2731	0.2598
37	0.7616	0.5131	0.1409	0.0372	0.7616	0.6737	0.2746	0.2640
38	0.7764	0.5254	0.1472	0.0397	0.7764	0.6767	0.2802	0.2697
39	0.7883	0.5368	0.1540	0.0418	0.7883	0.6810	0.2869	0.2714
40	0.7974	0.5471	0.1617	0.0449	0.7974	0.6861	0.2956	0.2777
41	0.8051	0.5581	0.1718	0.0495	0.8051	0.6932	0.3078	0.2881
42	0.8106	0.5701	0.1830	0.0550	0.8106	0.7033	0.3210	0.3005
43	0.8141	0.5809	0.1951	0.0604	0.8141	0.7135	0.3359	0.3096
44	0.8226	0.5957	0.2089	0.0667	0.8226	0.7242	0.3507	0.3193
45	0.8330	0.6127	0.2236	0.0737	0.8330	0.7355	0.3649	0.3296
46	0.8462	0.6314	0.2402	0.0812	0.8462	0.7462	0.3804	0.3381
47	0.8534	0.6450	0.2558	0.0876	0.8534	0.7558	0.3966	0.3425
48	0.8562	0.6567	0.2680	0.0941	0.8562	0.7670	0.4081	0.3511
49	0.8610	0.6694	0.2800	0.1010	0.8610	0.7775	0.4183	0.3607

Source: Authors' own calculations.

6.3. Graphical illustration

Figures 1 and 2 allow for more detailed comparison between the projected and observed age-specific PPR's by parity summarized in Tables 3, 5.

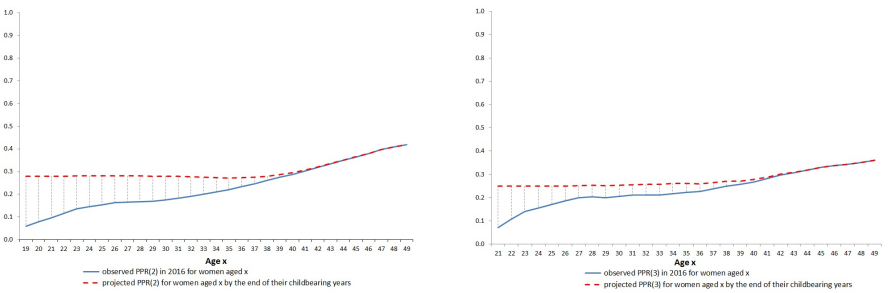
As expected, there are substantial differences between both types of PPR's, especially for young women, although differences vanish as age is getting older.



(a) Parity $i = 0$ (b) Parity $i = 1$

Figure 1: Projected and observed parity progression ratios for parities $i = 0, 1$

Source: Developed by the authors.



(a) Parity $i = 2$ (b) Parity $i = 3$

Figure 2: Projected and observed parity progression ratios for parities $i = 2, 3$

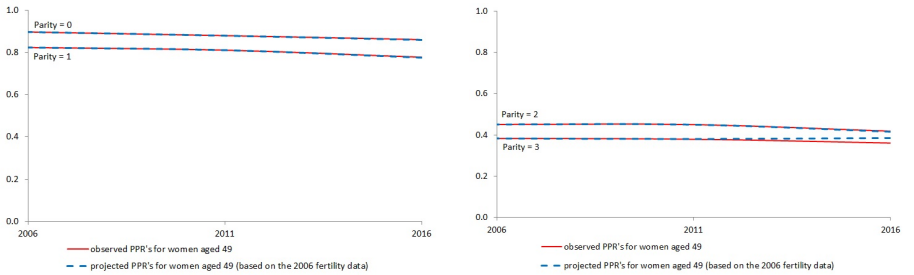
Source: Developed by the authors.

We can observe typical shapes of observed age-specific PPR's with curves increasing with age whereas projected PPR's tend to level. Both observed and projected PPR's decrease as parity increases.

6.4. Prediction

The major thrust of this section is to see if the projected parity progression ratios derived for younger women give a good prediction of their completed parity progression ratios. To achieve this, the projected ratios for parities $i = 0, 1, 2, 3$ (predictions based on the 2006 fertility data) are compared with completed ratios for women aged 49 in the years 2006, 2011 and 2016. Results are illustrated on Figure 3.

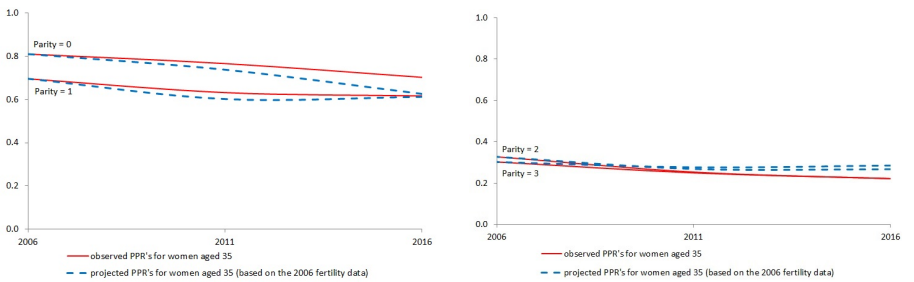
Analogous comparisons are made between the projected parity progression ratios (based on the 2006 fertility data) and the parity progression ratios for women aged 35 observed in the years 2006, 2011 and 2016. Figure 4 shows these two comparisons.



(a) (b)

Figure 3: Projected and observed (completed) PPR's for women aged 49

Source: Developed by the authors.



(a) (b)

Figure 4: Projected and observed (uncompleted) PPR's for women aged 35

Source: Developed by the authors.

Figure 3 shows that the projected and observed age-specific PPR's are very similar. Given that women aged 44 or 39 in 2006 are close to the end of their childbearing life and that there is little time for a significant fertility change, the projected ratios are almost the same as the completed PPR's observed for women aged 49 in 2011 or in 2016, respectively.

Based on the comparisons of the projected and observed PPR's for females attaining age 35 in 2011 or in 2016, we can conclude (see Figure 4) that the projections fit much better for the five-year forecast horizon compared to the ten-year horizon. Moreover, they underestimate the completed PPR's for parity $i = 0$ and overestimate the completed PPR's for parities $i = 2, 3$. This effect results from the fact that in this case the main assumption about time-invariant fertility rates is not satisfied.

In general, the projected parity progression ratios seem to provide a satisfactory prediction in the ten-year or shorter time horizon.

7. Conclusion

Parity progression ratios are important indicators explaining the pattern of fertility. They provide an alternative to conventional age-based studies of fertility trends. Traditional age-

specific fertility rates and their sum, i.e. the total fertility rate, use age as a main structural feature of the female population that may influence the number of births in a given period. However, another important structural feature is parity.

The parity analysis facilitates the interpretation of trends in the number of births and the age of women who decide to have a child. What is more, parity measures can be related more directly to behavioural factors, because a woman makes her decision about having a child not only based on how old she is but also how many children she already has.

In the paper age-specific parity progression ratios and projected parity progression ratios for the Polish female population were investigated in greater details. Based on the numerical results presented in Section 6 the following principal findings can be formulated: the decline in fertility in Poland in the near future will be caused by the gradual decrease in the propensity of women to have more than two children. There is still no problem with the desire to have one child. About 86% of young childless Polish women decide to have a child. Most of them have also a second child. The situation is much worse in the case of higher order births. The results obtained indicate that PPR's for higher parities drop down rapidly by more then half compared to the above mentioned rate.

REFERENCES

BILLARI, F. C., KOHLER, H.-P., (2004), Patterns of low and very low fertility in Europe, *Population Studies*, Vol. 58(2), pp. 161–176.

BHARDWAJ, S. B., SHARMA, G. C., KUMAR, A., (2010), Analysis of the Parity Progression Ratios, *Journal of Reliability and Statistical Studies*, Vol. 3(1), pp. 37–41.

BRASS W., (1985), Advances in Methods for Estimating Fertility and Mortality from Limited and Defective Data, London: Centre for Population Studies, London School of Hygiene and Tropical Medicine.

EUROSTAT. Fertility statistics. Data extracted in June 2019 .

FREJKA, T., CALOT, G., (2001), Cohort reproductive patterns in low fertility countries, *Population and Development Review*, Vol. 27(1), pp. 103–132.

FREJKA, T., (2008). Parity distribution and completed family size in Europe Incipient decline of the two-child family model, *Demographic Research*, Vol. 19(14), pp. 4–72.

FEENEY, G., (1983), Population Dynamics Based on Birth Intervals and Parity Progression, *Population Studies*, Vol. 37, pp. 75–89.

FEENEY, G., JINGYUAN, Yu, (1987), Period parity progression measures of fertility in China, *Population Studies*, Vol. 41, pp. 77–102.

HENRY, L., (1980), Fertility of marriages: A new method of measurement, *Population Studies Translation Series*, No. 3, United Nations (French Edition Published in 1953).

HUMAN FERTILITY DATABASE. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria).
Available at www.humanfertility.org (data downloaded on [06/06/2019]).

ISLAM, M. M., YADAVA, R. C., (1997), On the estimation of parity progression ratio, *Sankhya*, Vol. 58, series, B, pp. 200–208.

KOHLER, H.-P., BILLARI F. C., ORTEGA, J. A., (2002), The emergence of lowest-low fertility in Europe during the 1990s, *Population and Development Review*, Vol. 28(4), pp. 641–680.

MOULTRIE, T., DORRINGTON, R., HILL, A., HILL, K., TIMAUS I., ZABA B., (2013), Tools for demographic estimation, *International Union for the Scientific Study of Population*, Paris, France.

MYRSKYLA, M., GOLDSTEIN J., CHENG, Y.-H.A.,(2013), New cohort fertility forecasts for the developed world: rises, falls, and reversals, *Population and Development Review*, Vol. 39(1), pp. 31–56.

PARADYSZ, J., (1995), Birth Intervals as Period Measure Demographic Situation, [in:] *Demographic Situation Research*, University of Economics Press, Pozna"n, pp. 24–33.

PRESTON, S. H., HEUVELINE, P., GUILLOT, M., (2001), *Measuring and Modeling Population Processes*, Blackwell Publishing, UK.

SOBOTKA, T., (2013), Pathways to Low Fertility: European Perspectives, Expert Paper No. 2013/8, UN Department of Economic and Social Affairs, Population Division,
www.un.org/en/development/desa/population/publications/pdf/expert /2013-8 Sobotka Expert-Paper.pdf.

SRINIVASA, K., (1968), A set of analytical models for the study of open birth intervals, *Demography*, Vol. 5, pp. 34–44.

YADAVA, R. C., BHATTACHARYA, M., (1985), Estimation of parity progression ratios from closed and open birth interval data, Mimeo, Centre of Population Studies, Banaras Hindu University, Varanasi, India.

YADAVA, R. C., KUMAR, A., (2011), On the estimation of parity progression ratios, *Journal of Scientific Research*, Banaras Hindu University, Varanasi, Vol. 55, pp. 127–134.

YADAVA, R. C., PANDEY, A., SAXENA, N. C., (1992), Estimation of parity progression ratios from the truncated distribution of closed and open birth intervals, *Mathematical Biosciences*, Vol. 110, pp. 181–190.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 73–94, DOI 10.21307/stattrans-2020-005
Submitted – 10.10.2018; accepted for publishing – 30.09.2019

Using the ICAPM to estimate the cost of capital of stock portfolios: empirical evidence on the Warsaw Stock Exchange

Stanisław Urbański[1], **Jacek Leśkow**[2]

ABSTRACT

The aim of this paper is to present the method for estimating the cost of capital of typical portfolios available on the Warsaw Stock Exchange. The authors introduce the three factor Fama-French model and its two modifications. They also apply the bootstrap method to evaluate the variability of their estimation method. The cost of capital they refer to is related to portfolios of real options linked to projects. The market returns are generated both by stock companies running such projects and by real options modifying selected projects. The estimated cost of capital can serve as a valuable indicator for investors and for managers overseeing portfolios of stocks. Also, such an indicator can serve as a general reference while making business decisions related to new. The study demonstrated that the estimated cost of capital assumes highest values for value portfolios and stock companies with high financial indicators and, at the same time, low market prices compared to their book value. By the same token, the estimated cost of capital assumes low values for growth portfolios and for stock companies characterised by low financial indicators and, at the same time, high market prices compared to their book values.

Key words: ICAPM, cost of capital, risk premium, bootstrap method.
JEL: G11, G12

1. Introduction

The classical Capital Asset Pricing Model (CAPM), defined by Sharpe (1964) and Lintner (1965), plays a key role in the process of making investment decisions by managements of stock companies. It is widely applied to estimating the cost of capital and assessing the efficiency of the investment projects run. The research by Graham and Harvey (2001) or Welch (2008) provide some interesting examples of this use of the CAPMs. The former paper presents a survey of 392 Chief Financial Officers and the

[1] AGH University of Science and Technology, Faculty of Management, 30-067 Krakow, 10 Gramatyka Street, Poland. E-mail: surbansk@zarz.agh.edu.pl. ORCID: https://orcid.org/0000-0002-8020-8471.
[2] Institute of Teleinformatics, Cracow University of Technology, 31-155 Krakow, 24 Warszawska, Poland. E-mail: jleskow@pk.edu.pl.

capital budgeting they have supervised together with methods of estimating the cost of capital and assessing the structure of capital. This study suggests that the classical CAPM, the mean return and the multifactor CAPM, are the most popular methods of estimating the cost of capital. The method of dividend discounts, on the other hand, was quoted as the least popular. The authors state that: "… it is not clear that the model is applied properly in practice." (see: Graham and Harvey, 2001, p. 201). They also write that the main problem lies in the structure of the capital, for example in the effect of big or small companies. Some companies with higher capitalization more frequently apply the Net Present Value NPV or the CAPM method. This can cause price anomalies related to capitalization or to book to market value effects (see: Banz (1981), Rosenberg et al. (1985), Bhandari (1988) or Fama and French (1992)). Other anomalies impeding the CAPM-based pricing are analysed in papers of Lakonishok et al. (1994) or Jegadeesh and Titman (1993). Also, the research of Reinganum (1981) and Lakonishok and Shapiro (1986), whose results were confirmed by Fama and French (1992), is worth mentioning. The research advocates perceiving risk as a multidimensional factor. The applications of the Intertemporal Capital Asset Pricing Model (ICAPM) proposed by Fama and French (1993, 2015) or Carhart (1995) introduce capitalization-dependent risk factors such as book to market, profitability and investment. However, Welch (2008) favours the CAPM over the theoretical models such as the ICAPM or the Arbitrage Pricing Theory (APT).

Jagannathan and Wang (1996), Berk et al. (1999), Bernardo et al. (2007) and Zhi Da et al. (2012) attempt to explain the pricing that can be partially inconsistent with the pricing done via the CAPM model. These authors state that the stock companies frequently insure planned and carried out projects using the real options related to those projects. Therefore, the stock company can be described via a portfolio of current and future projects together with options related to such projects.

It is possible to assume that the CAPM-based estimate of the cost of the project is appropriate even in the case when pricing of the company is not consistent with CAPM. For example, Zhi Da et al. (2012, p. 205) clearly state that the risk premium and the beta factors related to them are related to the risk of such project and options related to projects. Therefore, they state that "… the CAPM could work well on the option-adjusted risk premium and beta." The quoted authors propose procedures to modify the relations between options and pricing and construct the option-adjusted beta, and option-adjusted stock returns.

Using the above literature survey we can state that if all projects of stock companies are not secured with real options then the necessary and sufficient condition to estimate the cost of capital using classical CAPM or ICAPM is to use pricing according to CAPM or ICAPM. Therefore, the estimate of the cost of capital will be more reasonable for the assets more resistant with the respect to price anomalies. Assuming that the correctness of CAPM-based pricing was established for portfolios, it is interesting to estimate the

capital cost for characteristic portfolios of a given market, see Cochrane (2001, p.445). Ferson and Locke (1998) indicate that the necessary condition for proper estimation of the CAPM-based cost of capital is proper estimation of the risk premium. This is more important than the estimation of betas, which are sometimes not adequate. Therefore, a precise estimation of the risk premium and pricing, which can be used when ICAPM or CAMP are appropriate, requires pricing that leads to generation of multifactor-efficient portfolios.

Our paper presents different methods of estimating of the cost of capital using stocks coming from the Warsaw Stock Exchange (WSE). We provide estimate of the cost of capital for the characteristic portfolios. In order to precisely estimate the risk premium we apply selected applications of ICAPM. Research regarding the Polish market is mainly focused on testing the classical CAPM. See also paper of Zarzecki et al. (2004-2005), and Czapkiewicz and Wójtowicz (2014) regarding the role of ICAPM in estimating the risk premium.

Our previous research (see Urbański et al. (2014) and Urbański (2015)) show that elimination of speculative and penny stocks enables generating multifactor-efficient portfolios using selected applications of ICAPM. In the first part of our work we show the precise and wide research in this direction. In order to obtain that, we apply the classical Fama-French model, a modified Fama-French model (see Urbański (2012)) and a new modification of the Fama-French model, which is presented in Section 2.1 of the paper.

Estimation of the cost of capital is also related to calculating the error of such estimation at a given significance level. To accomplish that, we present a method of building the confidence interval for the cost of capital. In our approach the cost of capital is the product of systematic risk and risk premium components. The risk and the risk premium components are defined as parameters of the corresponding regression models. In such regression models the monthly returns have a distribution that is close to normal. Therefore, the distribution of the regression parameter estimators should be also close to normal. However, as the capital cost is a nonlinear function of estimated normal distributions, it cannot be assumed to be normal. In order to deal with this difficulty, the bootstrap method is applied.

In Section 2 of our paper we present various methods of estimating the cost of capital using ICAPM. Section 3 describes the bootstrap of residuals as a method to investigate the distribution and variability of the estimator of the capital cost. Section 4 of our paper presents the results of estimation of the risk premium using pricing applications for different boundary conditions for penny and speculative stocks. In this Section we also show the distribution of the estimated cost of capital and the related confidence intervals for characteristic portfolios. Section 5 contains the summary and the conclusion of our work.

2. Cost of equity capital using the ICAPM applications

Our starting point is ICAPM expressed as follows:

$$E(r_i) = E(RF) + \sum_{k=1}^{K} \beta_{ik} E(F_k), \tag{1}$$

where $E(F_k)$ is the systematic risk premium vector for the analysed market, and β_{ik} is the systematic risk vector of stock (portfolio) i, $E(RF)$ is the expected return of risk free asset, and $E(r_i)$ is the expected return of analysed asset.

The corresponding econometric model, useful for estimating the parameters β from (1), is expressed as follows: (layer 1)

$$r_{it} - RF_t = \beta_{i0} + \sum_{k=1}^{K} \beta_{ik} F_{kt} + e_{it};\ t = 1,\ldots, T;\ \forall i{=}1,\ldots m, \tag{2}$$

where F_{kt} is the value of k factor model in the period t.

Once the estimators $\widehat{\beta_{ik}}$ are calculated using (2), then we use them to get the estimator $\widehat{\gamma_k}$ of the parameter γ_k using the following equation: (layer 2)

$$r_{it} - RF_t = \gamma_0 + \sum_{k=1}^{K} \gamma_k \widehat{\beta_{ik}} + e_{it};\ t = 1,\ldots, T;\ i{=}1,\ldots m, \tag{3}$$

We use the estimators γ_k to get the point value of capital cost ($Ccap_i$) for each analysed stock (portfolio), using the equation (1) (layer 3). The period of estimating systematic risk components β_{ik} is subjective. Betas were most often estimated using 60 monthly periods, due to the average duration of the business cycle. However, it seems reasonable to extend the beta estimation period due to the observable extended business cycles in the last two decades. Thus, we estimate $Ccap_i$ based on the T months, according to Eq. (4):

$$Ccap_i = E(RF) + \widehat{\gamma_0} + \sum_{k=1}^{K} \widehat{\gamma_k} \widehat{\beta_{ik}^T};\ \forall\ i{=}1,\ldots, m \tag{4}$$

The estimator $\widehat{\beta_{ik}^T}$ in equation (4) is obtained from the following equation:

$$r_{it} = \beta_{i0}^T + \sum_{k=1}^{K} \beta_{ik}^T F_{kt} + s_{it};\ t = 1,\ldots, T;\ \forall i{=}1,\ldots, m. \tag{5}$$

The purpose of this work is to study the variability of assessing the value $Ccap_i$. The most direct and comprehensive method to do this is to apply the bootstrap technique. For regression models like (2) and (3), the popular bootstrap method is bootstrapping the residuals. We will present details of this method in Section 3.

2.1. Three different ICAPM applications

We test the following three pricing applications to estimate the risk premium.

1) The classical three factor Fama-French model (see: Fama and French, 1993) - this application hereafter is denoted as FF model.
2) The modified three factor Fama-French model - this application hereafter is called M93FF, see Urbański (2012).
3) The modified three factor Fama-French model according to Fama and French (1995) work – this application hereafter is called M95FF. This application is presented below.

Based on the statements of Fama and French (1995), Urbański (2017, p. 84) assume that "The economic state variable that produces variation in the future earnings and returns related to size and *BV/MV* is a vector of structure of the past long-term differences in profitability." However, in fact, the results of Fama and French's (1995) research indicate that future returns are generated by changes in long-term relationships of past earnings to the book value of the company (see: Fama and French 1995, pp. 134-140, Figs 1 and 2, and Table 1).

Therefore, the pricing application, proposed by Urbański (2011), is modified. According to this new modification the adopted general state variable can be reflected by functional *FUN*, defined by equations (6), (7) and (8).

$$FUN = \frac{NUM}{DEN} = \frac{nor(ROE)\times nor(ASB)\times nor(APOB)\times nor(APNB)}{nor(MV/E)\times nor(MV/BV)}, \tag{6}$$

where

$$ROE = F_1; ASB = F_2 = \frac{\{\sum_{t=1}^{i}[S(Q_t)]\}/BV(Q_t)}{\sum_{t=1}^{i}\overline{SBV(nQ_t)}};\ \ APOB = F_3 = \frac{\{\sum_{t=1}^{i}[PO(Q_t)]\}/BV(Q_t)}{\sum_{t=1}^{i}\overline{POBV(nQ_t)}};$$
$$APNB = F_4 = \frac{\{\sum_{t=1}^{i}[PN(Q_t)]\}/BV(Q_t)}{\sum_{t=1}^{i}\overline{PNBV(nQ_t)}};\ MV/E = F_5;\ \ MV/BV = F_6. \tag{7}$$

F_j (j=1,…,6) are transformed to normalized areas $<a_j; b_j>$, according to Eq. (10):

$$nor(F_j) = [a_j + (b_j - a_j) \times \frac{F_j - c_j \times F_j^{min}}{d_j \times F_j^{max} - c_j \times F_j^{min} + e_j}]. \tag{8}$$

In Equations (6) and (7), the corresponding indications are as follows: *ROE* is return on book equity; $\sum_{t=1}^{i} S(Q_t), \sum_{t=1}^{i} PO(Q_t), \sum_{t=1}^{i} PN(Q_t)$ are values that are accumulated from the beginning of the year as net sales revenue (*S*), operating profit (*PO*) and net profit (*PN*) at the end of '*i*' quarter (Q_i); $\sum_{t=1}^{i}\overline{SBV(nQ_t)}$; $\sum_{t=1}^{i}\overline{POBV(nQ_t)}$; $\sum_{t=1}^{i}\overline{PNBV(nQ_t)}$ are average values, accumulated from the beginning of the year as *S/BV*, *PO/BV* and *PN/BV* at the end of Q_i over the last *n* years (the present research assumes that *n*=3 years); *BV* is the book value, *MV/E* is the

market-to-earning value ratio; E is the average earning for the last four quarters; MV/BV is the market-to-book value ratio; a_j; b_j; c_j; d_j; e_j are variation parameters. In equilibrium modelling, F_j (j=1,…,6) can be transformed into equal normalized area <1;2> (see Urbański, 2011).

In the case of the proposed multifactor model, as the modification of FF three factor model, the factors of equation (2) are defined as follows:

$$F_{1t} = RM_t - RF_t,\ F_{2t} = HLMN_t,\ F_{3t} = LMHD_t, \tag{9}$$

where $HMLN_t$ (high minus low) is the difference between the returns from the portfolio with the highest and lowest NUM_t values in period t; $LMHD_t$ (low minus high) is the difference between the returns from the portfolio with the lowest and highest DEN_t values in period t; $RM_t - RF_t$ is the market factor, defined as excess return of stock index – WIG (RM) over the risk-free rate (RF).

Considering (9), it can be shown that pricing model (1) can be written as follows:

$$E(r_i) = E(RF) + \beta_{i,M}E(RM - RF) + \beta_{i,HMLN}E(HMLN) + \beta_{i,LMHD}E(LMHD). \tag{10}$$

We analyse the research proposed by Kan and Zhang (1999) in order to check the possibility of incorrect specification of the model which is evidenced through the incorrect selection of factors (selection of useless factors). The results obtained by Kan and Zhang (1999) refer mainly to asymptotic cases, i.e. to study large samples of time series. Urbański (2011) carried out tests recommended by Kana and Zhang (1999), regarding the usefulness of the proposed factors, defining the applied M93FF applications and showed that the tested factors are not useless.

Our research regarding both M93FF and M95FF pricing applications cannot be considered as asymptotic. Our sample is at most 252 data points. It is frequently observed that even for samples of the size 200 in a very simple autoregressive AR(1) model the estimators do not necessarily achieve the normality of their distribution.

This is why we have decided to use the bootstrap approach. For small and medium-sized samples from time series, bootstrap is known to produce more reliable results when constructing confidence intervals and tests. The proposed new factors, of the M95FF application, result from a linear modification of the state functional, defining the factors of the M93FF application. The state functional, defining the factors of the M93FF application, is proposed in work of Urbański (2012, p. 555). The state functional, defining the factors of the M95FF application, is described by Eqs. 6, 7 and 8. An additional argument for the usefulness of the M95FF factors is the adjusting for errors-in-variables, by using Shanken's (1992) t-statistic. This is one of the tests

recommended by Kan and Zhang (1999). Therefore, it can be assumed that the proposed changes do not result in uselessness factors of the M95FF application.

In connection with the above-mentioned argumentation, we conclude that in our case carrying out the other uselessness factors tests suggested by Kan and Zhang (1999) is not required.

3. Residuals bootstrap and variability of capital cost

In this section we propose the bootstrap of residuals in each of the layers of the three layer model introduced in the previous Section. Bootstrap methods are widely described by Efron and Tibshirani (1993). Below, we present the bootstrapping algorithm that leads to assessment of the variability of *Ccap*.

STEP 1 Bootstrapping of $\widehat{\boldsymbol{\beta}_{ik}}$

In this step apply the GLS method to get the estimator $\widehat{\beta_{ik}}$ from Eq. 2.

Bootstrap replication $j = 1$, first run.

Sample with replacement $\left\{\widehat{e_{i1}^{*1}}, \ldots, \widehat{e_{iT}^{*1}}\right\}$ from the residuals $\{\widehat{e_{i1}}, \ldots, \widehat{e_{iT}}\}$ of Eq. 2.

Treating $\widehat{\beta_{ik}}$ as given, use Eq. 2 to get bootstrap replication of excess returns $\{(r_{i1} - RF_1)^{*1}, \ldots, (r_{iT} - RF_T)^{*1}\}$ corresponding to bootstrap replicates of $\left\{\widehat{e_{i1}^{*1}}, \ldots, \widehat{e_{iT}^{*1}}\right\}$.

Put values $\{(r_{i1} - RF_1)^{*1}, \ldots, (r_{iT} - RF_T)^{*1}\}$ back to the model (2) to get the first bootstrap replication of $\widehat{\beta_{ik}^{*1}}$.

STEP 2 Bootstrapping of $\widehat{\boldsymbol{\gamma_k}}$

In this step put $\widehat{\boldsymbol{\beta}_{\boldsymbol{ik}}^{*\mathbf{1}}}$ into Eq. 3 to get the bootstrap replicate $\widehat{\boldsymbol{\gamma}_{\boldsymbol{k}}^{*\mathbf{1}}}$ of the estimator $\boldsymbol{\gamma_k}$.

Repeat STEP1 and STEP2 j=1, …, B times. Again, taking at least B=1000.

As a result, obtain bootstrapped values of beta and gamma estimators:

$\left\{(\widehat{\boldsymbol{\beta}_{\boldsymbol{i1}}^{*\mathbf{1}}}, \ldots, \widehat{\boldsymbol{\beta}_{\boldsymbol{iK}}^{*\mathbf{1}}}), \ldots, (\widehat{\boldsymbol{\beta}_{\boldsymbol{i1}}^{*\boldsymbol{B}}}, \ldots, \widehat{\boldsymbol{\beta}_{\boldsymbol{iK}}^{*\boldsymbol{B}}})\right\}$, and $\left\{(\widehat{\boldsymbol{\gamma}_{\mathbf{1}}^{*\mathbf{1}}}, \ldots, \widehat{\boldsymbol{\gamma}_{\boldsymbol{K}}^{*\mathbf{1}}}), \ldots, (\widehat{\boldsymbol{\gamma}_{\mathbf{1}}^{*\boldsymbol{B}}}, \ldots, \widehat{\boldsymbol{\gamma}_{\boldsymbol{K}}^{*\boldsymbol{B}}})\right\}$.

STEP 3 Bootstrapping of $\boldsymbol{\beta_{ik}^{T}}$

In model (5) proceed identically as in STEP 1, and STEP 2. Bootstrapping residuals will give you B replicates of the estimator $\widehat{\beta_{ik}^{T}}$, that is

$\left\{(\widehat{\boldsymbol{\beta}_{\boldsymbol{i1}}^{\boldsymbol{T1}}}, \ldots, \widehat{\boldsymbol{\beta}_{\boldsymbol{iK}}^{\boldsymbol{T1}}}), \ldots, (\widehat{\boldsymbol{\beta}_{\boldsymbol{i1}}^{\boldsymbol{TB}}}, \ldots, \widehat{\boldsymbol{\beta}_{\boldsymbol{iK}}^{\boldsymbol{TB}}})\right\}$ for stock (portfolio) i and k=1, …, K factors.

STEP 4 Bootstrapping of $\boldsymbol{E(RF)}$

Focus now on independent risk free rates *RF*. The natural estimate of the parameter E(*RF*) is $\boldsymbol{\mu} = \mathbf{1}/_{\boldsymbol{T}} \sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{RF_t}$.

Obtain bootstrap replicates $\left\{\widehat{\boldsymbol{\mu}^{*\mathbf{1}}}, \ldots, \widehat{\boldsymbol{\mu}^{*\boldsymbol{B}}}\right\}$ drawing samples $\left\{\boldsymbol{RF}_{\mathbf{1}}^{*\boldsymbol{j}}, \ldots, \boldsymbol{RF}_{\boldsymbol{T}}^{*\boldsymbol{j}}\right\}$; j=1, …, B with replacement from $\{\boldsymbol{RF_1}, \ldots, \boldsymbol{RF_T}\}$,

here: $\widehat{\boldsymbol{\mu}^{*\boldsymbol{j}}} = \mathbf{1}/_{\boldsymbol{T}} \sum_{\boldsymbol{t=1}}^{\boldsymbol{T}} \boldsymbol{RF}_{\boldsymbol{t}}^{*\boldsymbol{j}}$.

STEP 5 Bootstrapping of $\widehat{Ccap_i}$

Use the replication $\left\{(\widehat{\gamma_1^{*1}}, \ldots, \widehat{\gamma_K^{*1}}), \ldots, (\widehat{\gamma_1^{*B}}, \ldots, \widehat{\gamma_K^{*B}})\right\}$ and $\left\{(\widehat{\beta_{i1}^{T1}}, \ldots, \widehat{\beta_{iK}^{T1}}), \ldots, (\widehat{\beta_{i1}^{TB}}, \ldots, \widehat{\beta_{iK}^{TB}})\right\}$ to create $\left\{Ccap_i^{*1}, \ldots, Ccap_i^{*B}\right\}$. Use the formula:

$$\widehat{Ccap_i^{*J}} = \widehat{\mu^{*J}} + \widehat{\gamma_0^{*J}} + \sum_{k=1}^{K} \widehat{\gamma_k^{*J}}\,\widehat{\beta_{ik}^{TJ}}; \qquad \forall i=1,\ldots,m, \qquad (11)$$

With the bootstrap sample $\left\{Ccap_i^{*1}, \ldots, Ccap_i^{*B}\right\}$, for each portfolio separately, we are able to assess the variability of $Ccap_i$ and also the sampling distribution of $Ccap_i$.

In the next Section of our paper we show practical applications of the procedure described above. Here, we would like to make an additional point regarding bootstrapping $\hat{\gamma}$ and $\widehat{\beta^T}$ in formula (4). In financial applications, one usually takes shorter samples to get the estimates of $\widehat{\beta^T}$. This is motivated by the stability requirements. On the other hand, to get estimates of $\widehat{\beta^T}$ one prefers longer samples, for the sake of accuracy. From the methodological point of view, this does not create problems as long as the shorter sample includes at least 100 observations. The residual bootstrap was proven to be a consistent procedure (see, e.g. Lahiri (2003)) in regression context as well as in the time series context.

4. Data and interpretation of results

We analyse monthly returns of the stocks listed on the WSE in 1995-2017. Data referring to the fundamental results of the inspected companies are taken from the database drawn up by Notoria Serwis Company. Data for defining returns on securities are provided by the WSE.

The cost of capital is evaluated using three ICAPM applications presented in Section 2.1. In the case of FF model the quintile portfolios are formed on *BV/MV*. These portfolios, in turn, are divided into other quintiles formed using the capitalization (*CAP*) calculated for each stock company. In the case of the M93FF and M95FF models the quintile portfolios are formed using the *NUM* function. Again, these portfolios are, in turn, divided into other quintiles formed using the *DEN* function. Our analysis is then conducted for the 25 most characteristic portfolios of the market.

4.1. Risk premium vector components

In order to test whether the pricing application generates multifactor-efficient portfolios ten modes of samples are analysed. Mode 1 considers all the WSE stocks except companies characterized by a negative book value. In modes: from 2 to 8 penny stocks, with market values lower than 0.5, 1.0, 1.5, 2.0, 3.0, 4.0 and 5.0 PLN are

eliminated, while modes 9 and 10 examine hypothetical cases of the exclusion speculative stocks. Mode 9 (indicated SPEC1) rejects the stocks that meet the following conditions: 1) $MV/BV > 100$; 2) $ROE < 0$ and $BV > 0$; and 3) $MV/BV > 30$ and $r_{it} > 0$, Mode 10 (indicated SPEC2) rejects the stocks meeting additional condition 4) $MV/E < 0$.

In Tables 1, 2 and 3 below we show the estimated values of the risk premium vector, estimated from second-pass regressions by the classical FF, M93FF and M95FF models, for eliminated penny stocks and speculative stocks.

Table 1. The values of the risk premium vector (γ) estimated from second-pass regressions for the classical Fama-French model

$$r_{it} - RF_t = \gamma_0 + \gamma_M \widehat{\beta_{i.M}} + \gamma_{HML} \widehat{\beta_{i.HML}} + \gamma_{SMB} \widehat{\beta_{i,SMB}} + \varepsilon_{it};\ \ t=1.\ldots.252;\ i=1.\ldots.25$$

Parameter	Excluded penny stocks below (PLN)								Excluded speculative stocks	
	0.0	0.5	1.0	1.5	2.0	3.0	4.0	5.0	SPEC1	SPEC2
	Mode 1	Mode 2	Mode 3	Mode 4	Mode 5	Mode 6	Mode 7	Mode 8	Mode 9	Mode 10
γ_0, %	-3.01	-3.16	-3.70	-3.48	-3.51	-2.96	-3.17	-3.24	-2.74	-2.81
t-stat	-1.94	-2.08	-2.57	-2.58	-2.62	-2.49	-2.74	-2.63	-1.77	-1.96
SH t-stat	-1.79	-1.90	-2.29	-2.32	-2.35	-2.28	-2.50	-2.38	-1.69	-1.82
p-value, %	7.39	5.76	2.22	2.04	1.88	2.25	1.25	1.76	9.18	6.83
γ_{HML}, %	2.11	1.99	1.89	1.54	1.32	1.80	1.32	1.43	-0.67	1.17
t-stat	3.44	3.40	3.32	2.53	2.10	2.91	2.20	2.35	-1.23	2.05
SH t-stat	3.57	3.50	3.42	2.58	2.11	2.91	2.21	2.33	-1.24	2.07
p-value, %	0.04	0.05	0.06	1.00	3.51	0.37	2.73	1.98	21.38	3.89
γ_{SMB}, %	-0.05	-0.02	0.09	0.35	0.46	0.24	0.31	0.24	-1.63	0.69
t-stat	-0.10	-0.05	0.19	0.75	0.99	0.51	0.66	0.49	-3.75	1.60
SH t-stat	-0.10	-0.05	0.20	0.78	1.03	0.52	0.67	0.50	-3.76	1.65
p-value, %	91.68	96.03	83.91	43.42	30.30	60.49	50.37	61.72	0.02	9.98
γ_M, %	2.58	2.75	3.29	3.15	3.17	2.66	2.87	2.96	1.76	2.46
t-stat	1.52	1.65	2.08	2.10	2.13	2.00	2.24	2.16	1.04	1.58
SH t-stat	1.41	1.52	1.87	1.90	1.93	1.85	2.06	1.97	0.99	1.48
p-value, %	15.79	12.97	6.19	5.74	5.36	6.49	3.93	4.84	32.34	14.02
GRS-F	1.83	1.38	1.50	1.41	1.67	1.29	1.99	1.40	3.29	1.18
p-value, %	1.16	11.21	6.70	10.23	2.84	16.92	0.46	10.49	0.00	25.58
$Q^A(F)$	1.00	1.05	0.99	0.92	1.07	0.83	0.80	0.86	2.33	0.77
p-value, %	46.91	39.98	47.13	56.85	38.71	67.80	71.52	64.20	0.12	75.43
R^2_{LL}	68.45	64.15	54.33	47.74	40.92	68.88	66.69	56.66	25.61	34.79

Note: This table presents Fama-MacBeth cross-sectional regressions using the excess returns on 25 portfolios sorted by *BV/MV* and capitalization *CAP*. 252 monthly periods are analysed from May 1995 through May 2017. RF_t is the 91-day Polish Treasury bill return. $\widehat{\beta_{i.M}}$ is the loading on the market factor estimated from first-pass time-series regressions. $\widehat{\beta_{i.HML}}$ and $\widehat{\beta_{i,SMB}}$ are loadings on the *HML* and *SMB* factors. GRS-*F* is F-statistic of Gibbons et al. (1989). $Q^A(F)$ reports *F*-statistic and its

corresponding *p*-value indicated below in brackets for Shanken's (1985) test that the pricing errors in the model are jointly zero. SH *t*-stat is Shanken's (1992) statistic adjusting for errors-in-variables. Following Lettau and Ludvigson (2001), R^2_{LL} is a measure that shows the fraction of the cross-sectional variation in average returns that is explained by each model and is calculated as follows: $R^2_{LL} = [\sigma_c^2(\bar{r}_i) - \sigma_c^2(\bar{\varepsilon}_i)] / \sigma_c^2(\bar{r}_i)$, where σ_c^2 denotes a cross-sectional variance, and variables with bars above denote time-series averages. The Prais-Winsten procedure for correction of first-lag autocorrelation is used. SPEC1 eliminates speculative stocks meeting one of the following boundary conditions: 1) $MV/BV > 100$; 2) $ROE < 0$ and $BV > 0$; and 3) $MV/BV > 30$ and $r_{it} > 0$, where *MV* is stock market value. *ROE* is return on book value (*BV*). r_{it} is return of portfolio *i* during period *t*. SPEC2 eliminates speculative stocks meeting additional condition 4) $MV/E < 0$, where *E* is average earning for last four quarters. Source: own research.

Table 2. The values of the risk premium vector (γ) estimated from second-pass regressions for the modified Fama-French model M93FF

$$r_{it} - RF_t = \gamma_0 + \gamma_M \widehat{\beta_{i.M}} + \gamma_{HMLN} \widehat{\beta_{i.HMLN}} + \gamma_{LMHD} \widehat{\beta_{i,LMHD}} + \varepsilon_{it};\quad t=1.....252;\ i=1..... 25$$

Parameter	Excluded penny stocks below (PLN)								Excluded speculative stocks	
	0.0 Mode 1	0.5 Mode 2	1.0 Mode 3	1.5 Mode 4	2.0 Mode 5	3.0 Mode 6	4.0 Mode 7	5.0 Mode 8	SPEC1 Mode 9	SPEC2 Mode10
γ_0, %	-3.72	-3.34	-3.02	-2.72	-2.42	-1.89	-1.18	-1.89	-9.85	-7.02
t-stat	-3.18	-2.94	-2.84	-2.46	-2.18	-1.41	-1.02	-1.36	-8.49	-7.35
SH t-stat	-2.75	-2.59	-2.56	-2.22	-2.00	-1.32	-0.97	-1.26	-4.53	-5.63
p-value, %	0.59	0.96	1.05	2.64	4.52	18.82	33.41	20.66	0.00	0.00
γ_{HMLN}, %	0.85	1.05	0.96	0.99	0.92	0.97	0.97	1.04	4.01	1.65
t-stat	2.81	3.47	3.12	3.23	3.18	3.44	3.32	3.71	14.85	6.22
SH t-stat	2.82	3.44	3.10	3.20	3.14	3.40	3.28	3.65	15.28	5.77
p-value, %	0.48	0.06	0.20	0.14	0.17	0.07	0.10	0.03	0.00	0.00
γ_{LMHD}, %	0.97	0.91	0.86	1.03	1.01	1.07	1.03	0.95	2.84	1.28
t-stat	3.13	2.93	2.75	3.29	3.39	3.76	3.60	3.38	9.93	4.87
SH t-stat	3.24	2.96	2.75	3.25	3.36	3.76	3.61	3.40	9.70	5.13
p-value, %	0.12	0.30	0.60	0.12	0.08	0.02	0.03	0.07	0.00	0.00
γ_M, %	3.65	3.25	2.88	2.67	2.24	1.59	0.89	1.68	9.48	5.12
t-stat	2.84	2.61	2.48	2.20	1.82	1.14	0.69	1.09	7.56	4.43
SH t-stat	2.50	2.33	2.26	2.01	1.69	1.07	0.66	1.02	4.22	3.50
p-value, %	1.26	2.00	2.38	4.50	9.11	28.57	50.70	30.85	0.00	0.05
GRS-F	3.64	2.82	3.02	2.63	2.52	2.34	1.93	2.42	9.84	7.92
p-value, %	0.00	0.00	0.00	0.01	0.02	0.06	0.67	0.03	0.00	0.00
$Q^A(F)$	2.27	2.08	1.94	1.32	1.87	1.61	1.37	1.70	2.59	1.27
p-value, %	0.17	0.47	0.96	16.45	1.36	4.78	13.19	3.14	0.03	19.66
R^2_{LL}	41.76	47.67	42.72	57.25	52.86	57.63	38.43	53.61	77.67	74.27

Note: see Table 1. Source: own research.

The M95FF model turns out to be the application generating the portfolio closest to the multifactor-efficient portfolio if penny stocks below 1.5 PLN are excluded. This is evidenced by the test results: GRS-F=1.53 with corresponding p-value=5.58%, and $Q^A(F)$=1.51 with p-value=7.47%, and γ_0=0.99 with p-value=53.52%. The use of the modified pricing application M95FF significantly improves the description of returns in comparison with the M93FF application. As opposed to M93FF, M95FF generates zero value intercepts at the significance level over 20% for almost all tested cases. In the light of ICAPM, the application M95FF gives a good description of returns if penny stocks below 2.0 PLN, 4.0 PLN and 5.0 PLN are excluded, as well as the classical FF model if penny stocks below 0.5 PLN are excluded. The hypothetical mode 10 (SPEC2) generates a good description of returns, especially using the M95FF application. The statistic $Q^A(F)$ and coefficient R^2_{LL} take small and high values: 0.70 and 87.82% respectively. However, for this case the intercept is significant, assuming high negative value -7.02% with corresponding p-values about 0.00%. In the case of mode 9 and mode 10, the prices of rejected speculative stocks vary over a wide range. However, 22.28% SPEC1 stocks and 20.35% SPEC2 stocks have a price of less than 2.00 PLN, and the largest number of speculative stocks is in the range from 1.00 PLN to 2.00 PLN, see Urbański et al. (2014).

Table 3. The values of the risk premium vector (γ) estimated from second-pass regressions for the modified Fama-French model M95FF

$$r_{it} - RF_t = \gamma_0 + \gamma_M \widehat{\beta_{i,M}} + \gamma_{HMLN} \widehat{\beta_{i,HMLN}} + \gamma_{LMHD} \widehat{\beta_{i,LMHD}} + \varepsilon_{it};\quad t=1.\ldots.252;\ i=1.\ldots.25$$

Parameter	Excluded penny stocks below (PLN)								Excluded speculative stocks	
	0.0 Mode 1	0.5 Mode 2	1.0 Mode 3	1.5 Mode 4	2.0 Mode 5	3.0 Mode 6	4.0 Mode 7	5.0 Mode 8	SPEC1 Mode 9	SPEC2 Mode10
γ_0, %	0.35	0.25	-0.69	0.99	1.69	0.22	-1.17	-2.24	-9.67	-7.02
t-stat	0.20	0.15	-0.46	0.66	1.00	0.12	-0.86	-1.40	-8.13	-8.07
SH t-stat	0.19	0.14	-0.45	0.62	0.92	0.12	-0.81	-1.27	-4.30	-5.96
p-value, %	84.95	88.55	65.35	53.52	35.87	90.58	41.90	20.50	0.00	0.00
γ_{HMLN}, %	0.74	0.83	0.79	0.87	1.02	0.80	0.87	0.78	3.45	2.58
t-stat	2.31	2.68	2.58	2.80	3.32	2.60	2.99	2.59	12.01	9.68
SH t-stat	2.30	2.68	2.57	2.78	3.32	2.60	2.98	2.56	14.43	9.63
p-value, %	2.21	0.79	1.07	0.58	0.10	0.99	0.31	1.10	0.00	0.00
γ_{LMHD}, %	0.83	0.88	0.73	0.84	0.83	0.90	1.18	1.18	2.80	0.99
t-stat	2.64	2.93	2.55	2.87	2.92	3.07	3.98	3.79	9.64	3.69
SH t-stat	2.67	2.92	2.55	2.86	2.89	3.04	3.92	3.72	9.80	3.69
p-value, %	0.81	0.38	1.13	0.45	0.41	0.26	0.01	0.02	0.00	0.01

Table 3. The values of the risk premium vector (γ) estimated from second-pass regressions for the modified Fama-French model M95FF (cont.)

$r_{it} - RF_t = \gamma_0 + \gamma_M\widehat{\beta_{i,M}} + \gamma_{HMLN}\widehat{\beta_{i,HMLN}} + \gamma_{LMHD}\widehat{\beta_{i,LMHD}} + \varepsilon_{it}$; t=1…..252; i=1….. 25

Parameter	Excluded penny stocks below (PLN)								Excluded speculative stocks	
	0.0 Mode 1	0.5 Mode 2	1.0 Mode 3	1.5 Mode 4	2.0 Mode 5	3.0 Mode 6	4.0 Mode 7	5.0 Mode 8	SPEC1 Mode 9	SPEC2 Mode10
γ_M, %	-0.72	-0.63	0.38	-1.42	-2.09	-0.58	1.05	2.17	9.68	5.07
t-stat	-0.37	-0.34	0.23	-0.87	-1.15	-0.30	0.70	1.20	7.28	4.67
SH t-stat	-0.36	-0.33	0.23	-0.82	-1.06	-0.29	0.66	1.10	4.01	3.58
p-value, %	71.71	74.22	82.08	41.37	29.04	77.02	51.11	27.40	0.01	0.07
GRS-F	1.59	2.33	1.62	1.53	1.86	2.32	2.00	1.66	9.33	8.39
p-value, %	4.24	0.06	3.57	5.58	0.99	0.06	0.44	2.90	0.00	0.00
$Q^A(F)$	1.42	1.54	1.68	1.51	1.32	1.67	1.35	0.94	1.87	0.70
p-value, %	10.82	6.64	3.52	7.47	16.22	3.73	14.63	54.18	1.42	83.14
R^2_{LL}	36.12	42.21	37.33	36.09	42.57	42.93	59.08	67.72	70.11	87.82

Note: see Table 1. Source: own research.

Rejecting the hypothetical modes SPEC1 and SPEC2, we decide that a deeper analysis should focus on the pricing application M95FF if penny stocks below 2.0 PLN are excluded. The risk prices γ_{HMLN} and γ_{LMHD} assume values of 1.02% and 0.83% (for monthly periods) with corresponding p-values 0.1% and 0.41%. Although the risk price γ_M assumes an insignificant negative value of -2.09% (with corresponding p-values 29.04%) it does not contradict the ICAPM assumptions (see Fama, 1996, pp. 456 and 463-464). This fact confirms the decisive impact of risk due to *HMLN* and *LMHD* factors. In this case the cross-section determination coefficient R^2_{LL} increases from 36.09% to 42.57%, and elimination of penny stocks below 4.0 PLN or 5.0 PLN significantly deviates from the upper limit 1.0 PLN set by the WSE.

In the case of the classical FF model (if penny stocks below 0.5 PLN are excluded) the intercept γ_0=-3.16 is insignificant on the borderline level 5.76%. Also, only risk premium component γ_{HML}=1.99 is statistically significant (with corresponding p-value=0.05%) for insignificant γ_{SMB}=-0.02 and γ_M=2.75 (with corresponding p-value=96.03%, and p-value=12.97%, respectively).

4.2. Distributions of capital cost of modelled portfolios

The controversial problem of capital cost assessment is the betas estimation method, and the number of monthly estimation periods. It seems appropriate to use linear multivariate regression in accordance with equation 7. However, one can also

consider the use of several mono-variable regressions, relative to the examined factors. In the case of commonly used financial applications for betas assessment, the samples with 60 monthly periods were most often applied. However, due to the well-known fact of extending the periods of business cycles from the beginning of the 21st century, it seems appropriate to extend the estimation period to 120 months.

We attempt to estimate the betas, and thus the capital cost, in two approaches. In Approach 1 the betas are estimated on the basis of last 120 months, from period 133 to period 252. In Approach 2, the procedure similar to the one proposed by Zhi Da et al. (2012) is applied, and the betas are estimated using (t-61, t-1) a sixty-month rolling window, with rolled step of one month, using the whole tested period of 252 months.

In Table 4 below we show the statistics of normality tests of bootstrapped capital cost, bootstrapped risk premium and systematic risk components, estimated by the M95FF model, for the portfolio formed on the highest *NUM* and the smallest *DEN* values.

In Figure 1 we show histograms of bootstrapped capital cost estimated in Approach 2 (a), and in Approach 1 (b).

The cost of capital, estimated in Approach 1, does not show normality of distribution. This is clearly confirmed by the results of the four normality tests (see: Table 4). The normality of the risk premium $\widehat{\gamma_{LMHD}}$ cannot be rejected only when using the Shapiro-Wilk and Lilliefors tests (with 5% level of significance), while the hypothesis of normality of $\widehat{\gamma_{HMLN}}$ should be rejected by using any of the used tests.

The distribution of the component $\widehat{\gamma_M}$ is not normal. Normality of the distribution of systematic risk $\widehat{\beta_{1,HMLN}}$, estimated in both approaches, confirms all four tests performed. However, no test performed confirmed the distribution normality of the $\widehat{\beta_{1,LMHD}}$ component, estimated in approach 1. On the other hand, the normal distribution of $\widehat{\beta_{1,LMHD}}$, estimated in Approach 2, is confirmed at the significance level of 0.16 by the Lilliefors test. Normality of the distribution of systematic risk $\widehat{\beta_{1,M}}$, estimated in approaches 1 and 2, also confirm all four tests performed.

It can be concluded that the risk components estimated in Approach 2 have distributions closer to the normal distribution, compared to the risk components estimated in Approach 1. It seems that this results in the normality of the capital cost distribution if the risk components are estimated in Approach 2.

In Table 5 we show the estimated values of the cost of capital, by the classical FF and M95FF models, for 25 model portfolios on the basis of Approach 2.[3] In step 1 of our algorithm, model (2) to estimate the components of systematic risk is used. In step 2, model (3) to estimate the risk premium components is used. In step 3, model (5) is used to estimate current values of betas for capital cost calculation.

[3] The estimated values of the cost of capital for 25 modelled portfolios on the basis of Approach 1, and the MFF93 application (in both approaches) are available from the authors on request.

Table 4. Statistics of normality tests of bootstrapped capital cost, risk premium and systematic risk of portfolio 1, formed on the highest *NUM* and the smallest *DEN*

Parameter	Normality test statistic (p-value, %)			
	Doornik-Hansen	Shapiro-Wilk	Lilliefors	Jarque'a-Bera
Capital cost Approach 1	100.87 (1.24e-020)	0.9973 (2.07e-010)	0.01861 (0)	127.09 (2.53e-026)
Risk premium $\widehat{\gamma_{HMLN}}$	12.625 (0.18)	0.9996 (1.97)	0.0089 (5.00)	13.7557 (0.10)
Risk premium $\widehat{\gamma_{LMHD}}$	6.9690 (3.07)	0.9997 (9.45)	0.0087 (6.00)	7.2862 (2.61)
Risk premium $\widehat{\gamma_{M}}$	53.5308 (2.38e-010)	0.9988 (6.88e-005)	0.0099 (2.00)	59.1272 (1.45e-0.11)
Systematic risk - Approach 1 $\widehat{\beta_{1,HMLN}}$	2.0549 (35.79)	0.9998 (63.17)	0.0051 (75.00)	1.9348 (38.01)
Systematic risk - Approach 1 $\widehat{\beta_{1,LMHD}}$	8.8843 (1.18)	0.9995 (0.78)	0.0092 (4.00)	9.3097 (0.95)
Systematic risk - Approach 1 $\widehat{\beta_{1,M}}$	5.5185 (6.33)	0.9998 (63.17)	0.0051 (75.00)	1.9348 (38.01)
Capital cost Approach 2	2.2508 (32.45)	0.9998 (59.02)	0.0062 (45.00)	2.1669 (33.84)
Systematic risk - Approach 2 $\widehat{\beta_{1,HMLN}}$	3.5136 (17.26)	0.9998 (40.03)	0.0079 (13.00)	3.4443 (17.87)
Systematic risk - Approach 2 $\widehat{\beta_{1,LMHD}}$	7.0586 (2.93)	0.9996 (2.43)	0.0077 (16.00)	7.1837 (2.75)
Systematic risk - Approach 2 $\widehat{\beta_{1,M}}$	0.2436 (88.53)	0.9998 (71.52)	0.0062 (47.00)	0.2526 (88.13)

Note: The vector components are estimated by the M95FF model. We analyse stock companies registered on the WSE in the period from May 1995 through May 2017 that were showing a positive *BV* and with market prices not lower than 2.00 PLN. The risk premium components are estimated by regression (3) using 252 monthly periods, while betas based on regression (5). In Approach 1 betas are estimated on the basis of last 120 months, from period 133 to period 252. In Approach 2 betas are estimated on the basis of a sixty-month rolling window, with rolled step of one month, using the whole tested period of 252 months. The bootstrap procedure is based on 10000 data resamples. Source: own research.

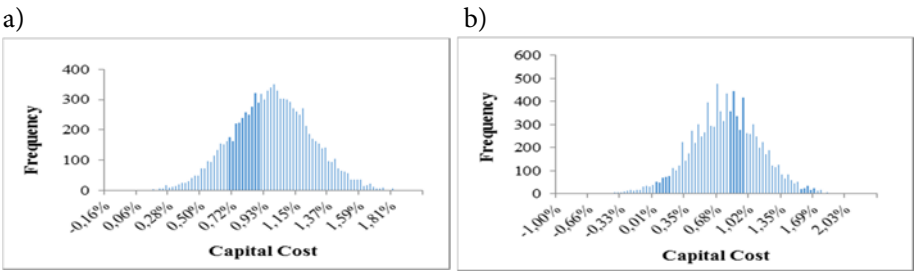


Figure 1. Histograms of bootstrapped capital cost of portfolio 1, formed on the highest *NUM* and the smallest *DEN*: a) Approach 2 - systematic risk components (betas) are estimated using a sixty-month rolling window, with rolled step of one month, using the whole tested period of 252 months, b) Approach 1 - betas are estimated using the last 120 months.

Note: see Figure 1. Source: own research.

Table 5. Percentage values of capital cost of modelled portfolios

Panel A: Classical Fama-French model

I pass: $r_{it} - RF_t = \beta_{i.0} + \beta_{i.M}(RM_t - RF_t) + \beta_{i.HML}HML_t + \beta_{i.SMB}SMB_t + e_{it}$; t=1,…, 252; $\forall i$=1,…, 25

II pass: $r_{it} - RF_t = \gamma_0 + \gamma_M\widehat{\beta_{i.M}} + \gamma_{HML}\widehat{\beta_{i.HML}} + \gamma_{SMB}\widehat{\beta_{..SMB}} + \varepsilon_{it}$; t=1…..252; i=1….. 25

Average betas: $r_{it} = \propto_i + \beta^t_{i,M}RM_t + \beta^t_{i,HML}HML_t + \beta^t_{i,SMB}SMB_t + e_{it}$; t=1….. 60 : 193…252

$Ccap_i = E(RF) + \widehat{\gamma_0} + \widehat{\beta^{av}_{i,M}}\widehat{\gamma_M} + \widehat{\beta^{av}_{i,HML}}\widehat{\gamma_{HML}} + \widehat{\beta^{av}_{i,SMB}}\widehat{\gamma_{SMB}}$; $\beta^{av}_{i,k} = \frac{1}{193}\sum_{t=1}^{193}\widehat{\beta^t_{i,k}}$

γ_0=-3.51%, SH t-stat=-2.35; BV/MV_i portfolios

CAP portfolios	Low Growth	2	3	4	High Value
	Monthly median values				
Small	Portfolio 21) -1.06 (-1.52÷-0.63)	Portfolio 16) -0.32 (-0.69÷0.00)	Portfolio 11) -0.09 (-0.46÷0.24)	Portfolio 6) 0.22 (-0.15÷0.56)	Portfolio 1) 1.45 (0.83÷2.08)
2	-0.35 (-0.64÷-0.10)	-0.28 (-0.60÷-0.01)	-0.08 (-0.32÷0.17)	0.35 (0.00÷0.68)	0.62 (0.19÷1.03)
3	-0.43 (-0.67÷-0.22)	-0.13 (-0.38÷0.10)	-0.27 (-0.58÷0.00)	-0.09 (-0.44÷0.23)	0.62 (0.26÷1.00)
4	-0.38 (-0.60÷-0.17)	-0.39 (-0.66÷-0.16)	-0.41 (-0.72÷-0.14)	0.07 (-0.15÷0.31)	0.37 (0.09÷0.66)
Big	Portfolio 25) -0.50 (-0.78÷-0.23)	Portfolio 20) 0.13 (-0.53÷0.30)	Portfolio 15) 0.15 (-0.15÷0.47)	Portfolio 10) 0.11 (-0.09÷0.33)	Portfolio 5) 0.90 (0.12÷1.64)

Table 5. Percentage values of capital cost of modelled portfolios (cont.)

Panel B: M95FF model

I pass: $r_{it} - RF_t = \beta_{i.0} + \beta_{i.M}(RM_t - RF_t) + \beta_{i.HMLN}HMLN_t + \beta_{i.LMHD}LMHD_t + e_{it}$;
t=1,…, 252;$\forall i$=1,…, 25

II pass: $r_{it} - RF_t = \gamma_0 + \gamma_M\widehat{\beta_{i.M}} + \gamma_{HMLN}\widehat{\beta_{i.HMLN}} + \gamma_{LMHD}\widehat{\beta_{i.LMHD}} + \varepsilon_{it}$

Average betas: $r_{it} = \propto_i + \beta^t_{i,M}RM_t + \beta^t_{i,HMLN}HMLN_t + \beta^t_{i,LMHD}LMHD_t + e_{it}$;
t=1….. 60 : 193…252

$Ccap_i = E(RF) + \widehat{\gamma_0} + \widehat{\beta^{av}_{i,M}}\widehat{\gamma_M} + \widehat{\beta^{av}_{i,HMLN}}\widehat{\gamma_{HMLN}} + \widehat{\beta^{av}_{i,LMHD}}\widehat{\gamma_{LMHD}}$; $\beta^{av}_{i,k} = \frac{1}{193}\sum_{t=1}^{193}\widehat{\beta^t_{i,k}}$

NUM portfolios; γ_0=1.69%; SH t-stat=0.92

Dynamics of increase of financial results

DEN portfolios	Low	2	3	4	High
	Monthly median values				
Small/Cheap	Portfolio 21) -0.42 (-0.87÷0.04)	Portfolio 16) -0.19 (-0.81÷0.44)	Portfolio 11) 0.80 (0.39÷1.23)	Portfolio 6) 0.61 (0.19÷1.06)	Portfolio 1 0.97 (0.44÷1.5)
2	-1.09 (-2.10÷-0.10)	0.00 (-0.32÷0.32)	0.22 (-0.06÷0.52)	0.42 (0.06÷0.77)	0.54 (0.25÷0.84)
3	-0.66 (-1.26÷-0.06)	-0.47 (-0.80÷-0.14)	0.06 (-0.18÷0.32)	0.81 (0.43÷1.20)	0.37 (0.08÷0.70)
4	-1.18 (-1.70÷-0.67)	-1.01 (-1.61÷-0.40)	0.23 (-0.06÷0.52)	-0.12 (-0.53÷0.26)	-0.08 (-0.35÷0.20)
Big/Priced	Portfolio 25) -0.98 (-1.42÷-0.55)	Portfolio 20) -0.97 (-1.46÷-0.51)	Portfolio 15) -0.52 (-0.97÷-0.12)	Portfolio 10) -0.52 (-1.02÷-0.06)	Portfolio 5) -0.58 (-0.96÷-0.18)

Note: In Panel A, 25 FF portfolios are investigated. Quintile portfolios are formed on *BV/MV* and *CAP*. In Panel B, 25 M95FF portfolios are investigated. Quintile portfolios are formed on *NUM* and *DEN*. The corresponding 95 confidence intervals appear in brackets. We analysed stock companies registered on the WSE in the period from May 1995 through May 2017 that are showing a positive *BV* and with market prices not lower than 0.5 PLN for the FF model, and 2.00 PLN for the M95FF model. The risk premium components are estimated using 252 monthly periods. The systematic risk components are estimated using a sixty-month rolling window, with rolled step of one month. The lower and the upper limit of the confidence intervals are calculated using the bootstrap distributions with 10000 iterations.

Source: Own research.

Panel A presents the values of capital cost estimated by the classical FF model. The capital cost assumes the positive estimate of 95% confidence interval for portfolios formed on the highest values of *BV/MV* (fifth quintile of *BV/MV*). For four portfolios of the fourth *BV/MV* quintile, the median confidence intervals are also positive. For the second quintile, the median confidence intervals are negative, except for the portfolio with the highest capitalization. On the other hand, the portfolios formed on the lowest *BV/MV* are characterized by the negative estimate of the lower and upper limits of the 95% confidence interval. Changes in the value of capital cost for portfolios with the growing capitalization are less regular. This can be explained by an insignificant non-zero risk premium due to the *SMB* factor. Nevertheless, the median of portfolios 20, 15, 10 and 5, formed on the biggest capitalization, assumes positive values.

The above results seem to be consistent with the Graham and Harvey (2001), who state that large companies more often estimate capital cost by CAPM. On the other hand, if company is seen by investors as a portfolio of current and future projects and by their real options then the returns are influenced by information reaching investors about the possibility of implementing a portfolio of projects with real options. In that case, the dependence of returns on risk factors may be non-linear, despite the fact that returns of projects without real options are consistent with CAPM, and the capital cost of these projects can be correctly estimated.

In our research, the cost of capital is estimated using portfolio market returns, that is, information that may take into account the impact of real options. The values of the estimated capital cost are influenced by all market information, so it can charge the actual capital cost of the company's projects. Then one can examine the impact of additional risk factors or other ICAPM applications. According to Zhi Da et al. (2012) research, the impact of real options on returns leads to price anomalies, which contradict the pricing consistent with CAPM. Therefore, we attempt to estimate the capital cost of the modelled portfolios using other ICAPM applications.

Panel B presents the values of capital cost estimated by the M95FF model. The capital cost assumes the positive estimate of 95% confidence interval for portfolios placed in the upper right corner of Table 5, that is for portfolios formed on the three highest values of *NUM* and three smallest values of *DEN*. Interestingly, there is a monotonous decrease in the cost of capital for portfolios formed at the highest values of *NUM* and diminishing *DEN* values. The capital cost assumes the negative estimate of 95% confidence interval for portfolios placed in the lower left corner of Table 5, that is for portfolios formed on the lowest values of *NUM* and the biggest values of *DEN*. Portfolios with the highest values of *NUM* are portfolios with the highest dynamics of financial results growth. Portfolios with the biggest values of *DEN* are portfolios with the priced stocks in relation to the book value and earning per share.

The portfolios assuming positive values of capital cost (estimated by the FF model) are commonly called value portfolios, and long-term investments in these portfolios generate high returns. On the other hand, the portfolios assuming negative values of capital cost, called growth portfolios, generate small returns (also see: FF, 1992). Similarly, the portfolios assuming positive values of capital cost, estimated by the M95FF model, are the most attractive for investors and generate above average returns (see: Urbański, 2011). However, as mentioned earlier, the estimated capital cost applies to projects with real options. Therefore, you can assume that, taking into account the market hyperactivity, the estimated capital cost according to the above-mentioned procedures is evaluated too high for value portfolios and portfolios formed on high *NUM* and small *DEN*. Similarly, the estimated capital cost is evaluated too low for growth portfolios and portfolios formed on low *NUM* and big *DEN*.

The reasoning presented in this way explains the negative values of the capital cost, estimated on the basis of portfolio market returns, for growth portfolios and portfolios formed on low *NUM* and big *DEN*.

5. Summary and conclusions

In our paper we present a method of estimation of the cost of capital for characteristic portfolios of stocks registered at the Warsaw Stock Exchange. In order to accomplish this, we apply three selected ICAPM applications: the classical Fama-French model (FF), the modified FF model, denoted as M93FF and proposed by Urbanski (2012), and, finally, the M95FF model, which is proposed in this research. Our analysis starts with 595 stocks from which we eliminate the penny stocks with value smaller than 0.50 PLN for FF and smaller than 2.0 PLN for M93FF and M95FF. Our methods allow generating portfolios that are close the multifactor-efficient. In order to estimate the confidence interval for the cost of capital we apply the bootstrap method. The estimated cost of capital, calculated using the market returns, is related to a hypothetical portfolio of investment projects as seen by the external investor. Such a portfolio is a combination of undergoing and planned projects, weighted with selected real options. According to the proposed procedure, the estimated cost of capital may be a valuable indicator for portfolio managers. Moreover, it allows one to estimate the capital returns for investors. However, such an estimate of the capital cost cannot be considered as a benchmark for making decisions regarding new investment projects of stock companies.

Our research leads to the following conclusions related with the estimation of the capital cost:

1) The ICAPM application allows one to estimate the *Ccap* of investment projects of stock companies from the perspective of an external investor.

2) The necessary condition for appropriate estimation of the *Ccap* is precise estimation of the risk premium. Here, the application of ICAPM should generate multifactor-efficient portfolios.
3) Application of the classical FF model and its two modifications (M93FF and M95FF) for stocks coming from the WSE allows one to generate portfolios that are close to multifactor-efficient, provided that the penny stocks are eliminated.
4) The *Ccap* is estimated using two different approaches. In the first approach, the betas are estimated using the most recent 120 months of observations. In the second approach, the betas are estimated using 60-month moving windows that roll with one-month step.
5) The application of the bootstrap method allows one to approximate the distribution of the systematic risk and the risk premium components as well as the distribution of the cost of capital.
6) The estimated *Ccap* is related to the project portfolio and open positions of real options related to these projects.
7) The estimated *Ccap* assumes positive values for value portfolios and portfolios formed on high values of *NUM* and low values of *DEN*.
   a) The *Ccap* of value portfolios varies from 0.37%±0.28% to 1.45±0.62% per month. The average width of the confidence interval for the *Ccap* is about 0.98%.
   b) The *Ccap* of portfolios formed on high *NUM* and low *DEN* varies from 0.37%±0.28% to 0.97%±0.53% per month. The average width of the confidence interval for the *Ccap* is about 0.78%.
8) The estimated *Ccap* takes negative values for growth portfolios and portfolios formed on low values of *NUM* and high values of *DEN*.
   a) The *Ccap* of growth portfolios varies from -0.35%±0.27% to -1.06%±0.45% per month. The average width of the confidence interval for the *Ccap* for such portfolios is about 0.56%.
   b) The *Ccap* of portfolios formed on low *NUM* and high *DEN* varies from -1.18%±0.52% to -0.47±0.33% per month. The average width of the confidence interval for the *Ccap* for such portfolios is about 1.06%.
9) In order to estimate the *Ccap* of stocks without real options one needs the components of the option-adjusted risk premium and option-adjusted systematic risk.

Acknowledgements

This work supported by the National Science Centre, Poland (Research Grant 2015/19/B/HS4/01294) is gratefully acknowledged.

REFERENCES

BANZ, R.W., (1981). The Relationship between Return and Market Value of Common Stock, Journal of Financial Economics, Vol. 9 (1), pp. 3–18.

BERK, J., GREEN, R., NAIK, V., (1999). Optimal investment, growth options, and security returns, Journal of Finance, Vol. 54, pp. 1553–1607.

BERNARDO, A., CHOWDRY, B., GOYAL, A., (2007). Growth options, beta, and the cost of capital, Financial Management, Vol. 36, 5–17.

BHANDARI, L. CH., (1988). Debt/Equity Ratio and Expected Common Stock Returns: Empirical Evidence, Journal of Finance, Vol. 43, 2, pp. 507–528.

CARHART, M. M., (1995). Survivor bias and persistence in mutual fund performance. (Unpublished doctoral dissertation), Graduate School of Business, University of Chicago.

COCHRANE, J., (2001). Asset Pricing. Princeton University Press, Princeton, New Jersey.

CZAPKIEWICZ, A., WÓJTOWICZ, T., (2014). The four-factor asset pricing model on the Polish stock market, Economic Research-Ekonomska Istraživanja, 27, 1, pp. 771–783.

EFRON, B., TIBSHIRANI, R. J., (1993). An Introduction to the Bootstrap, Chapman and Hall CRC, New York.

FAMA, E. F., (1996). Multifactor Portfolio Efficiency and Multifactor Asset Pricing, Journal of Financial and Quantitative Analysis, Vol. 31, 4, pp. 441–465.

FAMA, E. F., French, K. R., (1992). The Cross-Section of Expected Stock Returns, Journal of Finance, Vol. 47, 2, pp. 427–465.

FAMA, E. F., FRENCH, K. R., (1993). Common Risk Factors in the Returns on Stock and Bonds, Journal of Financial Economics, Vol. 33, 1, pp. 3–56.

FAMA, E. F., FRENCH K. R., (1995). Size and Book-to-Market Factors in Earnings and Returns, Journal of Finance, Vol. 50, 1, pp. 131–155.

FAMA, E. F., FRENCH K. R., (2015). A five-factor asset pricing model, Journal of financial Economics, Vol. 116, pp. 1–22.

FERSON, W., LOCKE, D. H., (1998). Estimating the cost of capital through time: An Analysis of the Sources of Error, Management Science, 44, 4, pp. 485–500.

GIBBONS, M. R., ROSS, S. A., SHANKEN, J., (1989). A Test of the Efficiency of a Given Portfolio, Econometrica, Vol. 57, 5, pp. 1121–1152.

GRAHAM, J. R. HARVEY, C. R., (2001). The theory and practice of corporate finance: evidence from the field, Journal of Financial Economics, Vol. 60, pp. 187–243.

JEGADEESH, N., TITMAN, S., (1993). Returns to Buying Winners and Selling Losers: Implications for Stock Market Efficiency, Journal of Finance, Vol. 48, 1, pp. 65–91.

JAGANNATHAN, R., WANG, Z., (1996). The Conditional CAPM and the Cross-Section of Expected Returns, Journal of Finance, Vol. 51, 1, pp. 3–53.

KAN R., ZHANG C., (1999). Two-Pass Tests of Asset Pricing Models with Useless Factors, Journal of Finance, Vol. 54, 1, pp. 203–235.

LAHIRI, S., (2003). Resampling Methods for Dependent Data. Springer-Verlag Inc., New York.

LAKONISHOK, J., SHAPIRO, A.C., (1986). Systematic Risk, Total Risk, and Size as Determinants of Stock Market Returns, Journal of Banking and Finance, Vol. 10, 1, pp. 115–132.

LAKONISHOK, J., SHLEIFER, A., VISHNY, R., (1994). Contrarian investment, extrapolation, and risk, Journal of Finance, Vol. 49, pp. 1541–1578.

LETTAU, M., LUDVIGSON, S., (2001). Resurrecting the (C) CAPM: A Cross-Sectional Test when Risk Premia Are Time-Varying, Journal of Political Economy, Vol. 109, 6, pp. 1238–1287.

LINTNER, J., (1965). The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets, Review of Economics and Statistics, Vol. 47, 1, pp. 13–37.

REINGANUM, M. R., (1981). A New Empirical Perspective on the CAPM, Journal of Financial and Quantitative Analysis, Vol. 16, 4, pp. 439–462.

ROSENBERG, B., REID, K., LANSTEIN, R., (1985). Persuasive Evidence of Market Inefficiency, Journal of Portfolio Management, Vol. 11, pp. 3, 9–16.

SHANKEN, J., (1985). Multivariate Tests of the Zero-Beta CAPM, Journal of Financial Economics, Vol. 14, pp. 327–348.

SHANKEN, J., (1992). On the Estimation of Beta-Pricing Models, The Review of Financial, Vol. 5, 1, pp. 1–33.

SHARPE, W. F., (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk, Journal of Finance, Vol. 19, 3, pp. 425–442.

URBAŃSKI, S., (2011). Modelowanie równowagi na rynku kapitałowym – weryfikacja empiryczna na przykładzie akcji notowanych na Giełdzie Papierów Wartościowych w Warszawie. [Pricing modelling on capital market - empirical verification on the example of stocks listed on the Warsaw Stock Exchange], Prace Naukowe Uniwersytetu Ekonomicznego w Katowicach, Katowice, Poland.

URBAŃSKI, S., (2012). Multifactor Explanations of Returns on the Warsaw Stock Exchange in Light of the ICAPM, Economic Systems, Vol. 36, pp. 552–570.

URBAŃSKI, S., JAWOR, P., URBAŃSKI, K., (2014). The Impact of Penny Stocks on the Pricing of Companies Listed on The Warsaw Stock Exchange in Light of the CAPM, Folia Oeconomica Stetinensia, Vol. 14, 2, pp. 163–178.

URBAŃSKI, S., (2015). The Impact of Speculation on the Pricing of Companies Listed on the Warsaw Stock Exchange in Light of the ICAPM, Managerial Economics, Vol. 16, 1, pp. 91–111.

URBAŃSKI, S., (2017). Comparison of modified and classic Fama-French model for the Polish market, Folia Oeconomica Stetinensia, Vol. 17, 1, pp. 80–96.

WELCH, I., (2008). The Consensus Estimate for the Equity Premium by Academic Financial Economists in December 2007, Unpublished manuscript, Brown University, Providence, United States.

ZHI, D., GUO, R. J., JAGANNATHAN, R., (2012). CAPM for estimating the cost of equity capital: Interpreting the empirical evidence, Journal of Financial Economics, Vol. 103, 1, pp. 204–220.

ZARZECKI, D., BYRKA-KITA, K., WIŚNIEWSKI, T., KISIELEWSKA, M.,, (2004-2005). Test of the Capital Asset Pricing Model: Polish and Developed Markets Experiences, Folia Oeconomica Stetinensia, Vol. 3–4, 11–12, pp. 63–85.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 95–122, DOI 10.21307/stattrans-2020-006
Submitted – 10.12.2018; accepted for publishing – 09.12.2019

KLEMS growth accounting implemented in Poland[1]

Dariusz Kotlewski[2], Mirosław Błażej[3]

ABSTRACT

The aim of the article is to present the main body of the KLEMS growth accounting recently implemented in Poland. The works on the KLEMS productivity accounting in Poland started in 2013 and focused on areas such as the development of methodology and the availability and assessment of data. These efforts enabled preparing KLEMS data sets pertaining to the Polish economy and moreover proved that unavailable data can be effectively estimated. Additionally, interesting but complex and debatable results were obtained, such as labour hoarding together with remunerations' freezing around the 2009 crisis, accompanied by a natural drop in the capital contribution growth and an increase in the MFP contribution, which most probably indicated effective reorganizations in the economy. In the years 2012–2014, increasing labour and capital contributions did not fully translate into gross value added growth, which led to negative MFP growths, as these are calculated residually. This, however, changed completely in the last two years of the time span covered by the research, namely in 2015-2016. An industry-level analysis became also possible, showing that the Polish economy was developing dynamically and undergoing intensive modernisation, which was obtained, however, with a debatable contribution of the State. To study the debatable features of the Polish economy in a greater detail, a further decomposition of the labour factor growth into four sub-factor contributions instead of two sub-factor contributions was performed. This additional analysis confirmed that labour hoarding phenomenon specific for Poland contributed to a softer impact of the 2007–09 financial crisis on this country's economy.

Key words: gross value added, decomposition, production factors, KLEMS, productivity.

1. Introduction

The aim of the article is to present KLEMS productivity growth accounting recently being implemented in Poland and discuss it. Although Poland is present in various

[1] Disclaimer: the views presented in this article are those of the authors and do not represent the official standpoint of Statistics Poland.

[2] Warsaw School of Economics, Dariusz Kotlewski is also an employee in Statistics Poland.
E-mail: dariusz.kotlewski@gmail.com, dariusz.kotlewski@sgh.waw.pl, d.kotlewski@stat.gov.pl.
ORCID: https://orcid.org/0000-0003-1059-7114.

[3] Statistics Poland. E-mail: m.blazej@stat.gov.pl.

releases of EU KLEMS database, no decomposition of gross value added growth or gross output growth into factor contributions and MFP contribution has been ever performed there because of missing input data (with the exception of 2007 EU KLEMS release, presently outdated) – the reason is that no sufficient data are being sent to Eurostat on the one hand (although theoretically they could be sent, which is a matter of co-operation agreements within Eurostat) and on the other hand that there are data which need to be imputed innovatively, since they are not straightforwardly available in Poland too. A growth accounting for Poland with decomposition as above mentioned has been performed by NBP[4] appointed researchers, based on a slightly different methodology (Gradzewicz et al., 2014, 2018), but not at the sectoral and industry level. As far as the authors of this article know, no one else has ever performed a decomposition of the above-mentioned kind at industry level for Poland (with the exception of the above mentioned release[5]).

The methodology framework based on a gross value added decomposition is outlined in the second section. In the third section, these results are discussed in an attempt to interpret them. In the fourth section a sample analysis at industry level is presented. In the fifth section, a developed labour factor decomposition is advanced allowing for more analytical insights in the Polish economy. Despite the specific data processing hurdles accompanying the works on the Polish KLEMS accounting, ample results of good quality were achieved (presently, they are available on Statistics Poland web page[6]). As they are to a large extent debatable, these outcomes remain open to further analyses and discussion. New avenues for developing KLEMS accounting and the conclusion are presented in the final section. The article is of a wide synthetic nature. It deals both with theoretical matters and technicalities, and presents also the results of KLEMS accounting in Poland with some interpretation.

2. Basic methodology

The basic methodology follows in general the growth accounting methodology developed by Dale W. Jorgenson and associates, as outlined in Jorgenson (1963), Jorgenson & Griliches (1967), Jorgenson, Gollop and Fraumeni (1987), Jorgenson

[4] The National Bank of Poland which is the Polish Central Bank.

[5] The EU KLEMS data set release of 2007 includes a decomposition for Poland with labour services' contribution being subdivided into hours worked and labour composition contributions, but with no subdivision of capital services' contribution into ICT and non-ICT capital contributions. This release covers the period of 1996–2004, therefore just preceding the present study time span. For the 2007 EU KLEMS release data were often extensively imputed (Timmer et al., 2007b, pp. 121–29) to a far greater degree than in the present study due to greater data shortages. The possible comparisons between the two studies can possibly be an avenue for further analysis.

[6] KLEMS Economic Productivity Accounts, https://stat.gov.pl/en/experimental-statistics/klems-economic-productivity-accounts/

(1989) and Jorgenson, Ho and Stiroh (2005)[7]. This methodology has been summarized by Timmer et al. (2007), and O'Mahony and Timmer (2009) for the EU KLEMS[8]. For Poland it has been developed and presented in Kotlewski & Błażej (2016, 2018a and 2018b)[9]. Hereinafter, only the basic idea and structure of the accounts is presented. It is based on the standard growth accounting decomposition of output into the contribution of input factors and MFP:

$$\Delta \ln Y_{jt} = \bar{v}_{jt}^{X} \Delta \ln X_{jt} + \bar{v}_{jt}^{K} \Delta \ln K_{jt} + \bar{v}_{jt}^{L} \Delta \ln L_{jt} + \Delta \ln A_{jt}^{Y} \quad (1)$$

where Y is the gross output, X – intermediate consumption, K – capital services[10], L – labour services[11] and where A^Y stands for multifactor productivity. These values are subscripted by j for industries and t for years. $\bar{v}$ with appropriate subscripts are average value shares[12] of the individual factors in the gross output (defined in the superscripts by X, K and L) for two discrete time periods t-1 and t, which are calculated through linear interpolation as $\bar{v} = (v_{t-1} + v_t)/2$ (for simplicity the subscripts of (1) have been omitted here). Since the growth of A^Y is residually calculated, the equation (1) is always met. As for most of the EU KLEMS countries performing the KLEMS growth accounting, the methodology has been reduced to a gross value added decomposition following the standard equation[13]:

$$\Delta \ln V_{jt} = \bar{w}_{jt}^{K} \Delta \ln K_{jt} + \bar{w}_{jt}^{L} \Delta \ln L_{jt} + \Delta \ln A_{jt}^{V} \quad (2)$$

[7] In the preparatory works, the OECD growth accounting methodology was studied as well for possible insights; see: OECD (2001, 2009 and 2013), Wölfl (2007).

[8] See also the large overview of the subject: Jorgenson, ed. (2009).

[9] The first of these works is addressed to the Polish audience not fully acquainted with KLEMS accounting, therefore methodological details were amply presented, whereas the second one is more focused on Polish specificities, more essential for the foreign audience. The third of these works presents development of the labour factor decomposition into additional sub-factors.

[10] It is assumed that the values of capital services are proportional to the values of capital stocks if these are separated into different kinds of capital stocks at industry level, which means that although capital stocks and capital services are different entities, their growths are the same at this level. These different kinds of capital stocks are then aggregated with the use of Törnqvist quantity index at industry level. Following: OECD (2001, 61), Timmer et al. (2007a, 32-33), OECD (2009, 60) and Timmer et al. (2010, eq. (3.6)).

[11] It is assumed that the values of labour services are proportional to the amounts of physical work engaged (in hours worked) if these are separated into different kinds of labour according to age, education attainment and sex.

[12] All value shares throughout the paper were taken from the national accounts, but they were adjusted for the self-employed before being used in calculations.

[13] This value added based growth accounting is not neutral to the substitution between intermediate inputs, and the production factors labour and capital, but this problem is considered to be less severe than vertical firm consolidation differences between different countries, when their growth accounting are compared. Moreover, there is the problem of producing intermediate inputs deflators that are required for gross output based growth accounting, which is a quite universal unsolved problem (see: EU KLEMS database). However, a separate methodology has been lastly developed for Poland to address this issue, which allowed to include intermediate inputs according to formula (1). This issue, however, does not contradict the present analysis and could be the subject of a separate paper.

where V is the gross value added and where A^V stands for multifactor productivity (further referred to as MFP[14]) in the gross value added based decomposition $\overline{w}$ with appropriate subscripts are average value shares of production factors' services in the gross value added (defined in the superscripts as K and L) for two discrete time periods *t-1* and *t*, which are calculated through linear interpolation in a similar way as $\overline{v}$ for the previous formula (1). The other symbols are the same as in equation (1). Replacing the decomposition (1) by (2) eases some data problems and the international comparability of the individual countries[15]. In practice the contribution of multifactor productivity $\Delta \ln A_{jt}^V$ is residually calculated as the subtraction between the other values, so the equation (2) is always met, similarly to equation (1). There is no need, therefore, to directly measure the levels of A.

The capital factor contribution (understood as capital services inputs) has been decomposed into two sub-factors contributions as follows:

$$\overline{w}_{jt}^K \Delta \ln K_{jt} = \overline{w}_{jt}^{KIT} \Delta \ln KIT_{jt} + \overline{w}_{jt}^{KNIT} \Delta \ln KNIT_{jt} \quad (3)$$

where *KIT* stands for ICT capital and *KNIT* for non-ICT capital services[16], treated as separate factors indeed, which is expressed also by their different shares (appropriately superscripted). In practice one of the three contributions, usually the non-ICT capital one, is residually calculated as the subtraction between the other values in the equation (3), in order to avoid mathematical tool problems (so the equation (3) is always met).

The labour factor contribution (understood in the standard KLEMS methodology as labour services' contribution) has been decomposed somehow differently as follows:

$$\overline{w}_{jt}^L \Delta \ln L_{jt} = \overline{w}_{jt}^L \Delta \ln H_{jt} + \overline{w}_{jt}^L \Delta \ln Q_{jt} \quad (4)$$

where L stands for the labour factor understood as hours worked aggregated with the use of Törnqvist quantity index on the one hand or compensations aggregated with the use of this index on the other hand[17], H stands for the (straightforward) sum of hours worked on the one hand or hours worked aggregated with the use of Törnqvist quantity index on the other hand[18], and Q for labour quality. The firsts of these options result in

[14] It can be considered as a variant of total factor productivity (TFP).

[15] Because of the vertical integration of firms of different intensity in different countries, which hinders international comparability of the countries, as far as the intermediate consumption is considered.

[16] Symbols taken from Timmer et al. (2007).

[17] Over standard KLEMS 18 kinds of labour according to sex, three age groups and three education attainment levels (2 X 3 X 3 = 18). The first of these options is a standard KLEMS procedure, the second is an option explored additionally in Statistics Poland.

[18] As above. Experiments done in Statistics Poland have finally led to a developed procedure for labour factor contribution decomposition presented in more details in Section 4.

labour quality contribution understood as labour composition contribution, whereas the seconds – as labour compensation change contribution. They are all treated as a single factor, which is expressed by their same share $\overline{w}_{jt}^{L}$ as for L. This difference in comparison with the capital factor decomposition (3) is however of no importance as far as the linear additivity of the sub-factor contributions in the gross value added growth is considered[19]. Here the growths of the so-called sub-factors sum up to the growth of the entire labour factor as:

$$\Delta \ln L_{jt} = \Delta \ln H_{jt} + \Delta \ln Q_{jt} \tag{5}$$

In practice, the labour quality related expressions in equations (4) and (5) are residually calculated through subtractions between the expressions related with L and H (this is possible because all expressions are in percentage points). So, there is no need to directly measure the value of Q and therefore the equations (4) and (5) are always met.

All data have been calculated after being converted initially into 2005 prices and presently into 2010 prices[20], following the same change in Eurostat transmission tables that happened during the works on Poland KLEMS. This change was found to be only technical and of little impact on the accounts. Information on data processing by other countries was studied for comparison and reference in Gouma and Timmer (2013). During the works the ESA'95 Eurostat system changed also into the ESA 2010 system[21], but this change was found to be of even lower importance than the above-mentioned one[22]. The problem of qualitative growth as discussed, e.g. by Diewert (1993), Hausman (2003) and Hulten (2009, 19–21) has been for the time being ignored, which is the general practice in the present KLEMS accounting[23]. However, one may argue that quality deflators for ICT capital services are to some extent considered, as the structure

[19] The equation (15) in O'Mahony and Timmer (2009, F378) expresses also this difference, but instead of the term "labour quality" (Q) used here, the terms "labour composition" (LC) was used there, more narrowly defined as only the standard KLEMS procedure. More details are presented in Section 4.

[20] This can be done directly comparing each year with the base year or through chaining (see Schreyer (2004) and Milana (2009). The results of the two methods are slightly different. As in other EU KLEMS countries chaining was applied to establish the base year. Establishing the base year is not necessary to perform KLEMS type growth accounting (pairs of years in the same prices are good enough), but such is the way the other EU KLEMS countries data were processed.

[21] These are European equivalents of SNA'93 and SNA 2008 systems.

[22] It concerns only the ICT capital services where mixed data on productive capital stocks had to be used. A test for shortened time series covering both ESA'95 and ESA 2010 comparing results arising from the use of these two systems showed that differences are negligible.

[23] And there seems to be no perspective to introduce it soon. It is quite controversial to use hedonic quality assessments based on price differentials. High quality growth industries should theoretically expand at the expense of low quality growth industries, but this is not necessarily the case. In the case of ICT industries it might be that prices are falling and product quality rising.

taken from SUT tables (ICT services related with software providing) was used to distribute productive capital stocks by industries[24].

The works on Poland KLEMS have a specific feature, which is open for further discussion. All results have been calculated and presented in four versions indicated on the Statistics Poland web page by capital letters A-D:

- A – Capital without residential capital (without dwellings), labour quality understood as labour composition
- B – Capital with residential capital (with dwellings), labour quality understood as labour composition
- C – Capital without residential capital (without dwellings), labour quality understood as labour compensation level
- D – Capital with residential capital (with dwellings), labour quality understood as labour compensation level

The inclusion of residential capital in the Polish KLEMS accounting is controversial because of the still opaque (to be understood as not fully liquid in economic terms) Polish dwelling market. But for international comparison it is preferable because all the other KLEMS countries (which are generally not transformation countries and therefore with well capitalised, liquid and transparent dwelling markets) do include it. Therefore, it was decided to make the calculations in both ways (to satisfy both contenders). There are perceivable difference on graphs, between these two options, although of no decisive importance[25].

The other dichotomy is the possibility of different understanding of the labour quality sub-factor, as solely related to labour composition change or, alternatively, as being entirely translatable into changes in labour compensation levels. Here also both ways of calculations have been performed. For international comparisons the versions B is the appropriate one. In Kotlewski & Błażej (2016 and 2018a) and in Kotlewski (2017) details on how data limitations have been overcome are extensively presented.

Figure 1 shows the results at the aggregate level for the above-mentioned four versions A to D. It can be seen that the differences are not of a fundamental nature. However, there are some benefits from doing KLEMS accounting in the four versions. The lower graphs (C and D) show the impact of remuneration level increase on the

[24] It was, as it seems quite sensibly, assumed that the ICT productive stocks (used to compute ICT capital services' contributions) are proportional to the related ICT services, such as software providing. Both ICT capital assets and related services provided on the market are increasing in quality and decreasing in prices, to some extent in parallel.

[25] The industry of residential building is usually excluded from market economy, because the output of this industry mostly reflect imputed housing rents instead of firms' production (Timmer et al., 26). Such is the case, e.g. in Germany. However, in Poland the situation is different since houses are generally owned by the users. Excluding housing from market economy is therefore also a controversial issue, although the Statistics Poland's KLEMS database provides also the data for this market economy aggregation. In Poland, houses are rather a kind of product sold in instalments, therefore excluding it from capital stocks finds supporters.

accounts, i.e. the years when this impact was huge can be identified by comparing the differences in the MFP level between these graphs and the graphs A and B situated above. Similarly, the years when the impact of dwellings' growth was huge can be identified by comparing the differences in the MFP level between graphs A and C with graphs B and D.

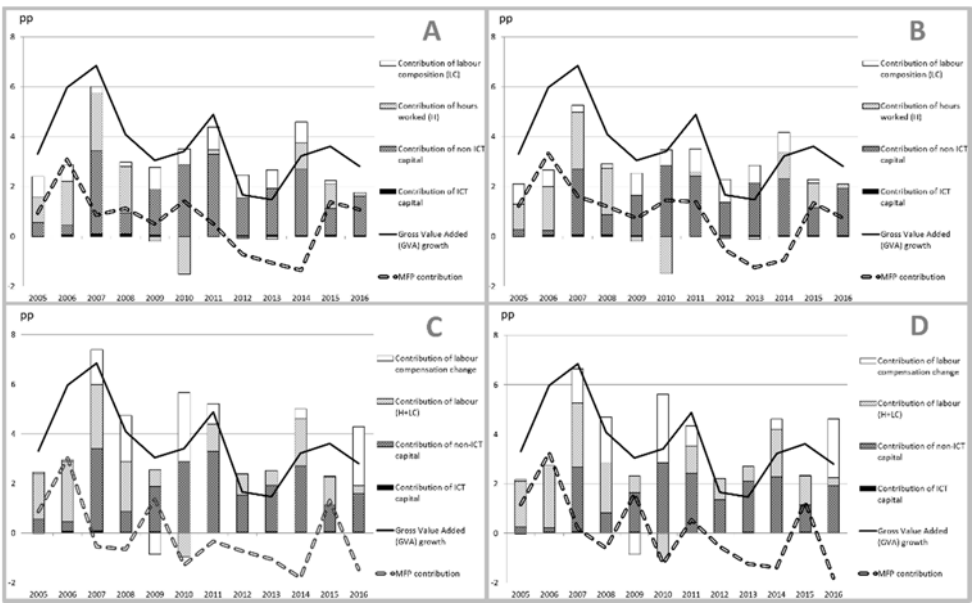


Figure 1. Decomposition of aggregate value added growth in Poland in four versions

Source: Statistics Poland web page.

3. Preliminary interpretation of aggregate results

The general picture of the Polish economy in the light of the present KLEMS accounting methodology, which is in line with other countries performing KLEMS productivity accounting (version B as indicated in section 2.), is presented in Figure 2. What can be seen is that the peaks of MFP contribution in 2006 and 2010 precede the peaks of aggregate gross value added growth in 2007 and 2011, which suggests at first that the increase in productivity due to some reorganizations or modernizations in the economy led to profit margin increases, launching forward the economy and that these impulses remained active for some time after.

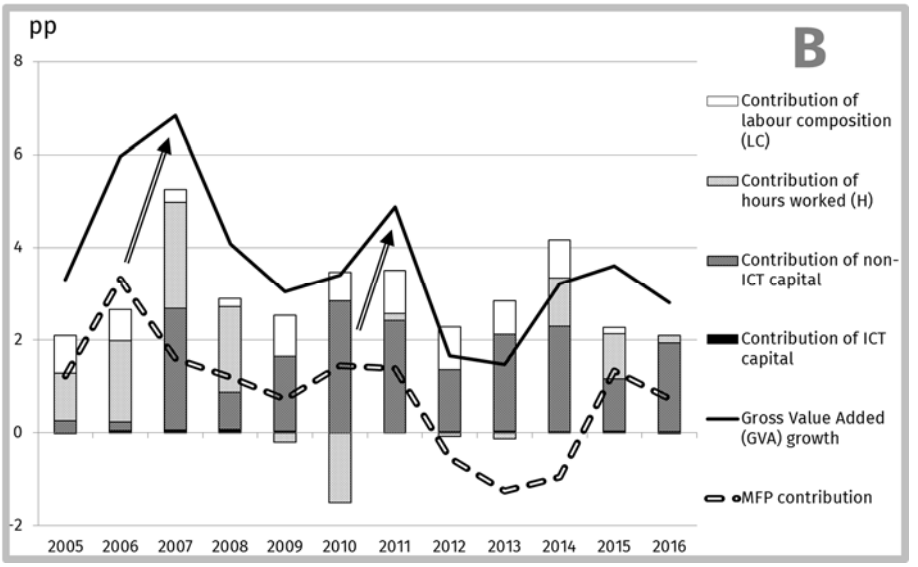


Figure 2. Decomposition of aggregate value added growth in Poland – version B conformable with EU KLEMS

Source: Own contribution based on Statistics Poland web page.

However, the evolution of the gross value added growth in Poland is also not in contradiction to its evolutions in some other economies, i.e. there was a slump around the year 2009 during the trough of the world-wide financial crisis, and the above-mentioned peaks before and after are somehow correlated with the world's economy as well, which suggests rather a business cycle effect. The mentioned slowdown has not, however, led the Polish economy into recession, which is its distinctive feature. These results correspond well with the results presented, e.g. in Havik, Leitner, Stehrer (2012) as far as the time series do overlap each other, but additionally MFP data are provided. Since yearly representations of growth accounting can be misleading, because it might be difficult to focus on long run effects only and eliminate the short-term demand effects, a cumulative growth analysis is provided in the fifth section. Exact quantitative yearly data are available on the Statistics Poland website[26].

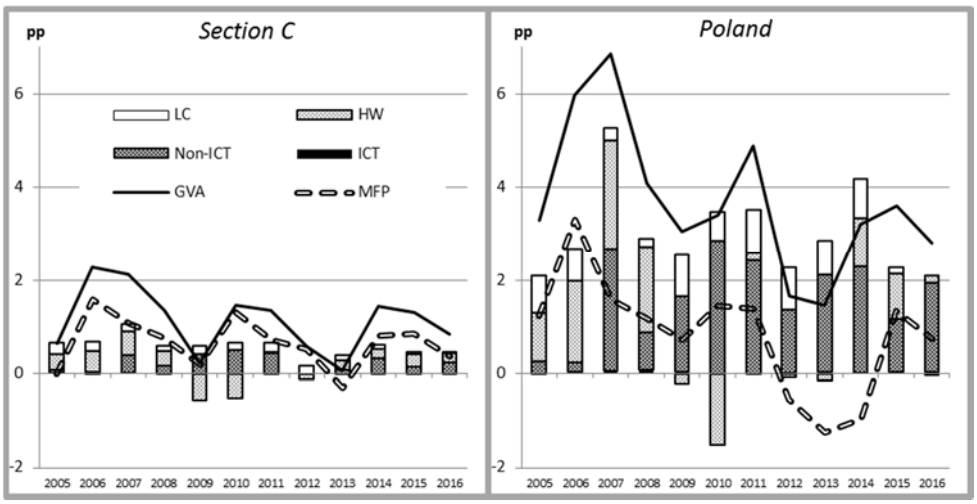
For the years 2012–2014 it can be observed that MFP contribution is negative. The possible explanation of this fact may be that the direct contribution to gross value added growth of growing capital stocks (at industry level and therefore capital services at the aggregate level within KLEMS methodology as explained before) is in relation with the growing demand for capital goods, which in Poland are largely imported. The increase in investments has led to "demand evasion" abroad through increased net imports of

[26] KLEMS Economic Productivity Accounts, https://stat.gov.pl/en/experimental-statistics/klems-economic-productivity-accounts/.

that capital goods, therefore we have a growth of capital contribution in the accounts, not accompanied, however, by a parallel domestic growth of demand for domestic capital goods – the contribution of capital services temporarily raises but with no additional raise in gross value added. This fact may be reflected in lower and even negative MFP contribution, because it is computed residually in the KLEMS methodology. This process was facilitated in those years by the fact that the Polish trade balance improved and even reached positive numbers in 2015, so the economy had finance to invest domestically and of course some role in increasing investments can be attributed to EU funds as well. In the light of developing KLEMS accounting this phenomenon could be better explained using the concept of Global Value Chains (GVC), which would allow to trace exactly and unequivocally (Timmer et al., 2013) the "migration" of gross value added growth together with some part of the residual contribution of MFP between the countries. This effect can also be considered as a tool problem or short-run demand disturbance (if MFP contribution was not calculated residually but computed somehow directly, it would not occur probably).

The last two years of the covered span of time show a huge increase in the contribution of MFP to gross value added growth, which is difficult to be explained solely by the business cycle upturn. It can be observed, however, that this phenomenon is accompanied by a fall in the contribution of the labour factor, i.e. both labour composition and hours worked, which is being offset, as far as its possible negative impact on economic growth is considered, by a surge in productivity growth represented by the MFP contribution.

KLEMS economic productivity accounting allows also to undertake an analysis of the contributions of industry aggregations to aggregate gross value added growth and to decompose this contribution into factor contributions as defined in KLEMS accounting. It can be observed that the NACE 2 classification industries, of which the contribution to aggregate gross value added growth is generally the most important, are industries belonging to section C (manufacturing) of this classification. Its evolution and the evolutions of its contributing factors, as shown in Figure 3 (left-hand side graph), exhibit a similar pattern to the pattern of evolutions observed for the aggregate economy (right-hand side graph).



Legend: ICT – ICT capital contribution, Non-ICT – non-ICT capital contribution, HW – hours worked contribution, LC – labour composition contribution, GVA – gross value added growth, MFP – multifactor productivity contribution.

Figure 3. Contribution of GVA growth of section C and its decomposition compared to GVA growth decomposition for the aggregate Polish economy

Source: Own contribution based on Statistics Poland web page data.

However, the falling trends of gross value added growth, particularly of MFP contribution in section C, are less conspicuous than for the aggregate economy. This indicates that section C supports economic growth. The role of manufacturing seems to strengthen particularly as far as MFP in section C is considered, since its contribution to section C growth is relatively much greater in comparison with the other factors' contributions than for the aggregate economy. This goes in line with the concept that reindustrialization of the Polish economy is on its way.

A lot of interest is due to the labour market, which in Poland has behaved very specifically in comparison with the other countries that have performed KLEMS accounting, as shown in Figure 2, and some argue that this fact lies behind the good comportment of the Polish economy during the trough of the world's financial crisis in 2009. This, of course, is very debatable.

However, we can imply that this is also due to a labour hoarding phenomenon of a both rational and psychological origin and it is caught by the right-hand side graph of Figure 6 later on in the article. The general government and firms refrained from laying off employees until the onslaught of the financial crisis passed away[27]. At the same time

[27] Although typical (see: Timmer et al., Spring 2011, 9–10), this labour hoarding lasted in Poland until 2010, that is about a year longer than in the other EU KLEMS countries.

the highest remunerations were curbed down in the Government administration. From 2008 wage increases in the Government administration were also officially frozen because of the regulations on budget deficit and debt that are in force in Poland[28]. In the next year (2010) the low base helped to reach a high labour quality contribution growth, when the labour market was freed again to some degree.

The data on KLEMS accounting posted on the Statistics Poland web page indicate that the NACE 2 sections C, G, J, K, L and M-N can be considered as trend settlers for the labour market. The sections O-U (representing mainly the wide public sector entities) contribute to some degree to the entire economy labour productivity evolution.

As far as the capital factor contribution to aggregate gross value added growth is considered, it can be seen in Figure 4 that it generally follows the evolution of the entire economy (i.e. the business cycle), but its relative contribution share increased in the long run and became prevalent in the years 2013–2014. In some countries this led to reduced capital productivity[29], but not in Poland. From some other analyses it is known that the ratio of residual capital income[30] to capital stock value is continuously (over the period of this analysis) and still growing. Following the standard theory, as long as the "golden rule"[31] shall be maintained there should not be any decrease in capital productivity. However, the relative role of capital contribution to gross value added growth decreased importantly in the years 2015v2016, which leads to a possible conclusion that a similar process of substitution happened to labour factor contribution substitution by MFP contribution. Generally, a decrease in factor contributions seems to be largely offset by an outstanding MFP contribution increase, much greater than expected to be induced solely by the business cycle itself.

[28] Timmer et al. (Spring 2011, 9) mention about a productivity going up together with massive layoffs. In Poland, instead of layoffs, remunerations were slashed in 2009.

[29] In other emerging markets the economic growth was similarly to Poland „increasingly investment driven with negative TFP", (see: Bart van Ark, 2016).

[30] Calculated by subtracting labour total compensation from gross value added (Lewandowski et al., 2015).

[31] According to the neoclassical theory of economic growth, an economy should invest about 20% of its GDP.

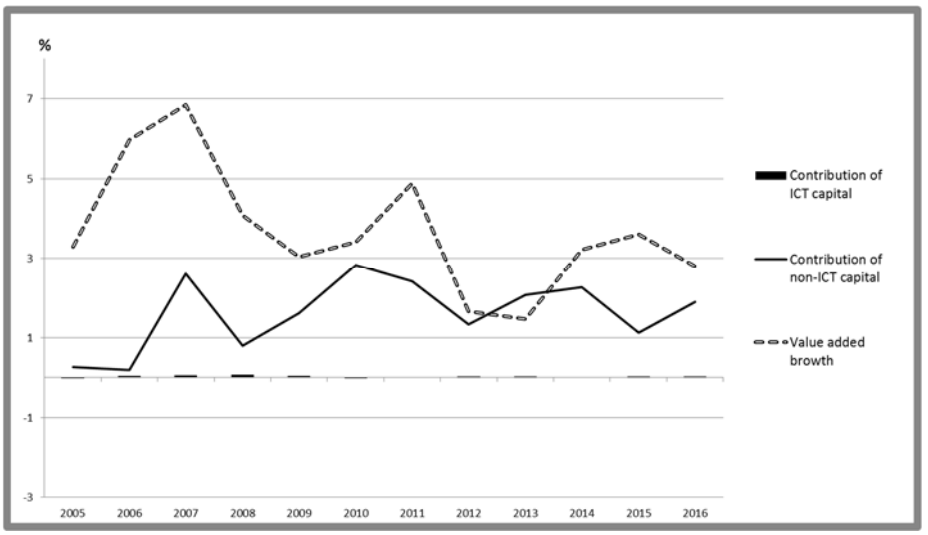


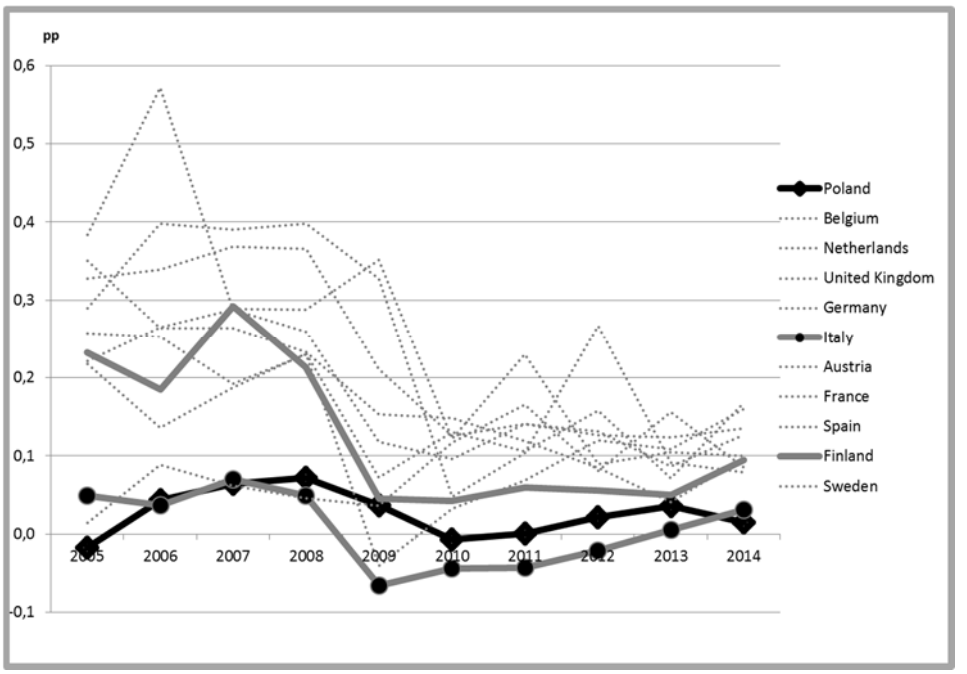
Figure 4. Capital sub-factors' contributions to aggregate GVA growth in Poland

Source: Own calculations based on Statistics Poland web page data.

However, this positive feature in general (with the exception of uncertain last two years 2015–2016), concerning the non-ICT capital services and therefore also the entire capital services as understood in KLEMS accounting, does not concern the ICT capital services, of which the contribution is almost imperceptible (see Figure 4). In comparison with some other countries this could be explained also by the fact that Poland is rather an importer of ICT goods (although Poland is an exporter of quite many ICT components and some software). But there can be also some other explanations that will be examined in Figure 5, where the scale on the vertical axis has been magnified to ease the observations.

As it can be seen in Figure 5, the ten countries that traditionally perform a decomposition of the KLEMS growth accounting type (presented on the EU KLEMS internet platform) greatly differ in the contribution of ICT capital to aggregate gross value added growth, although these are all developed and similar Western European economies, which suggests the possibility of a methodological problem[32].

[32] This has been discussed in Timmer et al. (2007a and 2007b), the definition of ICT capital is very different in the mentioned countries. The ICT capital can be solely consisting of computers and software, and some top telecommunication devices, or include many peripheral components and infrastructure that is related.



Note: 2015–2016 data for countries other than Poland are not entirely available presently, therefore the graph ends in 2014.

Figure 5. Comparisons of the contribution of ICT capital to aggregate value added growth between Poland and some other countries performing KLEMS accounting

Source: Own contribution based on Statistics Poland website data for Poland and EU KLEMS data for other countries.

Poland seems to be a "non-ICT country" just as Italy is, but this could be the result of narrow definitions of ICT capital used in these countries. The narrow definition of ICT capital in Poland was adopted because it can be assessed in this way from the supply and use tables (SUT), thanks to the structure of software services (see: Kotlewski & Błażej, 2016 and 2018a) and this technique cannot be reliably used if the ICT capital is defined widely[33]. At the same time, however, the case of Finland, where we have seen the fall of Nokia impacting the entire ICT industry in this country, suggests that there is some rationale behind the figures (after the fall of Nokia Finland became relatively a non-ICT country as can be seen on the graph).

[33] As Timmer et al. (2007a, 38) summarize it: The definitions of IT and CT assets have not been completely harmonized to date. In some countries IT has been defined broadly as office and computing equipment (CPA 30), whereas others have used the more narrowly defined category CPA 3002 computers only. Similarly, CT …etc.

The EU KLEMS site does not provide more information (e.g. on constant quality indices) about the countries represented on this site that could shed more light on this issue, but the results for Finland suggest some degree of credibility of these comparisons between the countries. It is possible, therefore, that not all countries do benefit from the explosive growth of productivity of ICT technologies (as suggested, e.g. in: Jorgenson, Ho, Stiroh (2005) and Ark, O'Mahony, Timmer (2008)).

The final observation presented here is that there are some premises to think that Poland in the long run is subject to *secular stagnation* just as the remaining world is (as observed through the EU KLEMS performing countries). The long-run trend, as can be seen on all the figures presented here, is a decreasing one. But the concept of *secular stagnation* does not take into consideration that qualitative growth may be replacing quantitative growth now[34]. In an environment of slowing down demographics, accompanied by a pressure on reducing fuel and material consumption and on shrinking sizes and weights of all end-user machinery, it might be that qualitative growth has become prevalent. This, however, remains debatable[35].

4. Preliminary interpretation of results at industry level

To carry on an analysis at industry level, a wide industry approach at NACE 2 classification sections has been applied here. Section A (agriculture, forestry and fishing) has been omitted because the KLEMS methodology is considered to be controversial for this economic activity. Also, NACE 2 sections not belonging to the so-called 'market economy' according to the standard approach (which includes sections L, O, P and Q) have been omitted. Because of their little importance, section T and U have been omitted too. However, NACE sections representing commercialized activities, but under strong public control or with heavy sovereign supports have been included (sections B, D, E, H and R). Industries represented by these sections have huge investment outlays, which are directly or indirectly sovereign supported, and which do not necessarily translate into gross value added growth.

In section B (mining and quarrying) we can observe a negative gross value added growth related to the process of mining industry restructuring carried out under sovereign control, so despite some investments the residual MFP contribution value is negative. In sections D (electricity, gas, steam and air conditioning supply) and E (water supply, sewerage, waste management and remediation activities) that concern network

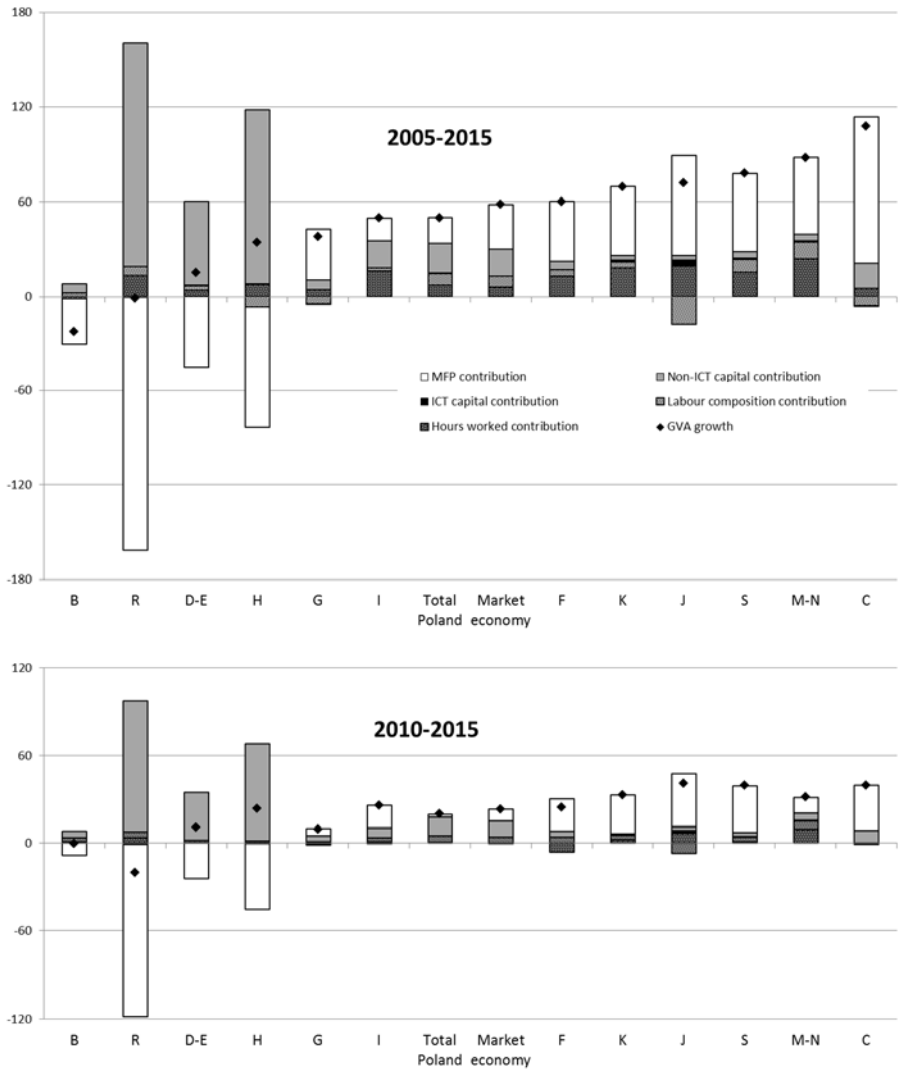
[34] A discussion about how to measure it and the possibility to use hedonic quality measures is discussed in Hulten (2009), among others.

[35] One may think that in the light of Krugman (2014, 61-68), the only way to combat secular stagnation is to appropriately reward qualitative growth, thanks to a monetary expansion that would allow prices to grow as an equivalent for improved quality of goods, but near zero policy interest rates do not give room for that. Perhaps only quantitative easing remains, which, however, remains scary for monetary economists.

services there are important upgrading outlays (probably necessary) that are not accompanied by substantial output growth, therefore here as well the contribution of residually computed MFP is negative. To some degree it is the same with section H (transportation and storage) in which also important public outlays are only partly accompanied by an increase in transport services – therefore we also observe a negative MFP contribution. Also a capital public support for section R (arts, entertainment and recreation) does not directly lead to production increase, which inevitably leads to negative MFP contribution.

Those negative MFP contributions are compensated by positive ones in the other industries included in the analysis, therefore the results for total Poland on the graph from Figure 6, are in the middle. Sections not included in the market economy contribute also negatively to the overall MFP contribution and that is why this category is situated to the right from total Poland.

One important observation in Figure 6 is that NACE 2 section C representing a group of industries related to manufacturing has the highest MFP cumulative contribution in the 2005–2016 period. Because this section is the largest in the Polish economy it weighs very importantly on the total economy MFP contribution in a situation where its level within the section is high as well. As they are very technical (as manufacturing generally is), it can be asserted that industries from this section have in general the greatest technological progress (with some possible but not numerous exceptions). It



Note: on the first graph (2005–2015) the NACE 2 classification sections are in order of growing cumulative gross value added growth from the left side to the right-hand side; on the second graph (2010–2015) this order has been maintained.

Figure 6. Decomposition of cumulative gross value added growth into factor and MFP contributions at selected NACE 2 classification sections in the light of KLEMS growth accounting (in pp).

Source: Own contribution based on Statistics Poland website data for Poland.

should translate to an important contribution of MFP into gross value added growth in this section, but the fact that this contribution dominates entirely over other

contributions is an important novelty observation. This suggests that manufacturing is being intensively upgraded (modernized) in Poland, and this is happening regardless of whether this upgrading is replicative (through imitation and acquiring of foreign technologies) or innovative. This upgrading seems to be the basic growth engine for the industries in this section, rather than new capital outlays. Section C in relative terms has the highest growth between all NACE 2 sections. This implies that a reindustrialization process is under way in Poland.

The second NACE 2 classification section that distinguishes itself by its high MFP contribution is section J (information and communication). In the period 2005-2016 the cumulative gross value added growth and the cumulative MFP contribution to that growth in that section was only a little lower than in section C, but in the second half of that period (2010–2016) as shown on the lower graph of Figure 6 it becomes the leader in both of these categories. Sections I (accommodation and food service activities) and S (other service activities) also have increased their importance thanks mainly to MFP contributions. In general we can observe that in all growth supporting activities it is MFP contribution that dominates. Therefore, MFP can be considered as the main *growth engine* in the economy and this domination remains in the second half of the analysed period (2010–2016). However, at the total economy level this important MFP contribution is being levelled by industries from the above-mentioned sections, which do not contribute importantly or positively to gross value added growth and this effect even strengthened in the 2010–2016 period.

The more general implication is that the Polish economy is developing well and intensively modernizing, particularly in industries from the well growing sections as shown in Figure 6 on the one hand. On the other hand, the share of the other industries that are not greatly contributing to gross value added growth or contributing negatively is to large (which would be a *government failure* paradigm supporters' view). However, some contenders might assert that there is no trouble at all since in general the MFP negative contributions of industries from some sections are well counterbalanced by MFP positive contributions of industries from other sections (which would be rather a *market failure* paradigm supporters' preferable interpretation).

The possible analyses at industry level are very numerous (by far trespassing the size of this article), so only a sample is provided here.

5. Developed labour factor decomposition

One of the key idea presented in this article is the possibility to further analyse the labour factor contribution by dividing it into three (and even four as shown latter on) instead of two sub-contributions. Considering KLEMS decomposition at industry

levels, this seems to open new possibilities for analysing the business cycle and the labour market itself, and can lead to promising linkages to other studies.

The standard KLEMS decomposition of the labour factor contribution (called labour services' contribution) into two sub-contributions is:

$$\overline{w}_{jt}^{L}\Delta \ln L_{jt} = \overline{w}_{jt}^{L}\Delta \ln LC_{jt} + \overline{w}_{jt}^{L}\Delta \ln H_{jt} \tag{6}$$

This formula is the same as formula (4) in section 2, but labour quality (Q) is understood here as labour composition (LC) solely, and is therefore calculated through a subtraction between hours worked H_{ljt} for labour kinds l growths aggregated with the use of the Törnqvist quantity index over industry j and hours worked H_{jt} growths simply added up for the given industry j:

$$\Delta lnLC_{jt} = \sum_{l} \bar{v}_{l,jt}\Delta lnH_{ljt} - \Delta lnH_{jt} \tag{7}$$

In the above-mentioned formulae $\overline{w}_{jt}^{L}$ stands for the average share of the labour factor remuneration (labour compensation together with the self-employed) in gross value added of industry j for two discrete periods *(t-1)* and t. $\Delta \ln L_{jt}$ stands for the relative growth of the labour factor, understood in the standard KLEMS accounting as labour services, in industry j between two discrete periods *(t-1)* and t; and $\Delta \ln H_{jt}$ for the relative growth of the number of hours worked in industry j, between these two discrete periods. $\Delta \ln LC_{jt}$ stands for the relative change in the so-called labour composition (otherwise called labour quality in standard KLEMS accounting) in industry j between two discrete periods *(t-1)* and t, understood as an effect of the change in the share structure of the labour factor by different labour kinds l. ΔlnH_{ljt} stands for the relative growth of the number of hours worked in industry j between two discrete periods *(t-1)* and t of different labour kinds l, whereas $\bar{v}_{l,jt}$ are average shares of labour kinds l in labour compensation in industry j between two discrete periods *(t-1)* and t. As seen from equations (6) and (7), only data on the hours worked and value shares are needed for the accounts (all data are according to the National Accounts).

In this way the traditionally understood (Solow, 1956 and 1957) contribution of the labour factor to economic growth as the contribution of hours worked solely is complemented in standard KLEMS accounting by the contribution of labour composition (otherwise called labour quality), which was contained before in the so-called Solow residual. Following this change, the labour inputs have been renamed as labour services' inputs in the standard KLEMS accounting. As mentioned before, in the KLEMS accounting 18 kinds of labour are defined, which arises from divisions into sexes, three age groups and three education attainment levels. Labour composition change is understood, therefore, as the effect of the change in the relative remuneration

shares of different 18 labour kinds at industry level (within industries)[36]. This effect is as conspicuous as the different labour kinds *l* are differently remunerated by hour.

This analysis of the labour factor contribution (labour services' contribution in standard KLEMS accounting) can be deepened further, however. The contribution of hours worked growth in formula (6) can be decomposed by changing this formula into:

$$\overline{w}_{jt}^{L}\Delta\, ln\, L_{jt} = \overline{w}_{jt}^{L}\Delta\, ln\, LC_{jt} + \overline{w}_{jt}^{L}\Delta\, ln\, M_{jt} + \overline{w}_{jt}^{L}\Delta lnH_{Mjt} \tag{8}$$

where:

$$\Delta lnH_{Mjt} = \Delta\, ln\, H_{jt} - \Delta lnM_{jt} \tag{9}$$

In the above-mentioned formulae ΔlnH_{Mjt} is the relative growth of hours worked *H* per employee *M* in industry *j* between two discrete periods *(t-1)* and *t*. In practice, it is calculated residually by subtracting the relative growth of the number of employees, i.e. ΔlnM_{jt}, from the relative growth of hours worked, i.e. $\Delta\, ln\, H_{jt}$ in industry *j* between two discrete periods *(t-1)* and *t*. This technique of residual calculations is the reason why the above formulae are always met in practice and is therefore a better technique for the accounts than to divide the number of hours by the number of employees and observe the changes of this ratio. More generally, the rationale behind this procedure is that the growth of hours worked at a given aggregation can be the result of two distinct processes. One is the possibility that the number of employees is increasing, the other is the possibility that the number of hours per employee is increasing, and it is assumed that these processes may not have exactly the same consequences.

A more detailed decomposition of hours worked contribution into the contributions of the number of employees and the number of hours per employee can be of some not negligible importance for using KLEMS results in analyses oriented at economic policy. In the case of a negative shock, the economy reacts usually by reducing the total number of hours worked. However, the case of reduced number of employees with stabilization (or even increase) of the number of hours per employees is well different from the case when the economic adjustment takes the shape of a decreased number of hours per employee with little employee reduction. In the first case, the social consequences are more severe, which results in reduced household consumption (a contagion alike consumer spending reduction effect) and the eventuality of high costs of bringing back the previous employment level (because of a hysteresis effect). In the second case, the social consequences are milder, which results in a lower decrease of the household consumption level and its quicker restauration

[36] The contribution of labour reallocation between industries with different levels of productivity can also be taken into account in theory (Timmer et al. 2010, 153, eq. (5.4)).

(because consumers tend to maintain their spending levels when their incomes decrease moderately).

This analysis can be a contribution to explaining the reasons of different reactions of the EU economies against the 2007–2009 crisis and of different paces of growth restauration. Some initial studies carried for some chosen EU economies according to a simplified methodology seem to demonstrate these different reactions in accordance with the rationale presented here (CSO Poland[37], 2014). This could be also interesting when regional (by provinces of the given country) growth decompositions will be performed[38]. On the right hand, in the side graph of Figure 7 we can see that in 2009 the contribution of hours per employee was negative, and thanks to that the contribution of the number of employees[39] remained positive even in the situation where the total number of hours worked decreased. This explains presumably why consumer spending did not decrease importantly in Poland in comparison with the other countries in that year. Obviously,

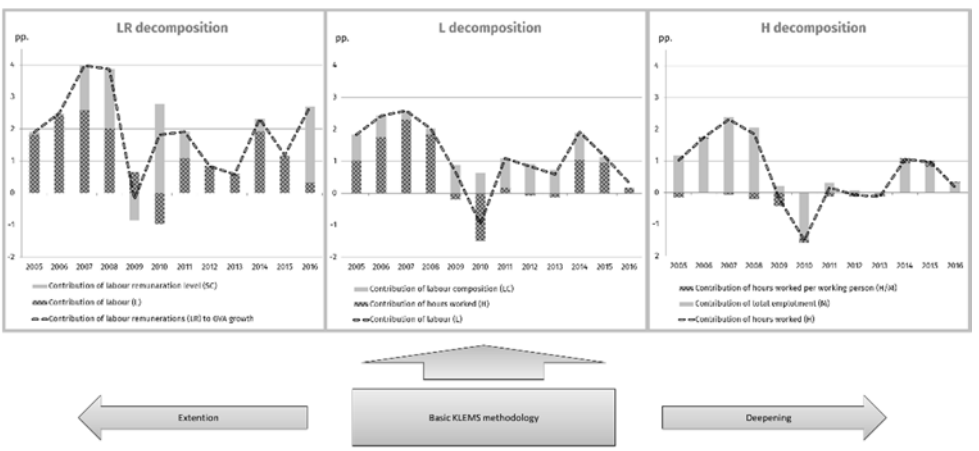


Figure 7. Developed labour factor decomposition

Source: Own contribution based on Statistics Poland web page.

although it helps to understand why Poland avoided recession in that year, it is certainly not the only reason for this wishful behaviour of the economy[40].

We have also a general wage increase phenomenon, which by a margin can be different than labour factor (*L*), understood as above, increase. In real terms (deflated)

[37] Actually Statistics Poland.

[38] As Spain did. In China it has been done also, see: Kang and Peng (2013).

[39] In the entire article employees are considered as including the self-employed.

[40] The other well identified reason is a floating currency that allowed for a betterment of the balance of payments to a large degree.

and marginally, including it should help to completely reflect labour quality growth, understood as marginal labour productivity growth. The resulting conclusion is that in an ideal situation the total labour quality effect should include the wage effect from formula (10) below. The analysis of the labour factor can be therefore extended. If the contribution of labour factor (*L*) growth, calculated as above-mentioned, is subtracted from the contribution of labour compensation (*LR*) growth then we receive the contribution of the change in the relative level of remunerations (*SC* – *soft composition*[41]) according to the following formula:

$$\overline{w}_{jt}^{L}\Delta \ln SC_{jt} = \overline{w}_{jt}^{L}\Delta \ln LR_{jt} - \overline{w}_{jt}^{L}\Delta \ln L_{jt} \tag{10}$$

and here as well (according to a technique often used in KLEMS accounting) we do not need to establish the value of *SC* directly, because the value of its contribution, i.e. the left-hand side of the equation (10), can be calculated residually from the other variables as the subtraction between their contributions (the value of labour compensation (*LR* – *labour remuneration*[42]) is available from the National Accounts). In such a case the contributions of all the above-mentioned four labour factor's sub-factors can be joined in a single formula:

$$\overline{w}_{jt}^{L}\Delta \ln LR_{jt} = \overline{w}_{jt}^{L}\Delta \ln SC_{jt} + \overline{w}_{jt}^{L}\Delta \ln LC_{jt} + \overline{w}_{jt}^{L}\Delta \ln M_{jt} + \overline{w}_{jt}^{L}\Delta \ln H_{Mjt} \tag{11}$$

In the KLEMS accounting labour composition *LC* is interpreted as the main manifestation of labour efficiency in the long run[43], which only to some degree translates into the actual remunerations' level. The remaining remunerations' level change *SC* can be attributed to the actual labour usage that mostly can be related with the business cycle but also with a reallocation effect between industries[44].

For clarity, on the Statistics Poland website the Excel tables concerning this developed decomposition of the labour factor are presented in a hierarchical way following the equation (11) divided into three equations:

$$\begin{aligned} \overline{w}_{jt}^{L}\Delta \ln LR_{jt} &= \overline{w}_{jt}^{L}\Delta \ln SC_{jt} + \overline{w}_{jt}^{L}\Delta \ln L_{jt} \\ \overline{w}_{jt}^{L}\Delta \ln L_{jt} &= \overline{w}_{jt}^{L}\Delta \ln LC_{jt} + \overline{w}_{jt}^{L}\Delta \ln H_{jt} \\ \overline{w}_{jt}^{L}\Delta \ln H_{jt} &= \overline{w}_{jt}^{L}\Delta \ln M_{jt} + \overline{w}_{jt}^{L}\Delta \ln H_{Mjt} \end{aligned} \tag{12}$$

[41] Own designation.

[42] Own designation – must be different from LC (labour composition) already used.

[43] The neoclassical premise is that labour is being remunerated according to its marginal productivity.

[44] This reallocation effect has been discussed in Stiroh (2002). Here, it is contained within SC.

They represent the three stages of the labour factor decomposition presented in Figure 7. It can be seen that a drop in remunerations' contribution in 2009 shown on the left-hand side graph probably delayed the trough in labour contribution to gross value added growth in 2010 shown on the middle graph, which was accompanied by some labour hoarding in 2009 shown on the right-hand side graph in the form of a negative hours per employee growth level.

6. Conclusions and a look into the future

To conclude, the KLEMS accounting methodology applied in Poland, although not solving all economic analysis dilemmas, is a valuable tool for economic ex post observation, which can provide non-negligible findings, also for the decision makers. It does not respond to similar Keynesian findings and controversies, but the processed final KLEMS data can possibly be used also for that purpose in some cases, although mainly applicable to similar neoclassical analyses oriented rather towards the long run economic paradigm.

The main findings are:

1) that missing data can be effectively assessed up to the level allowing to build up KLEMS type data sets of sufficient quality;
2) that the labour factor can be decomposed further, which is conducive to more informed analysis;
3) that during the 2009 crisis trough MFP contribution increased, resulting perhaps from some effective reorganisations in the economy undertaken to combat the crisis both at the aggregate level and at firms levels;
4) that a specific labour hoarding phenomenon was present in the economy during the 2009 onslaught of the crisis on the Polish economy, which prevented largely the incidence of huge demand slump;
5) that the industry level analysis exhibits an even more optimistic feature of the Polish economy than at the aggregate level – it shows that the Polish economy has strong sectoral fundamentals and that only the role of the State is an issue.

Some of these findings, i.e. four of the five above-mentioned items, have already been published in previous publications (Kotlewski &Błażej 2016, 2018a and 2018b), but here they were explained more extensively. Item 4) was only signalled in (Kotlewski & Błażej 2018b), whereas here it was well explained. Item 5) is a first time result publication.

The KLEMS accounting performed in Poland can be developed. One obvious possibility is to bring to live not only the gross value added growth decomposition but also the gross output growth decomposition, which includes intermediate consumption

contribution. The theoretical KLEMS methodology[45] includes a decomposition of intermediate consumption contribution into the contributions of its three sub-components, i.e. energy, materials and services inputs. As we know from the last release of National Accounts for Poland, the deflators for intermediate consumption by industries are now available[46], so the main obstacle standing against this procedure has been almost lifted (what remains is the issue to attribute the appropriate deflators to the three kinds of intermediate consumption, i.e. energy, materials and services). The international comparability problem, arising from differences in vertical integration of firms between the countries, will remain. But, performing gross output decomposition could be useful for intra-country analyses of energy and material footprints, and the scale of services' outsourcing also by industries. Lastly (June 2019), on Statistics Poland website gross output growth decomposition has been published, although without the subdivision into the above-mentioned three categories of intermediate inputs. This partial progress can lead, however, to some new analyses[47].

Considerably more important and complex is the idea to perform KLEMS accounting not only for the entire economy at the aggregate level and by industries, but also regionally, i.e. by individual Polish sixteen voivodships. This could allow to make analytical comparisons between voivodship regional economies (as some Polish voivodships are as large as some small European countries), also in relation to the aggregate economy as a whole, and could happen to become an important supporting tool for economic analyses oriented at regional economic policy.

REFERENCES

ARK BART VAN, (2016). Are Emerging Markets Still Emerging?, Forth World KLEMS Conference, Madrid 2016.

ARK BART VAN, O'MAHONY, M., TIMMER MARCEL P., (2008). The productivity gap between Europe and the United States, Journal of Economic Perspectives, 22(1), pp. 25–44.

BHAGWATI JAGHDISH N., (1982). Directly Unproductive Profit-seeking (DUP) Activities, Journal of Political Economy, Vol. 90, No. 5.

DIEWERT W., EDWIN, (1993). The Early History of Price Index Research, in: Diewert W.E., Nakamura A.O. (1993), Essays in Index Number Theory, Elsevier, B.V.

[45] Particularly comprehensively explained in O'Mahony and Timmer (2009).

[46] They are different as the price inflation for intermediate goods is different than for final goods.

[47] A paper about this issue is under elaboration.

GOUMA REITZE, MARCEL P. TIMMER, (2013a). EU KLEMS Growth and Productivity Accounts – 2013 update, Groningen Growth and Development Centre.

GOUMA, REITZE, MARCEL P. TIMMER, (2013b). WORLD KLEMS Growth and Productivity Accounts – 2013 update, Groningen Growth and Development Centre.

GRADZEWICZ, M., GROWIEC, J., KOLASA, M., POSTEK, Ł., STRZELECKI, P., (2014). Poland's Exceptional Growth Performance During the World Economic Crisis: New Growth Accounting Evidence. Working Paper, No. 186. NBP.

GRADZEWICZ, M., GROWIEC, J., KOLASA, M., POSTEK, Ł., STRZELECKI, P. (2018). Poland's Uninterrupted Growth Performance: new growth accounting evidence, in: Post-Communist Economies, Vol. 30, No. 2, pp. 238–272.

HAUSMAN, J., (2003). Sources of Bias and Solutions to Bias in the Consumer Price Index, Journal of Economic Perspective, Vol. 17, No. 1, pp. 23–44.

HAVLIK P., LEITNER, S., STEHRER, R. (2012). Growth Resurgence, Productivity Catching-up and Labor Demand in Central and Eastern European Countries, in: Matilde Mas, Robert Stehrer (Ed.), Industrial Production in Europe. Growth and Crisis, Edward Elgar, pp. 219–263.

HULTEN, C. R., (2009). Growth Accounting, NBER Working Paper 15341.

JORGENSON, D. W., (1963). Capital Theory and Investment Behavior, American Economic Review, 53(2), pp. 247–259.

JORGENSON, D. W., (1989). Productivity and Economic Growth, in Ernst R. Berndt and Jack E. Triplettt (eds.), Fifty Years of Economic Measurement, University of Chicago Press.

JORGENSON, D. W., (2012). The World KLEMS Initiative, International Productivity Monitor.

JORGENSON, D. W., ZWI GRILICHES, (1967). The explanation of Productivity Change, Review of Economic Studies, 34, pp. 249–83.

JORGENSON, D. W., ed., (2009). The Economics of Productivity, Edward Elgar Publishing.

JORGENSON, D. W., FRANK M. GOLLOP, BARBARA M. FRAUMENI, (1987). Productivity and U.S. Economic Growth, Cambridge, MA: Harvard University Press.

JORGENSON, D. W., MUN S. HO, KEVIN J. STIROH, (2005). Information Technology and the American Growth Resurgence, Cambridge, MA: The MIT Press.

JORGENSON, D. W., MUN S. HO, JON D. SAMUELS, (2014). Long-term Estimates of U.S. Productivity and Growth, Third World KLEMS Conference paper.

KANG LILI, FEI PENG, (October 2013). Growth Accounting in China 1978–2009, MPRA Paper No. 50827.

KOTLEWSKI, D., C. (2017). Problem cen w regionalnym rachunku produktywności, Wiadomości Statystyczne 12/2017.

KOTLEWSKI, D., BŁAŻEJ, M., (2016). Metodologia rachunku produktywności KLEMS i jego implementacja w warunkach polskich, Wiadomości Statystyczne 9/2016.

KOTLEWSKI, D., BŁAŻEJ, M., (2018a). Implementation of KLEMS economic productivity accounts in Poland, Lodz University Publication, Folia Oeconomica 2/2018.

KOTLEWSKI, D., BŁAŻEJ, M., (2018b). KLEMS Productivity Accounts – Poland 2005–2016, Statistics Poland.

KLEMS Economic Productivity Accounts, https://stat.gov.pl/en/experimental-statistics/klems-economic-productivity-accounts/

KOTLEWSKI, D., BŁAŻEJ, M., (updated regularly every year or every two years), KLEMS Economic Productivity Accounts, Statistics Poland, https://stat.gov.pl/en/experimental-statistics/klems-economic-productivity-accounts/methodology-of-decomposition-in-klems-productivity-accounts-for-the-polish-economy,2,1.html

KRUGMAN, P., (2014). Four observations on secular stagnation in Coen Teulings and Richard Baldwin, eds., Secular Stagnation: Facts, Causes, and Cures, CEPR Press, pp. 61–68.

LEWANDOWSKI, M., BANAŚ, M., KOTLEWSKI, D., KULCZYCKA, J., DONIEC, D., WITKOWSKI, G., and others, (2015). Metoda dekompozycji Produktu Krajowego Brutto (PKB) oraz Wartości Dodanej Brutto (WDB) w zastosowaniu do analizy struktury różnić regionalnych, Statistics Poland, http://stat.gov.pl/statystyka-regionalna/statystyka-dla-polityki-spojnosci/realizacja-prac-metodologicznych-analiz-ekspertyz-oraz-prac-badawczych-na-potrzeby-polityki-spojnosci/dezagregacja-wskaznikow-z-obszaru-rachunkow-narodowych-i-regionalnych .

MILANA, C., (2009). Solving the Index-Number Problem in a Historical perspective, EU KLEMS Working Paper 43.

O'MAHONY, MARY and MARCEL, P., TIMMER, (2009) Output, Input and Productivity Measures at the Industry Level: The EU KLEMS Database, The Economic Journal, 119 (June), F374–F403.

OECD, (2001) Measuring Productivity, OECD Manual.

OECD, (2009) Measuring Capital, OECD Manual.

OECD, (2013) OECD Compendium of Productivity Indicators 2013, OECD Publishing.

SCHREYER, P., (2004). Chain Index Number Formulae in the National Accounts, 8th OECD – NBS Workshop on National Accounts.

SOLOW R., .M. (1956). A Contribution to the Theory of Economic Growth, Quarterly Journal of Economics, Vol. 70, No. 1, pp. 65–70.

SOLOW R., M. (1957). Technical Change and the Aggregate Production Function, Review of Economics and Statistics, Vol. 39, No. 3, pp. 312–320.

STATISTICS POLAND (2014). Macroeconomic Situation in Poland in 2013 in the Context of the World Economic Processes, a yearly report.

STIROH, K., J., (2002). Information Technology and the U.S. Productivity Revival: What Do the Industry data Say? American Economic Review, 92(5), pp. 1559–76.

TIMMER M. P., MARY O'MAHONY, BART VAN ARK, (2011). Productivity and Economic Growth in Europe: A Comparative Industry Perspective, International Productivity Monitor.

TIMMER MARCEL P., INKLAAR, R., O'MAHONY, M., ARK BART VAN, (2010). Economic Growth in Europe, Cambridge University Press.

TIMMER, MARCEL P., ABDUL AZEEZ ERUMBAN, BART LOS, ROBERT STEHRER, GAAITZEN J. DE VRIES, (2013). Slicing Up Global Value Chains, Journal of Economic Perspectives, Vol. 28, No. 2, pp. 99–118.

TIMMER, MARCEL P., TON VAN MOERGASTEL, EDWIN STUIVENWOLD, GERARD YPMA, (Groningen Growth and Development Centre) and Mary O'Mahony, and Mari Kangasniemi (National Instutute of Economic and Social Research) (March 2007a) EU KLEMS Growth and Productivity Accounts – Metodology, EU KLEMS Consortium.

TIMMER, MARCEL P., TON VAN MOERGASTEL, EDWIN STUIVENWOLD, GERARD YPMA, (Groningen Growth and Development Centre) and MARY O'MAHONY, and MARI KANGASNIEMI (National Instutute of Economic and Social Research), (2007b). EU KLEMS Growth and Productivity Accounts – Sources by country, EU KLEMS Consortium.

WÖLFL, ANITA and DANA HAJKOVA, (2007) Measuring multifactor productivity growth, STI Working Paper 2007/5, OECD.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 123–136, DOI 10.21307/stattrans-2020-007
Submitted – 10.01.2019; accepted for publishing – 13.12.2019

Change-point detection in CO_2 emission-energy consumption nexus using a recursive Bayesian estimation approach

Olushina Olawale Awe[1], **Abosede Adedayo Adepoju**[2]

ABSTRACT

This article focuses on the synthesis of conditional dependence structure of recursive Bayesian estimation of dynamic state space models with time-varying parameters using a newly modified recursive Bayesian algorithm. The results of empirical applications to climate data from Nigeria reveals that the relationship between energy consumption and carbon dioxide emission in Nigeria reached the lowest peak in the late 1980s and the highest peak in early 2000. For South Africa, the slope trajectory of the model descended to the lowest in the mid-1990s and attained the highest peak in early 2000. These change-points can be attributed to the economic growth, regime changes, anthropogenic activities, vehicular emissions, population growth and industrial revolution in these countries. These results have implications on climate change prediction and global warming in both countries, and also shows that recursive Bayesian dynamic model with time-varying parameters is suitable for statistical inference in climate change and policy analysis.

Key words: dynamic model, Bayesian inference, CO_2, climate change, energy.

1. Introduction

A major reason for the burgeoning popularity of the recursive Bayesian estimation approach where new estimates are required each time a new measurement arrives in empirical science is the increasing prominence of numerical simulations by computational algorithms, which relies heavily on the Markov chain Monte Carlo methods (Ng and Young, 1990). Significant breakthroughs in the application of recursive Bayesian models with time-varying parameters in econometrics has been recorded by the works of authors like Pollock (2003), Chow et al. (2011), Del Negro and Otrok (2008) and Young (2011).

[1] Department of Mathematical Sciences, Anchor University, Lagos, Nigeria. E-mail: oawe@aul.edu.ng. ORCID: https://orcid.org/0000-0002-0442-4519.
[2] Department of Statistics, University of Ibadan, Ibadan, Nigeria. E-mail: pojuday@yahoo.com. ORCID: https://orcid.org/0000-0003-2368-4313.

Recursive Bayesian algorithms are mainly based on statistical dependence of random variables. Two random variables, say x and y, are statistically independent if and only if their joint distribution is equal to the product of their marginal distribution.

$$p(x, y) = p(x)p(y) \tag{1}$$

intuitively,

$$p(x, y) = p(x)p(y|x) = p(y)p(x|y) \tag{2}$$

represent the fact that the conditional distribution of one random variable, given the other, is not a function of what it is being conditioned on. So,

$$p(x|y) = p(x) \tag{3}$$

and

$$p(y|x) = p(y) \tag{4}$$

This definition can easily be extended to more than two random variables. The joint distribution of a collection of independent random variables is the product of their marginal distributions (Petris et al., 2009; Hillebrand and Koopman, 2016). Basically, all conditional distributions are independent of the random variables they are conditioned on. In statistical dependence, knowledge of x tells us something about y. Using Bayes' rule, we can easily show that the reverse is also true, i.e. knowledge of y also tells us something about x. Another important definition to consider is that of conditional independence. Two random variables x and y are conditionally independent given another random variable z if and only if:

$$p(x, y|z) = p(x|z)p(y|z) \tag{5}$$

or equivalently,

$$p(x|y, z) = p(x|z) \tag{6}$$

$$p(y|x, z) = p(y|z) \tag{7}$$

In Bayesian analysis, the simplest dependence structure is conditional independence. It can be assumed, in many applications, that the random observations $y_1, \ldots, y_n$ are conditionally independent and identically distributed given the parameter θ. Mathematically,

$$f(y_1, \ldots, y_n|\theta) = \prod_{t=1}^{n} (y_t|\theta) \tag{8}$$

In particular,

$$p(y_1, y_2) = \int p(y_1, y_2|\theta)p(\theta)d\theta$$
$$= \int p(y_1|\theta)p(y_2|\theta)p(\theta)d\theta$$

Suppose that the observations $y_1, \ldots, y_n$ provide us information about the unknown parameter θ, and through θ we can also obtain information about the next observation $y_{n+1,}$ then y_{n+1}depends on the past observations $y_1, \ldots, y_n$ in a probabilistic sense (Petris et al., 2009). Furthermore, the predictive densities for the case above can be computed as

$$(y_{n+1}|y_1, \ldots, y_n) = \int f(y_{n+1}, \theta|y_1, \ldots, y_n)d(\theta) \tag{9}$$

$$= \int f(y_{n+1}|\theta)\pi(\theta|y_1, \ldots, y_n)d(\theta) \tag{10}$$

where $\pi(\theta|y_1, \ldots, y_n)$ is the posterior density of θ, conditional on the data $(y_1, \ldots, y_n)$. The posterior density can be computed by Bayes' formula as noted earlier.

$$\pi(\theta|y_1, \ldots, y_n) = \frac{f(y_1, \ldots, y_n|\theta)\pi(\theta)}{p(y_1, \ldots, y_n)} \; \alpha \; \prod_{t=1}^{n} (y_t|\theta)\, \pi(\theta) \tag{11}$$

where the marginal density $p(y_1, \ldots, y_n)$ is playing the role of a normalizing constant and hence does not depend on θ. More so, with the assumption of conditional independence, the posterior distribution can be computed recursively. This implies that we do not need all the previous data to be kept in storage and processed every time a new measurement is taken.

At times (n – 1), the information available about the parameter θ can be described by the conditional density

$$\pi(\theta|y_1, \dots, y_{n-1}) \, \alpha \, \prod_{t=1}^{n-1} (y_t|\theta) \, \pi(\theta) \tag{12}$$

The density (12) above can then play the role of prior at time n. Once the new observation y_n becomes available, we would just compute the likelihood, which is given as

$$f(y_n|\theta, y_1, \dots, y_{n-1}) = f(y_n|\theta) \tag{13}$$

by the assumption of conditional independence and then update the prior $\pi(\theta|y_1, \dots, y_{n-1})$ by Bayes' rule to obtain

$$\pi(\theta|y_1, \dots, y_{n-1}, y_n) \, \alpha \, \pi(\theta|y_1, \dots, y_{n-1}) f(y_n|\theta) \, \alpha \, \prod_{t=1}^{n-1} f(y_t|\theta) \, \pi(\theta) f(y_n|\theta) \tag{14}$$

which is equivalent to equation (12). This recursive structure of the posterior distribution is crucial in the study of dynamic linear models with time-varying parameters, which is used to analyze climate variables in this present work. The treatment of recursive regression has a Bayesian flavour and relies on the calculus of conditional expectations, whose essentials have been provided in equation (1) – (7). Essentially, this article presents how the nexus of energy consumption and carbon dioxide emission can be studied and predicted using the recursive Bayesian estimation approach.

Essentially, the main objective of this study is to contribute to global warming and climate change research by investigating the change points in CO_2 emission-energy consumption nexus over time in two major African countries using a new modified recursive Bayesian estimation approach of (Awe and Adepoju, 2018), which has been found to be computationally less intensive. To achieve this objective, we apply a model where the observational variance in the Bayesian dynamic linear model of West and Harrison (1997) is constant and the evolution variance, which is time-varying, is represented as a fraction of the filtering variance. This present work is a continuation and application of our previous work.

2. State space representation of the Bayesian Dynamic Linear Model (DLM)

The Bayesian state space model has been defined by Petris et al. (2009), Awe et al. (2015) as one which consists of an R^p-valued time series $\beta_t: t = 1,2,\dots,T$ and an R^k-valued time series $y_t: t = 1,2,\dots,T$, which satisfies the following assumptions:

- β_t is a Markov chain
- Conditional on β_t, the $y_t's$ are independent and depend on β_t only.

State space models (dynamic linear models, in particular) are useful for modelling time-varying scenarios (Doh and Connolly, 2013), whose applications exist heavily in environmental science, economics and engineering. An archetypical dynamic linear (state space) model takes the following general form:

$$y_t = F_t'\beta_t + v_t \qquad v_t \sim N(0, V_t) \tag{15}$$

$$\beta_t = G_t\beta_{t-1} + w_t \qquad w_t \sim N_P(0, W_t) \tag{16}$$

$$\beta_0 \sim N_P(m_0, C_0) \tag{17}$$

where y_t is a vector of observed time series of dimension $m \times 1$.

Equation (15) is known as the observation equation, while equation (16) is a first order Markov process called the evolution equation. G_t and F_t are known matrices of order $p \times p$ and $m \times p$ respectively, which determine how the observation and state equation evolve in time (Lee, 2012). B_t are Markov-modulated time-varying parameters, which are to be estimated and whose structural behaviour we want to study. It also contains time-varying intercepts $\propto_t$, which were estimated. The matrices $F_t, V_t, G_t \; and \; W_t$ are known as the system matrices and contain non-random elements. If they do not depend deterministically on t, the state space system is time invariant, otherwise they are time varying. The initial state distribution is assumed to be normally distributed with parameters $m_0 \; and C_0$ as shown in (17), where $E(v_t\beta_t') = 0$, $E(w_t\beta_t') = 0$ for $t = 1,2,\dots,T$. It is often convenient to study the properties of a process when the model is in the state space form because of the Markovian property of the model which assumes that the knowledge of the present state is relevant to the predictions about the future of the system, although additional information about the past state is irrelevant.

3. Recursive estimation of model parameters

Assuming normality, we estimated βt and f_t by using the Kalman filter algorithm (Kalman, 1960). V_t is assumed to be fixed and distributed inverse-gamma a priori and is estimated using a Gibbs sampler. We propose the use of discount factors to estimate w_t in the spirit of Awe et al. (2015).

Basically, the estimation of the Bayesian state space model involves three important stages: prediction, filtering and smoothing. Prediction has to do with forecasting future values of the time-varying state parameters. Filtering makes the best estimate of the current values of the state from the record of observations including the current observation. Smoothing involves making the best estimate of past values of the states given the record of observations. Suppose data $d_t$obtained till time t is represented as $D_t = (D_{t-1}, y_t)$ meaning combination of data until time t – 1 and observations at time t. Using Bayes' formula, and supposing the parameter of interest is β_t, we have

$$\pi(\beta_t)|D_{t-1}, y_t) = \frac{\pi(y_t|\beta_t, D_{t-1})\pi(\beta_t|D_{t-1})}{\pi(y_t|D_{t-1})}$$

$$\pi(y_t|\beta_t|D_{t-1}) = \pi(y_t|\beta_t) \tag{18}$$

and

$$\pi(\beta_t|D_{t-1}, y_t)\ \alpha\ \pi(y_t|\beta_t)\pi(\beta_t|D_{t-1})$$

where $$\beta_t|D_t \sim N(m_t, C_t).$$

The prior distribution for β_t is

$$\beta_t|D_{t-1} \sim N(a_t, R_t),$$

The likelihood is

$$y_t|\beta_t \sim N(F_t\beta_t, V_t).$$

The posterior for β_t is

$$\beta_t|D_t \sim N(m_t, C_t).$$

m_t and C_t are iteratively computed for $t = 1, \ldots, n$ from the following Kalman filtering equations:

$$a_t = G_t m_{t-1}$$

$$f_t = F_t a_t \tag{19}$$

$$R_t = G_t C_{t-1} G_t' + w_t \tag{20}$$

$$Q_t = F_t' R_t F_t + V_t \tag{21}$$

$$e_t = y_t - f_t \tag{22}$$

$$A_t = R_t F_t Q_t^{-1} \tag{23}$$

$$m_t = a_t + A_t e_t \tag{24}$$

$$C_t = R_t - A_t Q_t A_t' \tag{25}$$

$$\tag{26}$$

From the algorithm, f_t is the predicted value of observation at time t+1. e_t represents the prediction error. In order to predict from the posterior estimate of the state parameter β_t, a one-step-ahead prediction is made as follows:

$$E(y_{t+1}|D_t) = E(F_{t+1}\beta_{t+1} + v_{t+1}|D_t) \tag{27}$$

$$= F_{t+1} E(\beta_{t+1}|D_t)$$

hence,

$$F_{t+1}' a_{t+1} = f_{t+1} \tag{28}$$

also,

$$V[y_{t+1}|D_t] = V[F_{t+1}\beta_{t+1} + V_{t+1}|D_t] \tag{29}$$

$$= F_{t+1} V[\beta_{t+1}|D_t] F_{t+1} + V_{t+1}) \tag{30}$$

$$= F_{t+1} R_{t+1} F_{t+1} + V_{t+1} \tag{31}$$

$$= Q_{t+1}$$

For the smoothing aspect, we used the Recursive Forward Filtering Backward Smoothing algorithm of Carter and Kohn (1994) and Awe et al. (2015), where the best estimates of past values of the states given the record of observations are obtained using the rule of total probability:

$$\pi(\beta_{t-1}|D_t) = \int \pi(\beta_{t-1}|\beta_t, D_t)\, \pi(\beta_t|D_t) d\beta_t \tag{32}$$

and

$$\pi(\beta_{t-1}|\beta_t, D_t) = \pi(\beta_{t-1}|\beta_t, D_{t-1}) \tag{33}$$

and

$$\pi(\beta_{t-1}|\beta_t, D_{t-1}) = \frac{\pi(\beta_t|\beta_{t-1}, D_{t-1})\pi(\beta_{t-1}|D_{t-1})}{\pi(\beta_t|D_{t-1})} \tag{34}$$

given that,

$$\beta_{t-1}|D_t \sim N(a_{t-1}, R_{t-1})$$

where

$$a_{t-1} = m_{t-1} + \beta_{t-1}(m_t - a_t) \tag{35}$$

$$R_{t-1} = C_{t-1} - \beta_{t-1}(R_t - C_t)\, B'_{t-1} \tag{36}$$

$$B_{t-1} = C_{t-1} G'_t R_t^{-1} \tag{37}$$

We denote $p(\theta_0, \dots, \theta_T|D_T = \prod_{t=0}^{T} p(\theta_t|\theta_{t+1}, \dots, \theta_T, D_T)$, then sample form $p(\theta_T|D_T)$ using the filtering density above.

By the Markov property,

$$p(\theta_t|\theta_{t+1}, \dots, \theta_T, D_T) = p(\theta_t|\theta_{t+1}, D_T)$$

And we proceed recursively until we have a complete sample from $p(\theta_0, \dots, \theta_T|D_T)$

A sample from the posterior state parameter was generated using the algorithm documented in Awe and Adepoju (2018).

4. Change point detection in carbon dioxide (CO_2) emission-energy consumption nexus

Our empirical application involves the estimation of time-varying parameters of the dynamic state space model with Markov-modulated structure presented in equations (15-17) above for analyzing the nexus of annual carbon dioxide emission and energy consumption in Nigeria and South Africa over time. The data were obtained from the World Development Index (WDI) database. Carbon dioxide (CO_2) series is the endogenous variable (y_t) while energy consumption is the exogenous variable. For the estimated Bayesian dynamic model, the techniques in the previous section were used and time-varying slope parameters (β_t) were estimated for both Nigeria and South Africa (two of the richest economies in Africa in the period before the fall of the global oil price (1970-2010). It is necessary to study past trends to aid future projections.

This study is important because energy consumption has been known to be an increasing function of CO_2 emission. In fact, CO_2 emission significantly depends on energy consumption and economic growth (Sulaiman and Abdul-Rahim, 2017). This implies that CO_2 emission increases with an increase in both energy consumption and economic growth. Carbon dioxide is also known as a greenhouse gas (GHG), a gas that absorbs and emits thermal radiation, thereby creating the 'greenhouse effect'. A limited number of countries are largely responsible for the African CO_2 emissions from fossil fuels. South Africa accounts for 38% of the continental total carbon dioxide emission, while 46% comes from Nigeria, Algeria, Egypt, Morocco and Libya combined. For every tonne of coal burned (energy consumption), approximately 2.5 tonnes of CO_2 are produced into the atmosphere (Aye and Edoja, 2017). The contribution of each of these sources has changed significantly through time, and still shows large differences by regions and countries. Rapid industrial development and growth of cities in major African countries have raised the quest for increasing understanding of the correlated relationship between pollution in the form of carbon dioxide emission, and energy consumption (burning of fossil fuels).

The time plot of the variables for the two countries shown in Figures 1 and 2 reveals a random walk nature during the period under study. The African continent is likely to be severely affected by climate change and global warming effects. Major catastrophes from climate change would affect the natural resources, energy consumption and economies of African nations (Aye and Edoja, 2017). Hence, there is a need to know the change points and dates in which the effect of energy consumption affects climate change in order to make adequate projections for proper future planning and policy recommendations based on past trends in the respective

countries. There are only few studies in the literature that monitor the year-to-year effect of energy consumption on CO_2 emission in Africa. Subjective Bayesian methods were proposed for use in climate modelling as early as 1997 (West and Harrison, 1997).

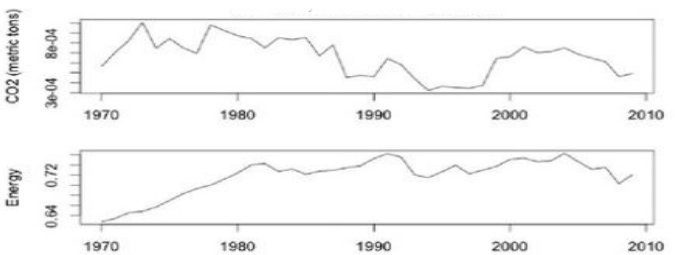


Figure 1. Time Plot of CO_2 Emission and Energy Consumption (Nigeria)

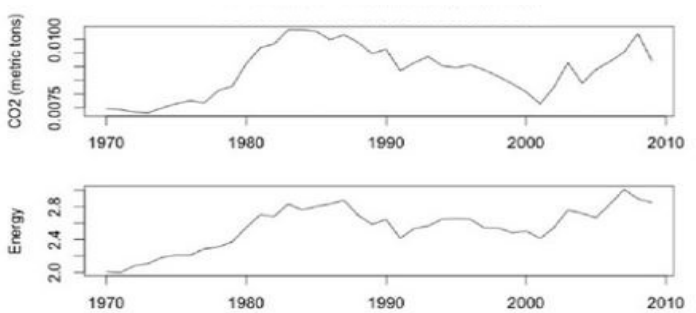


Figure 2. Time Plots of CO_2 Emission and Energy Consumption (South Africa)

The Bayesian model used in this study is therefore a suitable model which can help influence policy because many of the research statements made on climate change over the past few years took on an increasingly Bayesian flavour (Fienberg, 2011, IPCC, 2014). More so, the tradition of elicitation of expert judgments, which is what Bayesian model portends, is perfectly in line with climate policy making as suggested in the works of Morgan and Keith (1995), and Zickfeld et al. (2007). Knowing about the past trend of climate change with respect to change points and

periods of intervention can aid future policies. It has been well argued that a good predictive climate model is one that is able to adequately capture the past (Parker, 2011). Therefore, this study provides a historical time-varying perspective of how the relationship between CO_2 emissions and energy consumption have evolved over time. The probable causes of seasonal variation in concentration of atmospheric carbon dioxide due to energy consumption (burning of fossil fuels) are also discussed in the present study. A key advantage of the Bayesian method used in this work is that it allows us to combine estimates from study data with relevant past information (a prior probability distribution), to derive a posterior distribution. The prior probability distribution is capable of reflecting all information available to date on the model parameters. This is only possible and realistic via the Bayesian method adopted in this study.

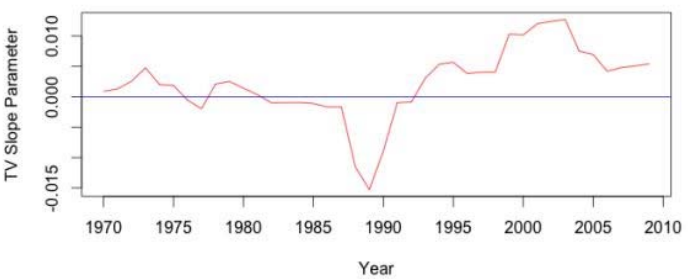


Figure 3. Time-Varying Slope (CO_2 Emission vs. Energy Consumption) for Nigeria

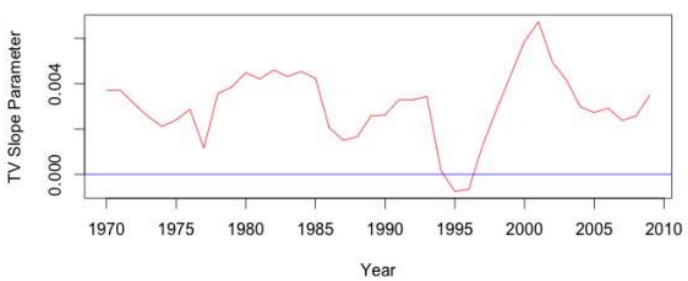


Figure 4. Time-Varying Slopes (CO_2 Emission vs. Energy Consumption) for South Africa

4.1. Result and discussion

The trend observed in Figures 3 and 4 is approximately a true reflection of the cross-sectional relationship of the observed climate variables over time across the two major countries considered. However, the trends differ for individual countries as

time evolved. If these trends are viewed over the timeline from 1990 onwards, we can see that there are large variations in the evolution of carbon intensities in both countries. Figure 3 shows that the relationship between energy consumption and CO_2 emission in Nigeria experienced the lowest peak in the late 1980s while that of South Africa (Figure 4) experienced the lowest peak in the mid-1990s. It, however, experienced the lowest peak in the mid-1990s around the period of independence in South Africa. Also, from early 2000, it can be shown that South Africa and Nigeria had a positive and significant impact on environmental degradation through the continuous release of carbon dioxide (CO_2) emission as revealed by the positive slopes. This increase (peak) within the studied period can be attributed to economic growth, anthropogenic activities, vehicular emission as a result of rural-urban migration, population growth, proliferation of industries and provision of social amenities with respect to new carbon emitting technologies. It also shows that the slopes exhibit a seasonal and random walk pattern from 1970 in both countries. These results agree with the use of CO_2 emissions as a pollution indicator, GDP and the production of nuclear electricity as economic indicators as reported by Baek and Pride (2014), who used the vector autoregressive co-integrated model and Johnsen cointegration during a study of countries in the major nuclear production within the period 1990–2011.

5. Concluding remarks

In this study, we have proposed a Bayesian time-varying parameter dynamic state space regression model for change-point detection in dynamic environmental and climatic processes. The model was estimated via a recursive Bayesian estimation approach for obtaining time-varying parameter shifts and change-point detection. It was observed that the contribution of energy consumption to CO_2 emission in two of the richest economies in Africa have changed significantly and erratically through time.

The relationship between energy consumption and CO_2 emission in Nigeria (Africa's largest economy) experienced the lowest peak in the late 1980s (after the Structural Adjustment Programme (SAP)), and the highest peak in mid-2000s (during a major regime change and industrial revolution) while the slope in South Africa (14^{th} highest emitter of CO_2 in the world and second largest economy in Africa (nominal GDP)) experienced the lowest peak in the mid-1990s (around the time of their independence) and the highest peak in early 2000 (during the time of their economic boom). Recursive Bayesian estimation of Bayesian state space models with time-varying parameters is useful for statistical inference on volatility of parameter estimates and change-point detection in climate change. The empirical applications

and findings in this study reveal that the increase and decrease in the relationship between carbon dioxide and energy consumption within this studied period can be attributed to economic growth, regime change, anthropogenic activities, vehicular emission as a result of rural-urban migration, population growth, proliferation of industries and provision of social amenities with respect to new carbon emitting technologies.

The need for low carbon technologies which are capable of plummeting carbon emissions and enhancing sustainable economic growth in South Africa and Nigeria is hereby recommended and emphasized. This may include policies to enhance efficiency of energy through modification from non-renewable energy to renewable energy, thereby reducing the impending greenhouse effect.

Acknowledgement

The authors acknowledge the helpful comments and efforts of two anonymous reviewers and Daniel Chinemeze Okolie for helping to typeset this article.

REFERENCES

AWE, O. O., ADEPOJU, A. A., (2018). Modified Recursive Bayesian Algorithm for Estimating Time-Varying Parameters in Dynamic Linear Models. Statistics in Transition, 19(2), pp. 239–258.

AWE, O. O., CRANDELL, I., ADEPOJU, A. A., (2015). A Time-Varying Parameter State-Space Model for Analyzing Money Supply-Economic Growth Nexus, Journal of Statistical and Econometric Methods, 4(1), pp. 73–95.

AYE, G. C., EDOJA, P. E., (2017). Effect of Economic Growth on CO_2 Emission in Developing Countries: Evidence from a Dynamic Panel Threshold Model, Cogent Economics & Finance, 5(1), 1379239.

BEAK, J., PRIDE, D., (2014). On the Income–Nuclear Energy–CO_2 Emissions Nexus Revisited, Energy Economics, 43, pp. 6–10. http://dx.doi.org/10.1016/j.eneco.2014.01.015.

CARTER, C. K., KOHN, R., (1994). On Gibbs Sampling for State Space Models, Biometrika, 81(3), pp. 541–553.

CHOW, S-M, ZU, J, SHIFREN, K., ZHANG, G., (2011). Dynamic Factor Analysis Models with Time-Varying Parameters, Multivariate Behavioral Research 46(2), 303-339. DOI: 10.1080/00273171.2011.563697.

DEL NEGRO, M., OTROK, C., (2008). Dynamic Factor Analysis Models with Time-Varying Parameters, FRB of New York Staff Report 326. DOI: 10.2139/ssrn.1136163.

DOH, T., CONNOLLY, M., (2013). The State Space Representation and Estimation of Time-Varying Parameter VAR with Stochastic Volatility, Springer, 2013.

FIENBERG, S. E., (2011). Bayesian Models and Methods in Public Policy and Government Settings, Statistical Science, 26(2), pp. 212–226.

HILLEBRAND, E, KOOPMAN, S. J., (2016). Dynamic Factor Models, ISBN: 978–1785603532.

IPCC, (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)], IPCC, Geneva, Switzerland, 151 pp. Available online.

KALMAN, R. E., (1960). A New Approach to Linear Filtering and Prediction Problems, Journal of Fluids Engineering, 82(1), pp. 35–45.

LEE, J., (2012). Measuring Business Cycle Co-Movement in Europe: Evidence from a Dynamic Factor Model with Time-Varying Parameters. Economic Letters, 115(3), pp. 438–440. DOI: 10.1016/j.econlet.2011.12.125.

MORGAN, M. G., KEITH, D. W., (1995). Subjective Judgments by Climate Experts, Environmental Science & Technology, 29(10), pp. 468A–476A.

NG, C. N., YOUNG, P. C., (1990). Recursive Estimation and Forecasting of Non-Stationary Time Series. Journal of Forecasting, 9(2), pp.173–204.

PARKER, W. S., (2011). When Climate Models Agree: The Significance of Robust Model Predictions, Philosophy of Science, 78(4), pp. 579–600.

PETRIS, G., PETRONE, S., CAMPNAGOLI, P., (2009). Dynamic Linear Models with R. Springer, 2009.

POLLOCK, D. S. G., (2003). Recursive Estimation in Econometrics, Computational Statistics & Data Analysis, 44(1), pp. 37–75.

SULAIMAN, C., ABDUL-RAHIM, A. S. (2017). The Relationship between CO_2 Emission, Energy Consumption and Economic Growth in Malaysia: a Three-Way Linkage Approach, Environmental Science and Pollution Research, 24(32), pp. 25204–25220.

WEST, M., HARRISON, P. J., (1997). Bayesian Forecasting and Dynamic Models. Springer-Verlag, New York, 2nd Edition.

YOUNG, P. C. (2011). Recursive Estimation and Time Series Analysis: An Introduction for the Student and Practitioner, Springer.

ZICKFIELD, K., LEVERMANN, A., MORGAN, M. G., KUHLBRODT, T., RAHMSTORF, S., & KEITH, D. W., (2007). Expert Judgements on the Response of the Atlantic Meridional overturning Circulation to Climate Change, Climatic Change, 82(3-4), pp. 235–265.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 137–157, DOI 10.21307/stattrans-2020-008
Submitted – 10.01.2019; accepted for publishing –13.12.2019

Robust estimation of wages in small enterprises: the application to Poland's districts

Grażyna Dehnel[1], **Łukasz Wawrowski**[2]

ABSTRACT

The paper presents an empirical study designed to test a small area estimation method. The aim of the study is to apply a robust version of the Fay-Herriot model to the estimation of average wages in the small business sector. Unlike the classical Fay-Herriot model, its robust version makes it possible to meet the assumption of normality of random effects under the presence of outliers. Moreover, the use of this version of the Fay-Herriot model helps to improve the precision of estimates, especially in domains where samples are of small sizes. These alternative models are supplied with auxiliary variables. The study seeks to present the characteristics of and differences among small business units cross-classified by selected NACE sections and district units of the provinces of Mazowieckie and Wielkopolskie. It was carried out on the basis of data from a survey conducted by the Statistical Office in Poznan and from administrative registers. It is the first study which attempts to produce estimates of average wages for this sector of the national economy.

Key words: small area estimation, indirect estimation, robust Fay-Herriot model, administrative registers, enterprise statistics

JEL Classification: C13, C51, M20

1. Introduction and motivation

Nowadays, it is widely known that small and medium-sized enterprises (SME) are a strong pillar of the economy. They play a crucial role in the economic and social sphere, not only for the country as a whole, but, even more importantly, at the regional level. This is because there is a strong correlation between the development of the SME sector and the regional development. The growth of the SME sector helps to eliminate regional differences, contributes to the improvement of living conditions of local communities and fosters creation of new jobs; in other words, it has a positive impact

[1] Poznan University of Economics and Business, Department of Statistics, Poznań.
E-mail: grazyna.dehnel@ue.poznan.pl. ORCID: https://orcid.org/0000-0002-0072-9681.
[2] Poznan University of Economics and Business, Department of Statistics, Poznań.
E-mail: lukasz.wawrowski@ue.poznan.pl. ORCID: https://orcid.org/0000-0002-1201-5344.

on the region's economic growth. In addition, entrepreneurs tend to locate their capital close to their places of residence, which enables them to rely on local resources and the local market (Strużycki 2004). The scope and intensity of investment depend on entrepreneurs' assessment of the degree of regional development, which in turn creates demand for regular publishing of economic information, such as the level of entrepreneurship, the labour market situation and investment activity.

In the context of regional development, it is small companies which play a most important role in this process. Despite that, they are usually treated as a mere component of the SME sector, and thus are rarely the subject of separate studies or analyses. The scope of available data on small businesses is limited, especially at low levels of aggregation. Such data are most often collected by sample surveys conducted by Statistics Poland. The study presented in this article is the attempt to partly fill this gap. The aim of the study is to estimate one indicator of entrepreneurship, namely average monthly wages in small enterprises at the level of districts. The analysis was limited to four NACE sections (manufacturing, construction, trade, transportation) and two provinces representing the highest level of entrepreneurial activity in Poland – Wielkopolskie and Mazowieckie. The province of Mazowieckie comprises 42 districts (including 5 cities) and the province of Wielkopolskie 35 districts (including 4 cities).

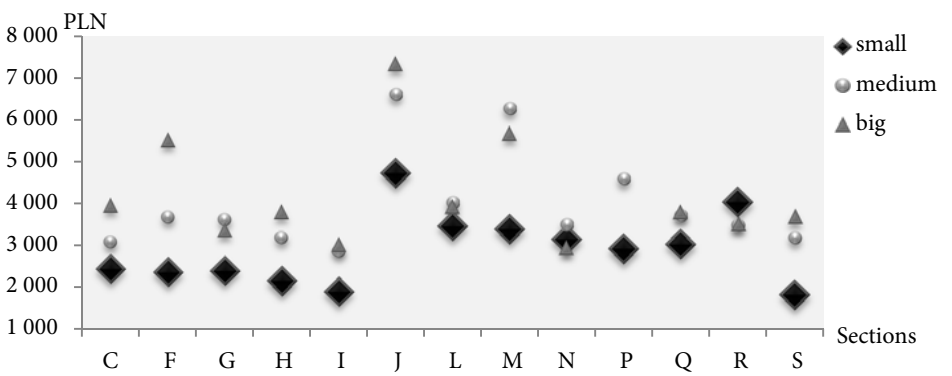
In the case of small companies, information on monthly financial results by NACE section is available only at the country and province levels. However, using the methods of indirect estimation offered by small area estimation (which are resistant to outliers), the authors managed to obtain estimates for a more detailed domain of interest, by cross-classifying NACE section with the territorial division into districts.

The article consists of five parts. The first part is devoted to the difficulties encountered while estimating average wages on the basis of Polish survey data. The second part describes data sources used for the estimation and provides additional details about the empirical study. The methodological considerations of the analysis are presented in the third part, the summary of the results and their interpretation in the fourth part, and the conclusions and suggestions for further work in the fifth part. The study presented in this paper is the continuation of the authors' previous research (Dehnel and Wawrowski 2018).

That study involved small enterprises employing from 10 to 49 employees. Owing to data availability, the analysis is limited to the year 2011, which was the time when the labour market was going through a downturn. The rate of registered unemployment grew to 12.5% compared to the previous year's 9.5% (according to the Labour Force Survey). The only positive trend that could be observed at that time was a rise in the average employment in the enterprise sector (a-7.6% growth in the construction section). The year 2011 also saw a slight increase in wages, but the rate of real wage growth was insignificant in view of the relatively large rise of consumer prices. The

average monthly wages was PLN 3481, which indicated the annual growth of 5.5%. However, the average monthly wages varied considerably across the groups of enterprises, depending on their size and type of activity. In small companies, the average gross wages totalled PLN 2583, ranging from PLN 1818 in companies involved in other service activities to PLN 4737 in information and communication companies; in medium-sized enterprises the average gross salary stood at PLN 3568, while in large enterprises at PLN 4255 (GUS 2013) (Figure1). In companies representing the four selected NACE sections, the average monthly salary reached similar levels, ranging from PLN 2150 in transportation enterprises to PLN 2460 in manufacturing companies.

There was also a considerable variation in average monthly salaries across provinces. The highest values were obtained for Mazowieckie province, which was an outlier. The average monthly wages in this province is considerably higher than the average monthly wages in the remaining provinces, regardless of the company size. Relatively high average monthly wages were also observed in Pomorskie, Dolnośląskie, Śląskie and Wielkopolskie.



C - Manufacturing
F - Construction
G - Trade and repair of motor vehicles
H - Transportation and storage
I - Accommodation and food service
J - activities
L - Information and communication
Real estate activities
M - Professional, scientific and technical
N - activities
P - Administrative and support service
Q - activities
R - Education
S - Human health and social work activities
Arts and recreation
Other service activities

Figure 1. Average monthly wage per 1 employee according to the enterprise's principal activity of size class in 2011

Source: Based on data by Statistics Poland (GUS 2013).

Figure 2 shows the average monthly wages for the whole subregion (NUTS 2), but does not account for the internal variation, which is explained, for example, by the classic theory of growth poles. In view of this fact, and also trying to satisfy the demand for detailed statistical information, an attempt was made to estimate the average monthly wages at a lower level of spatial aggregation, i.e. across districts.

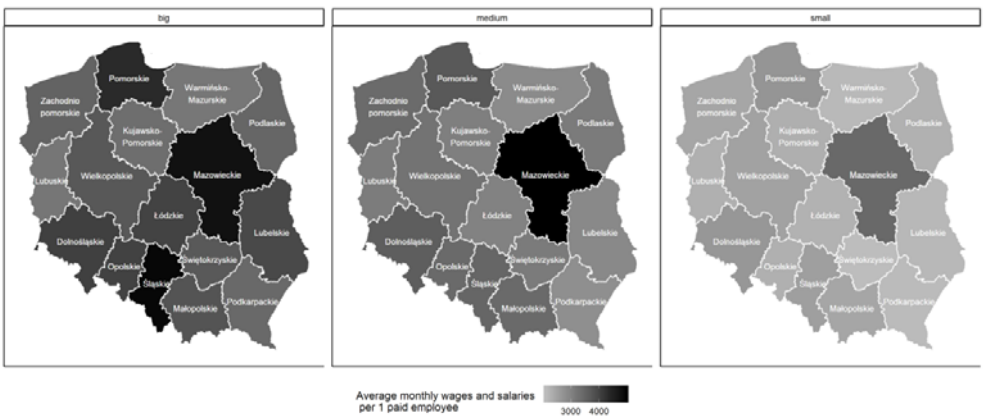


Figure 2. Average monthly wage per 1 employee across provinces by the enterprise's size class in 2011
Source: Based on data by Statistics Poland (GUS 2013).

2. Estimation methods

2.1. Direct estimation

The traditional approach in survey methodology to estimate population means and totals involves the use of the direct Horvitz-Thompson (1952) estimator. Let U denote a population consisting of N units divided into D domains $U_1, \ldots, U_D$ with the sample size denoted by N_d, where $d = 1, \ldots, D$. The sample denoted by s and $s \in U$ can also be divided into $s_1, \ldots, s_D$ with the sample size n_d for each domain.
Let y_{di} denote the value of the target variable for i-th unit in domain d. Each unit has a sampling weight w_{di}. The population mean in area d is denoted by θ_d. Thus, the Horvitz-Thompson estimator is expressed by the following formula:

$$\hat{\theta}_d^{HT} = \sum_{i=1}^{n_d} y_{di} w_{di}. \quad (1)$$

The Horvitz-Thompson (HT) estimator is unbiased and effective for large sample sizes n_d. However, when estimating detailed domains, sample sizes tend to be very small or even zero, which causes a big variance of the HT estimator and makes it impossible to obtain direct estimates.

2.2. Indirect estimation

To estimate population means in domains characterized by a small sample size n_d, it is necessary to use indirect estimators. This approach is also known as small area estimation, where the term *area* does not necessarily refer to a geographical unit. In this type of estimation, auxiliary variables from sources other than the survey are utilised. These variables should not contain sampling errors, so they should be taken from censuses or administrative registers.

The most common indirect approach is a model-based estimation, in which the target variable is the dependent variable in the linear mixed model. The methods within this approach can be divided into area-level and unit-level models. In an area-level model the dependent variable and auxiliary variables are aggregated at the target level of districts. Values of the dependent variables are often estimated by the Horvitz-Thompson (1952) estimator on the basis of survey data, while aggregated values of covariates come from the source that is not measured with error, such as the census or administrative registers (Guadarrama et al. 2016). The most popular example of an area-level model is the Fay-Herriot model (Fay and Herriot, 1979) and its numerous variants, e.g. spatial or robust (Rao and Molina 2015). On the other hand, there are unit-level models in which the dependent variable is taken directly from the survey and auxiliary variables also come from the census or administrative registers, but in a raw form. The Empirical Bayes method (Molina and Rao 2010) utilizes a nested error linear regression model and the Monte Carlo approximation to estimate the variable of interest in target domains, while the M-Quantile approach (Chambers and Tzavidis 2006) uses quantile regression to ensure robustness of that method. In both approaches, the random effect is usually assigned to the geographical area. Because access to unit-level data is limited, area-level models are used more frequently in practice.

Fay and Herriot used their model to estimate income in small geographical units in the USA (Fay and Herriot, 1979). However, the model's application is wider than that – e.g., it can be used for poverty estimation (Wawrowski 2016). Let us assume that the direct estimate of the population mean is the sum of true values of the parameter and random error, which is expressed by the formula:

$$\hat{\theta}_d^{HT} = \theta_d + e_d, \tag{2}$$

where $e_d \overset{ind}{\sim} N(0, \sigma_{ed}^2)$. In practice, variance σ_{ed}^2 is unknown and has been estimated on the basis of survey data.

Then, the true value of the parameter can be described by a linear model with area random effect:

$$\theta_d = x_d^T \beta + u_d, \tag{3}$$

where x_d is a vector of auxiliary information for area d, β is a vector of regression parameters and u_d is area random effect with distribution $u_d \overset{iid}{\sim} N(0, \sigma_u^2)$.

By combining the two equations above, we obtain the Fay-Herriot model:

$$\hat{\theta}_d^{HT} = x_d'\beta + u_d + e_d. \tag{4}$$

In order to obtain Empirical Best Linear Unbiased Predictor (EBLUP) of the Fay-Herriot model, it is necessary to estimate area random effect variance (σ_u^2). It can be done by various methods, e.g. Fay-Herriot method, Prasad-Rao method, REML or ML. Then, EBLUP is expressed by:

$$\hat{\theta}_d^{FH} = x_d^T \hat{\beta} + \hat{u}_d = \hat{\gamma}_d \hat{\theta}_d + (1 - \hat{\gamma}_d) x_d^T \hat{\beta}, \; d = 1, \ldots, D \tag{5}$$

where

$$\hat{\beta} = \left(\sum_{d=1}^{D} \hat{\gamma}_d \, x_d x_d^T \right)^{-1} \sum_{d=1}^{D} \hat{\gamma}_d \, x_d \, \hat{\theta}_d$$

and $\hat{\gamma}_d = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_{ed}^2}$.

It is worth remembering that EBLUP is a weighted average of direct and regression model estimates. The weight $\hat{\gamma}_d$ is a share of area random effect variance in the total variance and measures the uncertainty of the fitted model. For small values of sample variance ($\hat{\sigma}_{ed}^2$), a larger part of the final estimate will be contributed by the direct estimate, which will decrease as the sample variance increases (Rao and Molina, 2015).

The Fay-Herriot model is an example of a *shrinkage* estimator. Its drawback is the fact it cannot deal with outliers. For this reason, it is necessary to use robust small area estimation methods. In practice, this can be achieved by replacing $\hat{\beta}$ and $\hat{u}$ in the Fay-Herriot model with their outlier-resistant alternatives (Chambers et al., 2014).

Let us replace values of $\hat{\sigma}_{ed}^2$ and $\hat{\sigma}_u^2$ variances with covariance matrices Σ_e and Σ_u and let $V = \Sigma_e + \Sigma_u$. Then, the vector of fixed effects β is expressed by:

$$\beta = (X^T V^{-1} X)^{-1} X V^{-1} y \tag{6}$$

and random effects vector u is:

$$u = \Sigma_u Z^T V^{-1} (y - X\beta). \tag{7}$$

It can be noted that equations (6) and (7) could be transformed into:

$$X^T V^{-1} (y - X\beta) = 0 \tag{8}$$

and

$$\Sigma_u Z^T V - 1(y - X\beta) - u = 0. \tag{9}$$

Sinha and Rao (2009) proposed a robust version of equations (8) and (9):

$$X^T V^{-1} U^{1/2} \psi(U^{1/2}(y - X\beta)) = 0 \tag{10}$$

where $U = diag(V)$. A robust random effects vector is defined by:

$$\psi((y - X\beta)^T U^{\frac{1}{2}}) U^{\frac{1}{2}} V^{-1} (\partial V / \partial \theta) V^{-1} U^{\frac{1}{2}} \psi(U^{\frac{1}{2}}(y - X\beta)) = \mathrm{tr}(\mathrm{D}^{\Psi}(\partial \mathrm{V} / \partial \theta)), \tag{11}$$

where $\partial V / \partial \theta$ is the first order partial derivative of V with respect to the variance component θ and for $Z \sim N(0,1)$, $D^{\psi} = E(\psi^2(Z))V^{-1}$.

Moreover, Warnholz (2016) proposed a modification of the above equation in which only diagonal elements of the V matrix are used to standardise the residuals. In a robust Fay-Herriot model, this matrix is diagonal, but the transformation can be useful in models with correlated random effects, e.g. SAR(1) and AR(1) models, where calculations might be time-consuming.

The final equation of the robust Fay-Herriot model can be written as:

$$\hat{\theta}_d^{RFH} = x_d^T \hat{\beta}^{\psi} + \hat{u}_d^{\psi} \quad d = 1, \ldots, D. \tag{12}$$

The Fay-Herriot model and its robust version can also be used for estimation in non-sampled domains. In such cases, the estimated population mean is obtained from a regression model.

The mean square error (MSE) of the population means obtained with the help of the methods described above can be estimated using bootstrap methods. For the Horvitz-Thompson estimator, one can employ the standard replication weights procedure, whereas for the Fay-Herriot and robust Fay-Herriot models, the recommended option is parametric bootstrap proposed by González-Manteiga et al. (2008). Different estimates can be compared in terms of relative root mean square error (RRMSE), calculated as a square root of MSE divided by the estimate.

All methods described in this section are implemented in R language (R Core Team, 2018) in *survey* (Lumley, 2004), *sae* (Molina and Marhuenda, 2015) and *saeRobust* (Warnholz, 2018) packages.

3. Description of the DG1 dataset

The study is based on the data from the DG1 survey, which is the main source of information about Polish entrepreneurs. In the case of small companies, the survey uses a-10% sample of enterprises employing between 10 and 49 persons. These selected companies are then asked to complete a questionnaire about the basic company characteristics (Dehnel 2016).

The sampling design of the DG1 survey enables direct estimation while using the HT estimator, in order to obtain precise estimates at the level of province or for NACE sections. The structure of small companies in selected provinces and sections is presented in Table 1.

Table 1. Number of small enterprises in Mazowieckie and Wielkopolskie provinces in 2011

Province / NACE section	Wielkopolskie	Mazowieckie	Total	Mazowieckie (%)	Wielkopolskie (%)
Manufacturing	2782	2693	21583	12.9	12.5
Construction	1896	1423	12736	14.9	11.2
Trade	4038	2473	22677	17.8	10.9
Transportation	930	617	5000	18.6	12.3

Source: Based on data from the DG1 survey.

The number of manufacturing companies in both provinces is similar – 2782 in Mazowieckie and 2693 in Wielkopolskie, which accounts for about 13% of all enterprises in this section. Construction companies in Mazowieckie province account for 14.9% of the total number of units in the section, while in Wielkopolskie province for 11.2%. The biggest absolute and relative differences could be observed within the trade section – 4038 (17.8%) enterprises in Mazowieckie and 2473 (10.9%) in Wielkopolskie. Transportation is the smallest section, containing only 5000 companies from the whole country. Almost a fifth of them, 930 enterprises, are located in Mazowieckie province, whereas 617 in Wielkopolskie.

A detailed analysis of the number of companies at the level of district shows that there are no transportation enterprises in two districts (Lipski and Żuromiński) of Mazowieckie. These districts were excluded from the estimation process. Figure 3 shows the spatial distribution of population size at the level of district in each province.

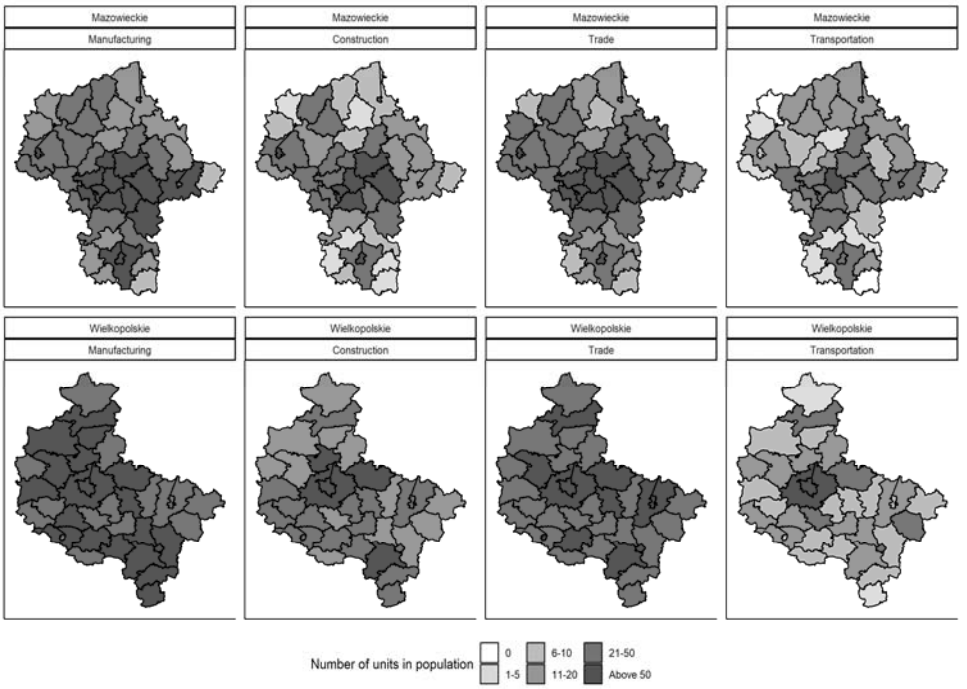


Figure 3. Number of enterprises in the population across districts in Mazowieckie and Wielkopolskie province

Source: Based on data from the DG1 survey.

Unit counts across districts indicate spatial variation as well as differentiation among and within the four NACE sections. The largest numbers of companies operate within the manufacturing and trade sectors, whereas the smallest in the construction and transportation. As already mentioned, the DG1 sample should include at least 10% of the population, and this condition is met. Table 2 shows the number of districts for different intervals of the sample size.

Table 2. Sample size at district level and NACE section

Province / NACE section	Number of units in the sample						
	Non-sampled	1	(1,5]	(5,10]	(10,20]	(20,50]	Above 50
Mazowieckie							
Construction	10	13	14	3	1	0	1
Manufacturing	3	3	15	15	3	2	1
Trade	4	5	21	5	5	1	1
Transportation	15	8	16	0	0	1	0

Table 2. Sample size at district level and NACE section (cont.)

Province / NACE section	Number of units in the sample						
	Non-sampled	1	(1,5]	(5,10]	(10,20]	(20,50]	Above 50
Wielkopolskie							
Construction	4	8	17	4	0	2	0
Manufacturing	0	0	6	14	12	2	1
Trade	0	1	13	14	5	1	1
Transportation	9	12	11	1	2	0	0

Source: Based on data from the DG1 survey.

As could be observed, the sample size in most districts lies within two intervals: (1-5] and (5-10]. There are also many unsampled domains. Interestingly, the number of unsampled districts is much higher in Mazowieckie province than in Wielkopolskie. The biggest sample sizes (above 20 enterprises) occurred in Warsaw, the capital of Poland, for the whole Mazowieckie province, for Poznań, the capital city of Wielkopolskie province, and for Poznański district that belongs to the areas surrounding Poznań (Figure 4). To facilitate the comparison, Figures 3 and 4 have the same legend. White areas denote non-sampled districts. The largest number of non-sampled districts can be observed while analysing the transportation section. This is mostly due to the small population sizes from this section.

In conclusion, the sample size across districts is usually small or even zero (Table 2). In this case, the use of direct estimation is either impossible or is likely to entail big values of the mean square error. One possible solution is to switch from the currently used methodology of estimation to new techniques of indirect estimation offered by small area estimation (SAE). The indirect estimation method was selected for this study mainly because of the characteristics of the enterprise population, i.e. its high degree of differentiation and the presence of outliers.

Figure 4. Number of units sampled across districts in Mazowieckie and Wielkopolskie province
Source: Based on data from the DG1 survey.

4. Estimation of average monthly wages at district level

The wage estimation process for selected domains consists of 4 stages: (1) direct estimation, (2) model fitting, (3) indirect estimation of population means for sampled and non-sampled domains and (4) MSE estimation. At the beginning, Horvitz-Thompson estimates were obtained for all sampled domains, together with their mean square errors. These estimates could be calculated for 201 of the 306 domains. For the remaining domains the sample was either zero or contained only one enterprise, which is not sufficient to estimate mean square error. The resulting direct estimates of the average wages ranged from PLN 2102 (section F, Mazowieckie province) to PLN 6494 (section G, Wielkopolskie). Using the rule of Tukey's fences (Hoaglin et al. 1986) for outlier identification, 11 outliers were detected, of which 8 in Wielkopolskie province. The outlying values represent estimated wages in the capitals of provinces, but also in districts where direct estimation was based on a very small sample, e.g. in Śremski district. The Horvitz-Thompson estimate, calculated there on the basis of just two units, amounted to PLN 6494. Consequently, the relative root mean square error was relatively high, at 40%. The above demonstrates that direct estimates calculated on the basis of small sample sizes are not reliable and cannot be analysed apart from their root mean squared errors.

There are no general precision thresholds for small sample domains. Eurostat (2013) recommends that they should be survey-specific, purpose-specific and should be determined taking users' needs into consideration. According to the guidelines by Statistics Poland, RRMSE of the estimates should not exceed 10%, and those above 20% should not be published (GUS 2013a), which is the usual publication practice (Tzavidis 2018). Table 3 presents domain frequencies for the specific RRMSE intervals.

Table 3. Number of districts with RRMSE of direct estimates within a given interval

Province / NACE section	Direct estimates RRMSE range		
	(0,10]	(10,20]	Above 20
Mazowieckie			
Construction	5	8	5
Manufacturing	16	16	5
Trade	11	19	4
Transportation	8	6	1
Wielkopolskie			
Construction	7	11	5
Manufacturing	23	8	3
Trade	17	13	4
Transportation	4	2	4

Source: Based on data from the DG1 survey.

For each NACE section, there is at least one district in each province where RRMSE of direct estimates exceeds 20%.

To improve the precision of direct estimates, the model-based approach was used. More than 20 variables measured at the level of district were examined as potential auxiliary variables for the model describing wages. The final model contains 4 variables: NACE section (X_1), income of enterprises drawn from the Ministry of Finance's registers (X_2), number of enterprises per 10 thousand of people (X_3) drawn from the Polish National Business Register (REGON), and the average monthly gross wages (X_4). Variable X_1 is categorical and the other three are continuous. Table 4 presents fixed effects of the Fay-Herriot and robust Fay-Herriot model and their standard errors (in brackets).

Table 4. Beta coefficients and their standard errors in the Fay-Herriot and the robust Fay-Herriot model

Variable	*Beta coefficients*	
	Fay-Herriot	Robust Fay-Herriot
Intercept	96.2421 (207.2802)	1011.2566 (231.1063)***
X_1 Construction	270.4663 (88.0113)**	1002.5664 (93.5293)***
X_1 Trade	207.8377 (208.0932)	208.7219 (75.2006)**
X_1 Transportation	-275.6738 (94.4806)**	-424.0788 (98.4990)***
X_2	0.0068 (0.0020)***	0.0071 (0.0021)***
X_3	1.1833 (0.1352)***	1.4082 (0.1607)***
X_4	0.3071 (0.0851)***	-0.0986 (0.1011)
P value significance codes: < 0.001 - ***; < 0.01 - **; < 0.05 - *		

Source: Based on data from the DG1 survey.

Compared to manufacturing, monthly wages in the construction sector were on average PLN 270 higher when estimated using the FH model, and PLN 1002 higher when using the RFH model. Compared to the trade section, they were higher by PLN 208 and PLN 209, respectively. A significant difference between the results yielded by the two models occurred also while estimating average wages in the transportation section. The average wages in that section was estimated at PLN 275 less than in the manufacturing section when using the FH model, and at PLN 424 less than in the manufacturing section when using the RFH model. Higher values of revenue and the number of enterprises per 10 thousand people in districts are correlated with higher wages. In the case of Fay-Herriot model, the average monthly gross wages can be interpreted in the same way. This covariate is insignificant in the robust Fay-Herriot model.

The beta coefficients presented in Table 4 were then used to estimate the average wages for non-sampled domains. Figure 5 presents the distribution of the estimation results.

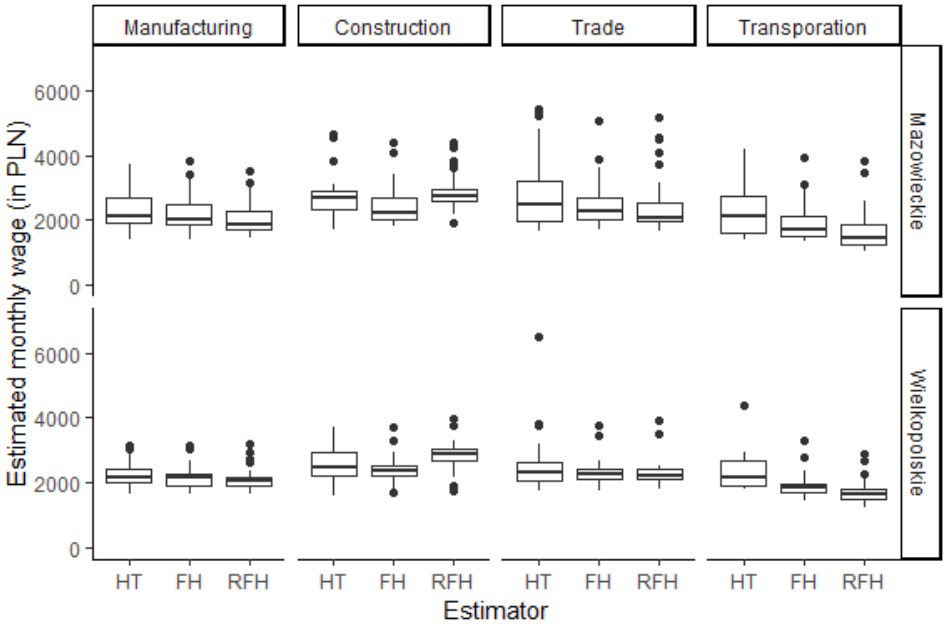


Figure 5. Distribution of estimates of average monthly wages obtained by using three approaches: Horvitz-Thompson estimator (HT), Fay-Herriot model (FH) and robust Fay-Herriot model (RFH)

Source: Based on data from the DG1 survey.

Unlike direct estimation, small area estimation methods make it possible to obtain estimates of an average monthly wages for all target domains (districts). For almost all domains, estimates based on the robust Fay-Herriot model are characterized by the lowest coefficient of variation, except for the trade and transportation sections in Mazowieckie province. The reduced impact of outliers is especially evident in the trade and transportation sections in Wielkopolskie province. The lowest wages are estimated for the transportation section, whereas the highest – over PLN 5000 – for the trade section in Warsaw. The comparison of average wages between the provinces indicates small differences. For example, the estimated average monthly wages in manufacturing companies in Mazowieckie province equals PLN 2060, while in Wielkopolskie it is PLN 2106. Bigger differences could be observed in the maximum wages – in the transportation section in Wielkopolskie it is PLN 2879, while in Mazowieckie PLN 3832.

Another aspect worth analysing is the precision of estimates measured by relative root mean square error. Distributions of these values are presented in Figure 6.

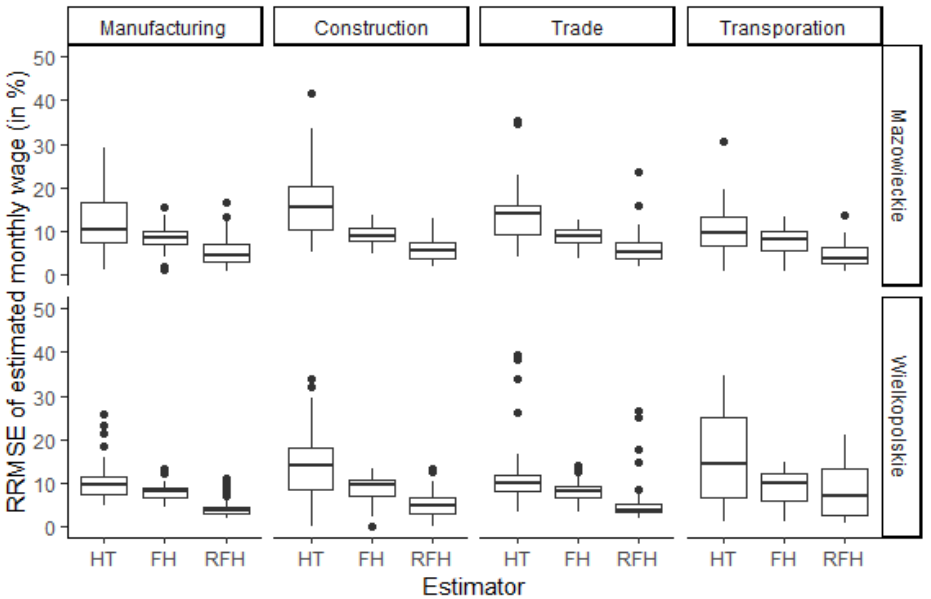


Figure 6. Distribution of RRMSE of the estimates of average monthly wage obtained through the three approaches - Horvitz-Thompson estimator (HT), Fay-Herriot model (FH) and robust Fay-Herriot model (RFH)

Source: Based on data from the DG1 survey.

It should be noted that both the Fay-Herriot model and its robust version are characterized by smaller values of RRMSE than the Horvitz-Thompson estimator. Because of many outliers in the trade section, RRMSE values for the RFH model in a few districts are higher than those for the FH model. However, descriptive statistics indicate that, on the whole, this approach yields more precise estimates than direct estimation. The median value of RRMSE for the robust Fay-Herriot model is at least half the value obtained for the Horvitz-Thompson, in all considered domains. The maximum value of RRMSE exceeds the 20% threshold only for 4 districts. Figure 7 shows a comparison of direct and indirect estimates using scatterplots.

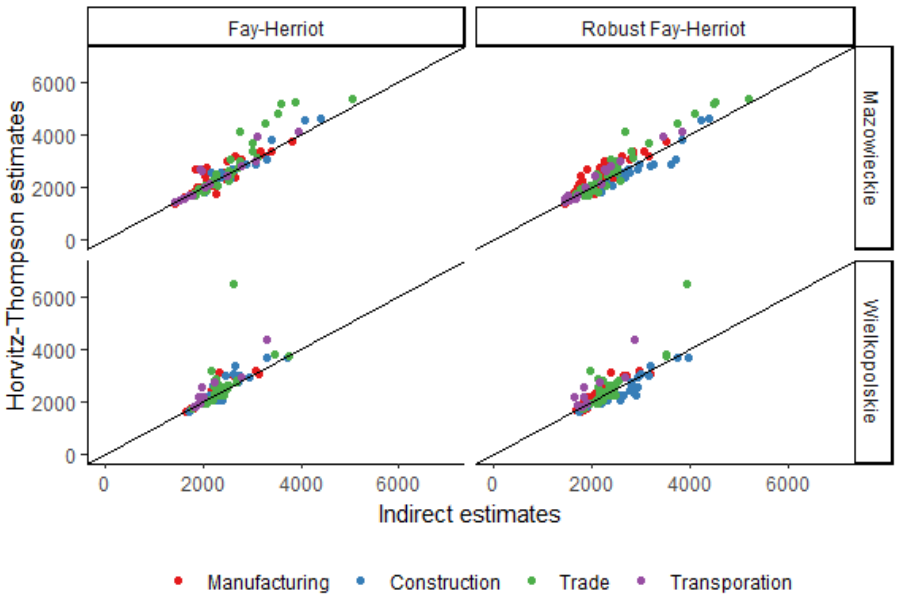


Figure 7. Comparison of estimates of average monthly wage at district level by NACE section
Source: Based on data from the DG1 survey.

The direct and indirect estimates are similar in most cases. The similarity of estimates measured by Pearson's coefficient of linear correlation is the highest for Mazowieckie and equals $r = 0.923$ for the Fay-Herriot model and $r = 0.926$ for the robust Fay-Herriot model. In the case of Wielkopolskie, the values are smaller, mainly due to the impact of the outlier (Śremski district). Pearson's coefficient of linear correlation for direct estimates and the FH model equals $r = 0.758$ and $r = 0.801$ for the robust model.

Figure 8 shows the correlation between RRMSE values obtained using different approaches. There is a difference in the shape of scatterplots for the two indirect approaches. In the case of the Fay-Herriot model, the precision of direct and indirect estimates is similar up to the level of 10%. For higher values of RRMSE, however, the FH model outperforms the direct estimator in terms of precision; this is the result of the γ weight, which accounts for the precision of the direct estimator. The use of the robust FH model improves the precision of most estimates but because values of fixed and random effects differ from those in the FH model, the resulting RRMSE values are relatively higher.

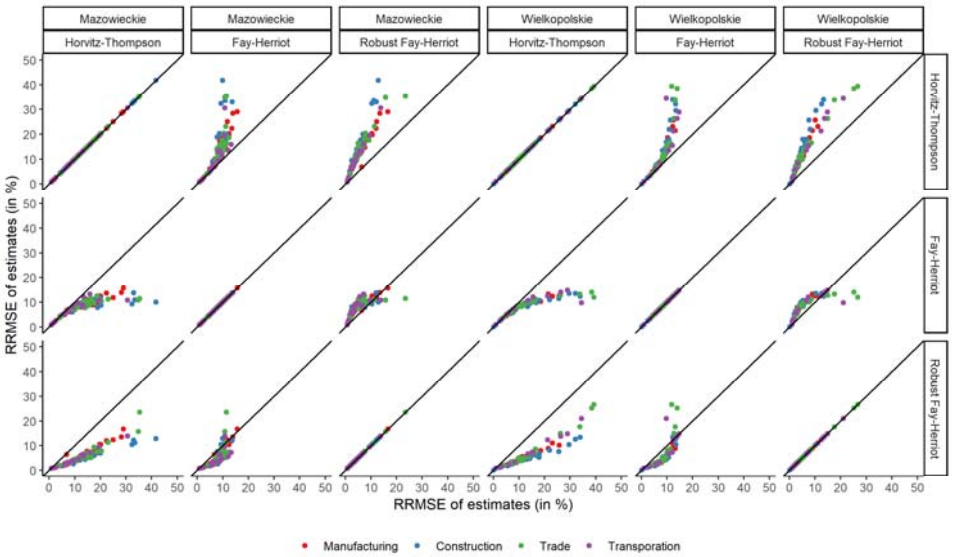


Figure 8. Comparison of RRMSE of estimates of average monthly wage at district level by NACE section

Source: Based on data from the DG1 survey.

As is demonstrated above, the results obtained by means of the Fay-Herriot model and its robust version are, in most cases, more precise in terms of RRMSE than direct estimates. The robust FH model is characterised by the highest average precision. Figure 9 shows mean estimates obtained using this model across districts.

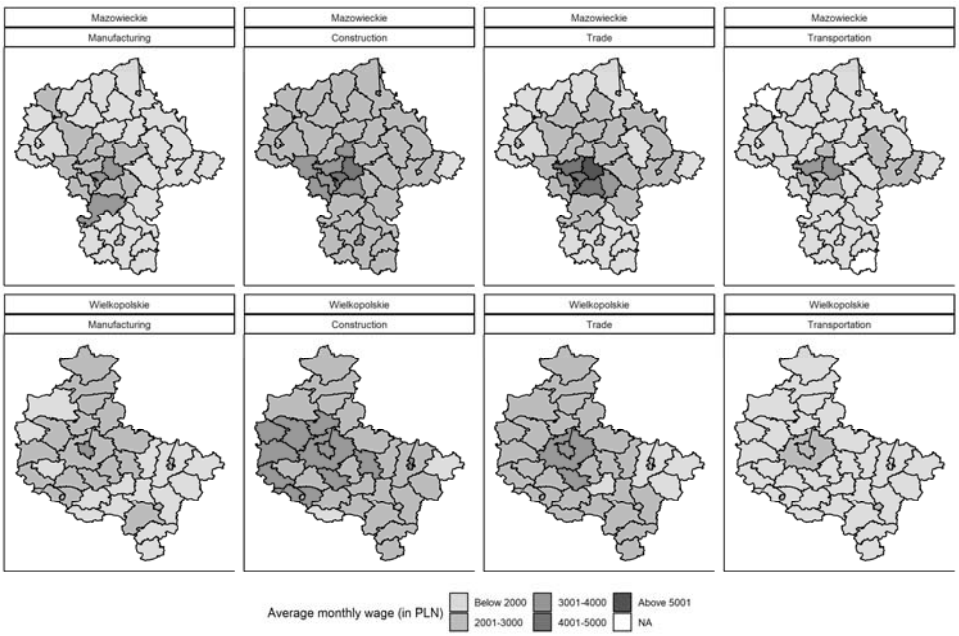


Figure 9. Spatial variation in average monthly wages at district level
Source: Based on data from the DG1 survey.

The highest estimates of average monthly wages were obtained for large cities and the neighbouring territorial units. In Mazowieckie province, the level of average wages was higher than in Wielkopolskie province.

The graphic presentation facilitate the identification of groups of similar districts characterised by similar values of the target variable and those that represent different level of average wages. The biggest difference between the units of Mazowieckie and Wielkopolskie provinces could be observed for province capitals and the neighbouring districts. Using the section of industry as the differentiating criterion, the highest average wages was observed in the group of construction companies, while the lowest in the group of transportation companies.

5. Conclusions

The article describes a study whose aim is to estimate the average monthly wages in the districts of Wielkopolskie and Mazowieckie provinces, for companies representing four major industry sections: manufacturing, construction, trade and transportation. Two methods of indirect estimation were applied: the FH model and its robust version. The analysis focused on districts for which no official estimates had been published before. A potential problem with performing analysis at such a low

level of aggregation is that sample sizes in districts are relatively small. However, thanks to the use of small area estimation methods, it was possible to obtain relatively precise estimates, i.e. with lower values of RRMSE compared to the results obtained using direct estimates. Robust estimation affected outlier values of monthly wages and decreased the range of estimates. Moreover, the use of auxiliary variables for indirect estimation made it possible to obtain estimates of the average wages in non-sampled domains. The study boasts a degree of novelty, because it made it possible to estimate monthly wages in small enterprises according to NACE sections, at the level of districts, for the very first time. The highest estimates of average wages were obtained for large cities and their neighbouring districts in all the four main NACE sections. The results of the study moreover indicate that both Warsaw and Poznań serve as poles of growth for the neighbouring districts.

Further work into the subject will focus on the application of robust area-level models with spatial autocorrelation (Warnholz 2016), and, depending on data availability, unit-level models (Chambers and Tzavidis 2006).

Acknowledgements

The project is financed by the Polish National Science Centre DEC-2015/17/B/HS4/00905.

REFERENCES

CHAMBERS, R., TZAVIDIS, N., (2006). M-quantile models for small area estimation, Biometrika, 93(2), pp. 255–268.

CHAMBERS, R., CHANDRA, H., SALVATI, N., TZAVIDIS, N., (2014). Outlier robust small area estimation, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(1), pp.47–69.

DEHNEL, G., (2016). M-estimators in business statistics. Statistics in Transition, New Series, 2016, Vol. 17, No. 4, pp. 749–762, ISSN 1234-7655, http://dx.doi.org/10.21307/stattrans-2016-050.

DEHNEL, G., WAWROWSKI. Ł., (2018). Robust Estimation Of Revenues Of Polish Small Companies By NACE Section And Province, In: M. Papież and S. Śmiech (Eds.), The 12th Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Conference Proceedings, Foundation of the Cracow University of Economics, Cracow, pp. 110–119.
URL http://dx.doi.org/10.14659/SEMF.2018.01.11.

EUROSTAT, (2013). Handbook on precision requirements and variance estimation for ESS households surveys, European Commission, Belgium, DOI:10.2785/13579.

FAY III, R. E., HERRIOT, R. A., (1979). Estimates of income for small places: an application of James-Stein procedures to census data, Journal of the American Statistical Association, 74(366a), pp. 269–277.

GONZÁLEZ-MANTEIGA, W., LOMBARDIA, M., MOLINA, I., MORALES, D., SANTAMARIA, L., (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. Computational Statistics and Data Analysis, 52(12), pp. 5242–5252, http://EconPapers.repec.org/ RePEc:eee:csdana:v:52:y:2008:i:12:p:5242-5252.

GUADARRAMA, M., MOLINA, I., RAO, J. N. K., (2016). A comparison of small area estimation methods for poverty mapping, Statistics in Transition new series, 1(17), pp. 41–66.

GUS, (2013). Działalność przedsiębiorstw niefinansowych w 2011 r., Zakład Wydawnictw Statystycznych, Warszawa, (Activity of Non-financial Enterprises in 2011), Central Statistical Office of Poland, Warszawa. URL, http://stat.gov.pl/obszary-tematyczne/podmioty-gospodarcze-wyniki-finansowe/przedsiebiorstwa-niefinansowe/dzialalnosc-przedsiebiorstw-niefinansowych-w-2016-r-,2,12.html [Accessed 14 November 2018].

GUS, (2013a). Narodowy Spis Powszechny Ludności i Mieszkań. Ludność. Stan i struktura demograficzno-społeczna, Zakład Wydawnictw Statystycznych, Warszawa, URL http://stat.gov.pl/spisy-powszechne/nsp-2011/nsp-2011-wyniki/ludnosc-stan-i-struktura-demograficzno-spoleczna-nsp-2011,16,1.html [Accessed 14 November 2018].

HOAGLIN, D. C., IGLEWICZ, B., TUKEY, J. W., (1986). Performance of some resistant rules for outlier labeling, Journal of the American Statistical Association, 81(396), pp. 991–999.

HORVITZ, D. G., THOMPSON, D. J., (1952). A generalization of sampling without replacement from a finite universe, Journal of the American statistical Association, 47(260), pp.663–685.

LUMLEY, T., (2004). Analysis of complex survey samples, Journal of Statistical Software 9(1): pp.1–19

MOLINA, I., MARHUENDA, Y., (2015). "sae: An R Package for Small Area Estimation", The R Journal, 7(1), pp. 81–98. https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf

MOLINA, I., RAO, J. N. K., (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics, 38(3), pp. 369–385.

R Core Team, (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RAO, J. N. K., MOLINA, I., (2015). Small area estimation, John Wiley & Sons.

SINHA, S. K., RAO, J. N. K., (2009). Robust small area estimation, Canadian Journal of Statistics, 37(3), pp.381–399.

STRUŻYCKI, M., (2004). Małe i średnie przedsiębiorstwa w gospodarce regionu, PWE, Warszawa.

TZAVIDIS, N., ZHANG, L. C., LUNA, A., SCHMID, T., ROJAS-PERILLA, N., (2018). From start to finish: a framework for the production of small area official statistics, Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(4), pp. 927–979.

WARNHOLZ, S., (2016). Small Area Estimation Using Robust Extensions to Area Level Models, Doctoral dissertation, Freie Universität Berlin.

WARNHOLZ, S., (2018). saeRobust: Robust Small Area Estimation. R package version 0.2.0, https://CRAN.R-project.org/package=saeRobust.

WAWROWSKI, Ł., (2016). The Spatial Fay-Herriot Model in Poverty Estimation, Folia Oeconomica Stetinensia, 16(2), pp. 191–202.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 159–168, DOI 10.21307/stattrans-2020-009
Submitted – 17.10.2019; accepted for publishing – 30.01.2020

Generalized exponential estimators for the finite population mean

Tolga Zaman[1]

ABSTRACT

This study proposes a new class of exponential-type estimators in simple random sampling for the estimation of the population mean of the study variable using information of the population proportion possessing certain attributes. Theoretically, mean squared error (MSE) equations of the suggested ratio exponential estimators are obtained and compared with the Naik and Gupta (1996) ratio and product estimators, the ratio and product exponential estimator presented in Singh et al. (2007) and the ratio exponential estimators presented in Zaman and Kadilar (2019a). As a result of these comparisons, it is observed that the proposed estimators always produce more efficient results than the others. In addition, these theoretical results are supported by the application of original datasets.

Key words: ratio-exponential estimators, auxiliary attribute, mean square error, efficiency.

Mathematical classification: 62D05

1. Introduction

When there is a positive correlation between the study variable and the auxiliary variable in the simple random sampling method, ratio-type estimators are used to estimate the population mean. But when the correlation coefficient between the study and auxiliary variables is negative, the product estimator is used. Many authors have proposed estimators based on auxiliary attribute. Bahl and Tuteja (1991) estimator and a family of estimators considered by Jhajj et al. (2006), Singh et al. (2008), Koyuncu (2012), Malik and Singh (2013), Shabbir and Gupta (2010), Solanki and Singh (2013), Zaman (2018), Zaman and Kadilar (2019b) suggested a class of estimators by using auxiliary attributes. In this paper, a new class of exponential type estimators is proposed to estimate the population mean for the study variable using information on auxiliary attributes.

[1] Çankırı Karatekin University, Faculty of Science, Department of Statistics, 18100 Çankırı, Turkey.
E-mail: tolgazaman@karatekin.edu.tr. ORCID: https://orcid.org/0000-0001-8780-3655.

Let y_i be i-th characteristic of the population and ϕ_i is the case of possessing certain attributes. If ith unit has the desired characteristic, it takes the value 1; if not then the value 0. That is:

$$\phi_i = \begin{cases} 1 & , \quad if\ ith\ unit\ of\ the\ population\ possesses\ attribute \\ 0 & , \quad otherwise. \end{cases}$$

Let $A = \sum_{i=1}^{N} \phi_i$ and $a = \sum_{i=1}^{n} \phi_i$ be the total count of the units that possess certain attribute in the population and the sample, respectively. And $P = \frac{A}{N}$ and $p = \frac{a}{n}$ are the ratio of these units, respectively.

The Naik and Gupta (1996) estimator for the population mean $\bar{Y}$ of the variate of study, which makes use of information concerning the population proportion possessing certain attribute, is defined by

When the relationship between the study variable and the auxiliary attribute is positive, the ratio estimator of the population proportion of the study variable,

$$\bar{y}_{NG1} = \frac{\bar{y}}{p} P \tag{1.1}$$

When the relationship between the study variable and the auxiliary attribute is negative, the product estimator of the population proportion of the study variable,

$$\bar{y}_{NG2} = \frac{\bar{y}}{P} p \tag{1.2}$$

where it is assumed that the population proportion P of the form of attribute ϕ is known.

The expressions of MSE of the Naik and Gupta estimators are

$$MSE(\bar{y}_{NG1}) \cong \frac{1-f}{n} \bar{Y}^2 \left(C_y^2 - 2\rho_{pb} C_y C_p + C_p^2\right) \tag{1.3}$$

$$MSE(\bar{y}_{NG2}) \cong \frac{1-f}{n} \bar{Y}^2 \left(C_y^2 + 2\rho_{pb} C_y C_p + C_p^2\right) \tag{1.4}$$

Considering Bahl and Tuteja (1991), Singh et al. (2007) suggested the following ratio estimator when the study variable and the auxiliary attribute are positively correlated.

$$t_1 = \bar{y} exp\left(\frac{P-p}{P+p}\right) \tag{1.5}$$

Singh et al. (2007) suggested the following ratio estimator when the study variable and the auxiliary attribute are negatively correlated.

$$t_2 = \bar{y} exp\left(\frac{p-P}{p+P}\right) \tag{1.6}$$

The MSEs of these estimators are

$$MSE(t_1) \cong \frac{1-f}{n}\bar{Y}^2\left(C_y^2 - \rho_{pb}C_yC_p + \frac{C_p^2}{4}\right) \tag{1.7}$$

$$MSE(t_2) \cong \frac{1-f}{n}\bar{Y}^2\left(C_y^2 + \rho_{pb}C_yC_p + \frac{C_p^2}{4}\right) \tag{1.8}$$

Zaman and Kadilar (2019a) suggested modified exponential ratio estimators using auxiliary attribute information for estimating $\bar{Y}$ as

$$t_{ZKi} = \bar{y}exp\left[\frac{(kP+l)-(kp+l)}{(kP+l)+(kp+l)}\right] \quad i = 1,2,\dots,9 \tag{1.9}$$

where $k(\neq 0)$ and l are either the real number or the functions of the known parameters of the attribute, C_p, $\beta_2(\phi)$ and the known parameter of the attribute with the study variable, ρ_{pb}. Note that the sum of k and l is not necessarily equal to one.

The MSE of these estimators

$$MSE(t_{ZKi}) \cong \frac{1-f}{n}\bar{Y}^2\left[\theta_i^2C_p^2 - 2\theta_i\rho_{pb}C_yC_p + C_y^2\right] \ ,i = 1,\dots,9 \tag{1.10}$$

$$\theta_1 = \frac{P}{2(P+\beta_2(\phi))};\ \theta_2 = \frac{P}{2(P+C_p)};\ \theta_3 = \frac{P}{2(P+\rho_{pb})};\ \theta_4 = \frac{\beta_2(\phi)P}{2(\beta_2(\phi)P+C_p)};\ \theta_5 = \frac{C_pP}{2\left(C_pP+\beta_2(\phi)\right)}$$

$$\theta_6 = \frac{C_pP}{2(C_pP+\rho_{pb})};\ \theta_7 = \frac{\rho_{pb}P}{2(\rho_{pb}P+C_p)};\ \theta_8 = \frac{\beta_2(\phi)P}{2(\beta_2(\phi)P+\rho_{pb})};\ \theta_9 = \frac{\rho_{pb}P}{2\left(\rho_{pb}P+\beta_2(\phi)\right)}.$$

where $f = \frac{n}{N}$; N is the number of units in the population; C_p is the coefficient of population variation of the form of attribute and C_y is the coefficient of population variation of the study variable. ρ_{pb} is the point biserial correlation coefficient. $\beta_2(\phi)$ is the coefficient of population kurtosis of the auxiliary attribute.

2. Suggested estimators

Following Ozel (2016), the improved class of estimator $\bar{y}_{pr}$ for the population is proposed as follows

$$\bar{y}_{pri} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{(kP+l)-(kp+l)}{(kP+l)+(kp+l)}\right]; i = 1,2,\dots,10 \tag{2.1}$$

A class of new estimators generated from Equation (2.1) is listed in Table 1.

Table 1. Proposed Estimators

Estimators	Values of	
	k	l
$\bar{y}_{pr1} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{P-p}{P+p}\right]$	1	0
$\bar{y}_{pr2} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{P-p}{P+p+2\beta_2(\phi)}\right]$	1	$\beta_2(\phi)$
$\bar{y}_{pr3} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{P-p}{P+p+2C_p}\right]$	1	C_p
$\bar{y}_{pr4} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{P-p}{P+p+2\rho_{pb}}\right]$	1	ρ_{pb}
$\bar{y}_{pr5} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{\beta_2(\phi)(P-p)}{\beta_2(\phi)(P+p)+2C_p}\right]$	$\beta_2(\phi)$	C_p
$\bar{y}_{pr6} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{C_p(P-p)}{C_p(P+p)+2\beta_2(\phi)}\right]$	C_p	$\beta_2(\phi)$
$\bar{y}_{pr7} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{C_p(P-p)}{C_p(P+p)+2\rho_{pb}}\right]$	C_p	ρ_{pb}
$\bar{y}_{pr8} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{\rho_{pb}(P-p)}{\rho_{pb}(P+p)+2C_p}\right]$	ρ_{pb}	C_p
$\bar{y}_{pr9} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{\beta_2(\phi)(P-p)}{\beta_2(\phi)(P+p)+2\rho_{pb}}\right]$	$\beta_2(\phi)$	ρ_{pb}
$\bar{y}_{pr10} = \bar{y}\left(\frac{p}{P}\right)^{\alpha} exp\left[\frac{\rho_{pb}(P-p)}{\rho_{pb}(P+p)+2\beta_2(\phi)}\right]$	ρ_{pb}	$\beta_2(\phi)$

In Table 1, C_p, $\beta_2(\phi)$ and ρ_{pb} are coefficient of variation, coefficient of population kurtosis of the form of the auxiliary attribute and the point biserial population correlation coefficient between the auxiliary attribute and the study variable, respectively. $\bar{y}$ and p are the sample mean belonging to the study variable and the sample proportion possessing certain attributes, respectively.

To obtain the MSE expression of these estimators $\bar{y}_{pri}$, let $\bar{y} = \bar{Y}(1+e_0)$ and $p = P(1+e_1)$ such that $E(e_0) = E(e_1) = 0$ and $E(e_0^2) = \frac{1-f}{n}C_y^2$, $E(e_1^2) = \frac{1-f}{n}C_p^2$ and $E(e_0e_1) = \frac{1-f}{n}C_{yp} = \frac{1-f}{n}\rho_{pb}C_yC_p$.

Expressing the estimator $\bar{y}_{pri}$ in terms of e_j, $(j = 0, 1)$ and shortened terms up to the first two and the first three components of e's as follows

$$\bar{y}_{pri} = \bar{Y}(1+e_0)(1+e_1)^{\alpha}exp\{-\theta e_1(1+\theta e_1)^{-1}\}$$

$$\bar{y}_{pri} = \bar{Y}(1+e_0)\left[1+\alpha e_1+\frac{\alpha(\alpha-1)}{2}e_1^2+\cdots\right][1+(-\theta e_1)(1+\theta e_1)^{-1}]$$

$$= \bar{Y}(1+e_0)\left[1+\alpha e_1+\frac{\alpha(\alpha-1)}{2}e_1^2+\cdots\right][1-\theta e_1+\theta^2 e_1^2] \tag{2.2}$$

Expanding the right-hand side of (2.2) to the first order approximation and subtracting $\bar{Y}$ from both sides, the following expression is obtained

$$\bar{y}_{pri}-\bar{Y} \cong \bar{Y}\left[-\theta_i e_1+\theta_i^2 e_1^2+\alpha e_1-\alpha\theta_i e_1^2+\frac{\alpha(\alpha-1)}{2}e_1^2+e_0-\theta_i e_0 e_1+\alpha e_0 e_1\right] \tag{2.3}$$

Squaring both sides of equation (2.3) and then taking the expectation of both sides, the following expression of MSE of the estimator $\bar{y}_{pri}$ is obtained

$$MSE(\bar{y}_{pri}) \cong \frac{1-f}{n}\bar{Y}^2\left[\alpha^2 C_p^2+\theta_i^2 C_p^2-2\alpha\theta_i C_p^2+C_y^2+2\alpha\rho_{pb}C_y C_p-2\theta_i C_y C_p\right]; i = 1,2,\ldots,10 \tag{2.4}$$

Using the following equations, the optimal value of α can be obtained:

$$\frac{\partial MSE(\bar{y}_{pri})}{\partial\alpha}=\frac{1-f}{n}\bar{Y}^2\left(2\alpha C_p^2-2\theta_i C_p^2+2\rho_{pb}C_y C_p\right)=0$$
$$\alpha_{opti}=\frac{\theta_i C_p-\rho_{pb}C_y}{C_p}; i = 1,2,..,10 \tag{2.5}$$

The minimum MSE of the proposed estimator can be calculated using α equations in (2.5). The suggested estimators have the same minimum MSE as follows:

$$MSE_{min}(\bar{y}_{pr}) \cong \frac{1-f}{n}\bar{Y}^2 C_y^2\left(1-\rho_{pb}^2\right) \tag{2.6}$$

3. Efficiency comparisons

The expressions of MSE of classical estimators $\bar{y}$, $\bar{y}_{NG1}$, $\bar{y}_{NG2}$, t_1, t_2 and zt_i are compared with the MSE of the suggested estimator $\bar{y}_{pri}$.

Under simple random sampling without replacement (SRSWOR) the variance of the sample mean is

$$V(\bar{y})=\frac{1-f}{n}\bar{Y}^2 C_y^2 \tag{3.1}$$

From (2.6) and (3.1), we have

$$\begin{gathered} MSE_{min}(\bar{y}_{pri}) < V(\bar{y}) \\ \frac{1-f}{n}\bar{Y}^2 C_y^2(1-\rho_{pb}^2) < \frac{1-f}{n}\bar{Y}^2 C_y^2 \\ \rho_{pb}^2 > 0 \end{gathered} \tag{3.2}$$

From (2.6) and (1.3), we have

$$\begin{gathered} MSE_{min}(\bar{y}_{pri}) < MSE(\bar{y}_{NG1}) \\ \frac{1-f}{n}\bar{Y}^2 C_y^2(1-\rho_{pb}^2) < \frac{1-f}{n}\bar{Y}^2(C_y^2 - 2\rho_{pb}C_yC_p + C_p^2) \\ (C_y\rho_{pb} - C_p)^2 > 0 \end{gathered} \tag{3.3}$$

From (2.6) and (1.4), we have

$$\begin{gathered} MSE_{min}(\bar{y}_{pri}) < MSE(\bar{y}_{NG2}) \\ \frac{1-f}{n}\bar{Y}^2 C_y^2(1-\rho_{pb}^2) < \frac{1-f}{n}\bar{Y}^2(C_y^2 + 2\rho_{pb}C_yC_p + C_p^2) \\ (C_y\rho_{pb} + C_p)^2 > 0 \end{gathered} \tag{3.4}$$

From (2.6) and (1.7), we have

$$\begin{gathered} MSE_{min}(\bar{y}_{pri}) < MSE(t_1) \\ \frac{1-f}{n}\bar{Y}^2 C_y^2(1-\rho_{pb}^2) < \frac{1-f}{n}\bar{Y}^2\left(C_y^2 - \rho_{pb}C_yC_p + \frac{C_p^2}{4}\right) \\ \left(C_y\rho_{pb} - \frac{C_p}{2}\right)^2 > 0 \end{gathered} \tag{3.5}$$

From (2.6) and (1.8), we have

$$\begin{gathered} MSE_{min}(\bar{y}_{pri}) < MSE(t_2) \\ \frac{1-f}{n}\bar{Y}^2 C_y^2(1-\rho_{pb}^2) < \frac{1-f}{n}\bar{Y}^2\left(C_y^2 + \rho_{pb}C_yC_p + \frac{C_p^2}{4}\right) \\ \left(C_y\rho_{pb} + \frac{C_p}{2}\right)^2 > 0 \end{gathered} \tag{3.6}$$

From (2.6) and (1.10), we have

$$MSE_{min}(\bar{y}_{pri}) < MSE(t_{ZKi})$$
$$\frac{1-f}{n}\bar{Y}^2C_y^2(1-\rho_{pb}^2) < \frac{1-f}{n}\bar{Y}^2\left[\theta_i^2C_p^2 - 2\theta_i\rho_{pb}C_yC_p + C_y^2\right]$$
$$\left(C_y\rho_{pb} - \theta_iC_p\right)^2 > 0 \tag{3.7}$$

We infer that the suggested exponential estimators $\bar{y}_{pr}$ perform better than the competing estimators in all conditions because the condition given in (3.2)-(3.7) is always satisfied.

4. Empirical study

For empirical study, we use the data given in Sukhatme and Sukhatme (1970) and Zaman et al. (2014).

The values of the required parameters for these data sets are given in Tables 2 and 3.

Population 1: The data is defined as follows:

$$y = \text{the number of villages in the circles}$$
$$\phi_i = \begin{cases} 1 & , \text{ if a circle consists of more than five villages} \\ 0 & , \text{ otherwise.} \end{cases}$$

Table 2. Data Statistics of Population 1

N=89	$\overline{Y} = 3.3596$	θ_1 =0.0171	θ_5 =0.0433	θ_9 =0.0132
n= 20	$P = 0.1236$	θ_2 =0.0221	θ_6 =0.1508	
$\beta_2(\phi) = 3.492$	$C_y = 0.6008$	θ_3 =0.0695	θ_7 =0.0171	
$\rho_{pb} = 0.766$	$C_p = 2.6779$	θ_4 =0.0694	θ_8 =0.1802	

Population 2: The data is defined as follows:

$$y = \text{the number of teachers}$$
$$\phi_i = \begin{cases} 1 & , \text{ if the number of teachers is more than 60} \\ 0 & , \text{ otherwise.} \end{cases}$$

Table 3. Data Statistics of Population 2

N=111	$\overline{Y} = 29.279$	θ_1 =0.0146	θ_5 =0.0382	θ_9 =0.0117
n= 30	$P = 0.117$	θ_2 =0.0203	θ_6 =0.1441	
$\beta_2(\phi) = 3.898$	$C_y = 0.872$	θ_3 =0.0640	θ_7 =0.0164	
$\rho_{pb} = 0.797$	$C_p = 2.758$	θ_4 =0.0709	θ_8 =0.1819	

Table 4. MSE Values of the Classical and Suggested Estimators

Estimator	MSE	
	Population 1	Population 2
$\bar{y}$	0.1579	15.8557
$\bar{y}_{NG1}$	2.2168	94.532
$\bar{y}_{NG2}$	4.3742	254.4077
t_1	0.4030	15.5403
t_2	1.4817	95.4782
t_{ZK1}	0.1404	14.7247
t_{ZK2}	0.1357	14.2948
t_{ZK3}	0.0981	11.3891
t_{ZK4}	0.0982	10.9827
t_{ZK5}	0.1171	13.0318
t_{ZK6}	0.0667	7.6304
t_{ZK7}	0.1404	14.5909
t_{ZK8}	0.0654	6.5614
t_{ZK9}	0.1442	14.9436
$\bar{y}_{pr}$	**0.0652**	**5.7840**

From the Table 4, it is observed that the suggested exponential estimators $\bar{y}_{pri}$, $(i = 1,2,\ldots,10)$ perform better than the usual unbiased estimator $\bar{y}$, ratio and product estimators of Naik and Gupta (1996), ratio and product estimators suggested by Singh et al. (2007) and the ratio exponential estimators presented in Zaman and Kadilar (2019a). Finally, it is inferred that the suggested estimators perform better than the considered ratio estimators in all conditions, because the conditions given in Section 3 are always satisfied.

5. Conclusion

This paper proposes exponential families of estimators and presents a theoretical argument. All of the suggested estimators have the same minimum MSE equation. Considering MSE equations, the suggested exponential estimators are always more efficient than those of the sample mean, the ratio and product estimators of Naik and Gupta (1996), the ratio and product estimators of Singh et al. (2007) and the ratio exponential estimators of Zaman and Kadilar (2019a), under all the conditions. The results presented here support these conclusions by theoretical development and numerical analysis. In forthcoming studies, we hope to extend the suggested class of estimators presented in this article to the stratified random sampling.

REFERENCES

BAHL, S., TUTEJA, R. K., (1991). Ratio and product type estimator, Information and Optimization Science, Vol. XII, pp. 159–163.

JHAJJ, H. S., SHARMA, M. K., GROVE, L. K., (2006). A family of estimators of population mean using information on auxiliary attribute, Pakistan Journal of Statistics 22 (1), pp. 43–50.

KOYUNCU, N., (2012). Efficient estimators of population mean using auxiliary attributes, Applied Mathematics and Computation, 218(22), pp. 10900–10905.

MALIK, S., SINGH, R., (2013). An improved estimator using two auxiliary attributes, Applied Mathematics and Computation, 219(23), pp. 10983–10986

NAIK, V. D., GUPTA, P. C., (1996). A note on estimation of mean with known population proportion of an auxiliary character, Jour. Ind. Soc. Agr. Stat., 48 (2), pp. 151–158.

OZEL, K. G., (2016). A New Exponential Type Estimator for the Population Mean in Simple Random Sampling, Journal of Modern Applied Statistical Methods, 15(2), pp. 207–214

SHABBIR, J., GUPTA, S., (2010). Estimation of the finite population mean in two phase sampling when auxiliary variables are attributes, Hacettepe Journal of Mathematics and Statistics, 39(1).

SINGH, R., CHAUHAN, P., SAWAN, N., SMARANDACHE, F., (2007). Ratio-product type exponential estimator for estimating finite population mean using information on auxiliary attribute, in "Auxiliary Information and a Priori Values in Construction of Improved Estimators" edited by Singh, R., Chauhan, P., Sawan, N. and Smarandache, F., Renaissance High Press, pp. 18–32.

SINGH, R., CHAUHAN, P., SAWAN, N., (2008). On linear combination of Ratio-product type exponential estimator for estimating finite population mean, Statistics in Transition, 9(1), pp. 105–115.

SOLANKI, R. S., SINGH, H. P., (2013). Improved estimation of population mean using population proportion of an auxiliary character, Chilean journal of Statistics, 4(1), pp. 3–17.

SUKHATME, P. V., SUKHATME, B. V., (1970). Sampling Theory of Surveys with Applications, Iowa State University Press, Ames, U.S.A.

ZAMAN, T., SAGLAM, V., SAGIR, M., YUCESOY, E., ZOBU, M., (2014). Investigation of some estimators via Taylor series approach and an application, American Journal of Theoretical and Applied Statistics, 3(5), pp. 141–147.

ZAMAN, T., (2018). New family of estimators using two auxiliary attributes, International Journal of Advanced Research I Engineering & Management (IJAREM), 4(11), pp. 11–16.

ZAMAN, T., KADILAR, C., (2019a). Novel family of exponential estimators using information of auxiliary attribute, Journal of Statistics and Management Systems, pp. 1–11.

ZAMAN, T., KADILAR, C., (2019b). New class of exponential estimators for finite population mean in two-phase sampling, Communications in Statistics – Theory and Methods, pp. 1–16.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1, pp. 169–176, DOI 10.21307/stattrans-2020-010
Submitted – submitted 31.12.2017; accepted for publishing – 28.11.2019

Alternative approach to moments of order statistics from one-parameter Weibull distribution

Piyush Kant Rai[1], **Anu Sirohi**[2]

ABSTRACT

The Weibull distribution is used to describe various observed failures of phenomena and widely used in survival analysis and reliability theory. Sometimes it is very difficult to compute moments of such distributions due to various reasons for e.g. analytical issues, multi parameter cases etc. This study presents the computation of the moments and the expected value of the product of order statistics in the sample from the one-parameter Weibull distribution. An alternative approach in connection to survival function is used to obtain these moments and expected values. In addition the characteristic function of the above distribution is also obtained in the form of gamma functions. Further an illustration is shown to find the first two moments and expected value of the product of order statistics by using this approach.

Key words: order statistics, survival function, moments, characteristic function.

1. Introduction

Order statistics deals with the properties and applications of ordered random variables. The concept of order statistics is widely used and play a very important role particularly in life testing experiments, in a variety of practical situations where some of the observations in the sample are censored. Since these experiments may take a long time to complete, usually it is desirable to stop after the first r failure out of n items under the test. The observations are basically the times of r failures, unlike most situations, and are obtained in order by the method of experimentation. Using the obtained data we can estimate the required parameter of interest such as true mean lifetime.

Other occurrences arise in the reliability theory and survival analysis where, a system of n components is called r out of n system if at least r components occur. For components with independent lifetime distributions $F_1, F_2, ..., F_n$ the time to failure of the system is seen to be the $(n-r+1)th$ order statistics from the set of underlying heterogeneous distributions $F_1, F_2, ..., F_n$. Several studies reveal the use of it in the characterization problems, linear estimation, detection of outliers, study of system reliability, survival analysis, life testing, data compression, among others (Arnold and Balakrishnan, 1989; Balakrishnan and Cohen, 1991). Lieblein (1953) derived a formula for variance and covariance of order statistics in sample from the extreme-value distribution (asymptotic distribution of the largest value) in terms of a certain tabulated functions. Balakrishnan and his co-author (1986) have looked

[1]Dept. of Statistics, B.H.U. Varanasi, India. E-mail: raipiyush5@gmail.com.
[2]Dept. of Mathematics and Statistics, Banasthali Vidyapith, Rajasthan, India. E-mail: bsc.ashori@gmail.com.

into the order statistics from various types of generalized logistic population. The theory of order statistics is available in the excellent book by David and Nagaraja (2003).

The order statistics from the Weibull distribution is considered in the present study. The Weibull distribution is widely used in reliability and survival analysis due to its versatility. This distribution is used to model the variety of life behaviours depending on the values of the parameters. This paper is not concerned with the properties of this distribution, a detailed account on the estimation of parameter and applications of this distribution can be seen from many studies and available researches, e.g. the books written by Prabhakar Murthy et al. (2004) and Rinne (2008). Let us consider the random samples of size n from the Weibull distribution with probability density function (pdf),

$$f(x) = \gamma x^{(\gamma-1)} e^{-x^{\gamma}}, \qquad x \geq 0, \qquad \gamma > 0 \tag{1}$$

The corresponding cumulative distribution function (cdf) and survival function (sf) are $F(x) = 1 - e^{-x^{\gamma}}$ and $S(x) = e^{-x^{\gamma}}$ respectively. This distribution has been applied extensively, mainly by Weibull (Weibull, 1939a; Weibull, 1939b; Weibull, 1951; Weibull, 1952) and will be referred to by his name. Lieblein (1955), in his later paper, derived first two moments of order statistics in terms of incomplete beta and gamma function for the Weibull distribution. Sometimes the derivation is not simple and the obtained expressions are not in a very analytical form. We derive higher order moments of order statistics in the sample from one-parameter Weibull distribution using an alternative expectation formula and also their joint and characteristic functions.

2. Moments of order statistics

Let n values from one-parameter Weibull distribution after ordering in size are denoted by

$$x_1, x_2, ..., x_n, \qquad x_1 \leq x_2 \leq ... \leq x_n$$

where, $x_r, r = 1, 2, ..., n$ is called the rth order statistics and we seek the moments of these order statistics. The probability density function of the rth order statistics is given by

$$p(x) = \frac{n!}{(r-1)!(n-r)!} [F(x)]^{r-1} [1 - F(x)]^{n-r} f(x), \qquad -\infty < x < \infty \tag{2}$$

where, $x = x_r$. Using (1)

$$p(x) = \frac{n!}{(r-1)!(n-r)!} \gamma x^{\gamma-1} (1 - e^{-x^{\gamma}})^{(r-1)} e^{-x^{\gamma}(n-r+1)}, \qquad x \geq 0 \tag{3}$$

Now, from (3) survival function of rth order statistics is

$$S(x) = \frac{n!}{(r-1)!(n-r)!} \sum_{u=0}^{r-1} (-1)^{u\ r-1} C_u \, e^{-x^{\gamma}(n-r+u+1)} / (n-r+u+1) \tag{4}$$

The joint pdf of rth and sth order statistics x_r, x_s, is

$$p(x,y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}[F(x)]^{r-1}[F(y)-F(x)]^{s-r-1}[1-F(y)]^{n-s} \\ f(x)f(y), \qquad -\infty < x \le y < \infty \tag{5}$$

where $x = x_r, y = x_s, r < s, r, s = 1,2,...,n$. From d.f. (1) we obtain

$$p(x,y) = \frac{n!\ \gamma^2\ x^{\gamma-1}y^{\gamma-1}}{(r-1)!(s-r-1)!(n-s)!}(1-e^{-x^\gamma})^{r-1}(e^{-x^\gamma}-e^{-y^\gamma})^{s-r-1} \\ e^{-x^\gamma}\ e^{-y^\gamma(n-s+1)} \tag{6}$$

Now, joint survival function of rth and sth order statistics is

$$S(x,y) = P(X \ge x, Y \ge y)$$

$$S(x,y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!}\sum_{u=0}^{r-1}\sum_{v=0}^{s-r-1}(-1)^{u+v}\ {}^{r-1}C_u\ {}^{s-r-1}C_v \\ e^{-x^\gamma(s-r-v+u)}\ e^{-y^\gamma(n-s+v+1)}/[(s-r-v+u)(n-s+v+1)] \tag{7}$$

The joint pdf and survival function of $x_{r_1}, x_{r_2}, ..., x_{r_k}$ order statistics are given below in (8) and (9) respectively.

$$p(x_{r_1}, x_{r_2}, ..., x_{r_k}) = \frac{n!\ \gamma^k\ (x_{r_1}^{\gamma-1}\ x_{r_2}^{\gamma-1}...\ x_{r_k}^{\gamma-1})}{(r_1-1)!(r_2-r_1-1)!...(n-r_k)!}(1-e^{-x_{r_1}^\gamma})^{r_1-1} \\ (e^{-x_{r_1}^\gamma}-e^{-x_{r_2}^\gamma})^{r_2-r_1-1}(e^{-x_{r_2}^\gamma}-e^{-x_{r_3}^\gamma})^{r_3-r_2-1}... \\ e^{-x_{r_k}^{\gamma(n-r_k)}}\ e^{-(x_{r_1}^\gamma+x_{r_2}^\gamma+...+x_{r_k}^\gamma)} \tag{8}$$

$$S(x_{r_1}, x_{r_2}, ..., x_{r_k}) = P(X_1 \ge x_{r_1}, ..., X_k \ge x_{r_k})$$

So,

$$S(x_{r_1}, x_{r_2}, ..., x_{r_k}) = \frac{n!}{(r_1-1)!(r_2-r_1-1)!...(n-r_k)!}\sum_{u_1=0}^{r_1-1}\sum_{u_2=0}^{r_2-r_1-1}...\sum_{u_{k-1}=0}^{r_{k-1}-r_{k-2}-1} \\ (-1)^{u_1+u_2+...+u_{k-1}}\ {}^{r_1-1}C_{u_1}\ {}^{r_2-r_1-1}C_{u_2}...\ {}^{r_{k-1}-r_{k-2}-1}C_{u_{k-1}} \\ \frac{e^{-x_{r_1}^\gamma(r_2-r_1-u_2+u_1)}...e^{-x_{r_{k-2}}^\gamma(r_{k-1}-r_{k-2}-u_{k-1}+u_{k-2})}e^{-x_{r_{k-1}}^\gamma(u_{k-1}+1)}e^{-x_{r_k}^\gamma(n-r_k+1)}}{(r_2-r_1-u_2+u_1)...(r_{k-1}-r_{k-2}-u_{k-1}+u_{k-2})(u_{k-1}+1)(n-r_k+1)} \tag{9}$$

where $r_1, r_2, ..., r_k = 1,2,...,n$, such that $r_1 < r_2 < ... < r_k$. This note utilizes survival function to find the moments of order statistics from one-parameter Weibull distribution. Recently, this technique of finding moments has gained more importance in practical situation and Hong called it an alternative expectation formula (Feller, 1966; Nadarajah and Mitov, 2003; Hong, 2012). For a continuous non-negative random variable X, the mean or the first

moment can be expressed as

$$E(X) = \int_0^\infty (1 - F(x))dx = \int_0^\infty S(x)dx \tag{10}$$

Hong (2012) derived this formula to find the mean. Later he established the formula for the expected value of the joint variables (Hong, 2015). If $X_1, X_2, ..., X_n$ are non-negative random variables and $S(x_1, x_2, ..., x_n)$ is their joint survival function then expected value of joint variables (10) is

$$E(X_1, X_2, ..., X_n) = \int_0^\infty ... \int_0^\infty S(x_1, x_2, ..., x_n)dx_1dx_2...dx_n \tag{11}$$

Chakraborti et al. (2017) described higher order moments using the same technique. The mth moments about the origin, $E(X^m)$, of a continuous random variable X is

$$E(X^m) = \int_0^\infty mx^{m-1}S(x)dx \tag{12}$$

From the survival function (4) of rth order statistics we obtain

$$E(X_r^m) = \frac{n!}{(r-1)!(n-r)!} \sum_{u=0}^{r-1} \frac{(-1)^u \ {}^{r-1}C_u}{(n-r+u+1)} \int_0^\infty mx^{m-1}e^{-x^\gamma(n-r+u+1)}dx \tag{13}$$

Making the transformation $x^\gamma = z$ in (13) and performing some algebraic simplification we obtain

$$E(X_r^m) = \frac{n!}{(r-1)!(n-r)!} \sum_{u=0}^{r-1} (-1)^u \ {}^{r-1}C_u \frac{(m/\gamma)\Gamma(m/\gamma)}{(n-r+u+1)^{m/\gamma+1}} \tag{14}$$

From (14) we can find moments about origin of the order statistics from the Weibull distribution. Combining (11) and (7) we obtain

$$E(X,Y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \sum_{u=0}^{r-1} \sum_{v=0}^{s-r-1} (-1)^{u+v} \frac{{}^{r-1}C_u \ {}^{s-r-1}C_v}{(n-s+v+1)(s-r-v+u)} \int_0^\infty \int_0^y e^{-x^\gamma(s-r-v+u)} \ e^{-y^\gamma(n-s+v+1)}dxdy \tag{15}$$

Making the transformation $x^\gamma = z$ and performing some algebra we get

$$E(X,Y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \sum_{u=0}^{r-1} \sum_{v=0}^{s-r-1} (-1)^{u+v} \frac{{}^{r-1}C_u \ {}^{s-r-1}C_v}{\gamma(n-s+v+1)(s-r-v+u)} \int_0^\infty G(1/\gamma, y^\gamma) \ e^{-y^\gamma(n-s+v+1)}dy \tag{16}$$

where,

$$G(1/\gamma, y^{\gamma}) = \int_0^{y^{\gamma}} e^{-z^{(s-r-v+u)}} z^{(1/\gamma)-1} dz \tag{17}$$

On multiplying $e^{y^{\gamma}(s-r-v+u)}$ both sides and making transformation $z = y^{\gamma}t$, and performing some algebra in (17) we get

$$G(1/\gamma, y^{\gamma}) = \sum_{w=0}^{\infty} \frac{(s-r-v+u)^w}{w!} B(1/\gamma, w+1)(y^{\gamma})^{(1/\gamma)+w} e^{-y^{\gamma}(s-r-v+u)} \tag{18}$$

where $B(1/\gamma, w+1)$ is the beta function. Combining (16) and (18) and making another transformation $y^{\gamma} = p$ and performing some algebraic simplification we get

$$E(X,Y) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \sum_{u=0}^{r-1} \sum_{v=0}^{s-r-1} (-1)^{u+v} \frac{{}^{r-1}C_u \; {}^{s-r-1}C_v}{\gamma^2 (n-s+v+1)(s-r-v+u)} \sum_{w=0}^{\infty} \frac{(s-r-v+u)^w}{w!(n-r+u+1)^{(2/\gamma+w)}} B(1/\gamma, w+1)\Gamma((2/\gamma)+w) \tag{19}$$

In the same manner the expected value of the product of order statistics $x_{r_1}, x_{r_2}, ..., x_{r_k}$ from the Weibull distribution is

$$E(X_{r_1}, X_{r_2}, ..., X_{r_k}) = \int_0^{\infty} \int_0^{x_{r_k}} \int_0^{x_{r_{k-1}}} ... \int_0^{x_{r_2}} S(x_{r_1}, x_{r_2}, ..., x_{r_k}) \, dx_{r_1} dx_{r_2} ... dx_{r_k} \tag{20}$$

After simplifications,

$$\begin{aligned} E(X_{r_1}, X_{r_2}, ..., X_{r_k}) = {} & \frac{n!}{\gamma^k (r_1-1)!(r_2-r_1-1)!...(n-r_k)!} \sum_{u_1=0}^{r_1-1} \sum_{u_2=0}^{r_2-r_1-1} ... \sum_{u_{k-1}=0}^{r_{k-1}-r_{k-2}-1} \\ & (-1)^{u_1+u_2+...+u_{k-1}} \; {}^{r_1-1}C_{u_1} \; {}^{r_2-r_1-1}C_{u_2} ... \; {}^{r_{k-1}-r_{k-2}-1}C_{u_{k-1}} \\ & \left(\sum_{w_1=0}^{\infty} \frac{(r_2-r_1-u_2+u_1)^{w_1} B(1/\gamma, w_1+1)}{w_1!(r_2-r_1-u_2+u_1)} \right) ... \\ & \left(\sum_{w_{k-2}=0}^{\infty} \frac{(r_{k-2}-r_1-u_{k-2}+u_1)^{w_{k-2}} B((k-2/\gamma)+w_1+...+w_{k-3}, w_{k-2}+1)}{w_{k-2}!(r_{k-2}-r_1-u_{k-2}+u_1)} \right) \\ & \left(\sum_{w_{k-1}=0}^{\infty} \frac{(r_{k-1}-r_1-u_{k-1}+u_1)^{w_{k-1}} B((k-1/\gamma)+w_1+...+w_{k-2}, w_{k-1}+1)}{w_{k-1}!(u_{k-1}+1)} \right) \\ & \frac{\Gamma((k/\gamma)+w_1+...+w_{k-1})}{(n-r_k+1)(n-r_k+r_{k-1}-r_1+u_1+2)^{(k/\gamma)+w_1+...+w_{k-1}}} \end{aligned} \tag{21}$$

3. characteristic function of order statistics

The characteristic function of rth order statistics x_r is given as

$$\phi_r(t) = E(e^{itx}) \tag{22}$$

where, $x = x_r$, from (2) we obtain

$$\phi_r(t) = \frac{n!}{(r-1)!(n-r)!} \int_0^\infty e^{itx} \gamma x^{\gamma-1} (1-e^{-x^\gamma})^{(r-1)} e^{-x^\gamma(n-r+1)}$$

$$= \frac{n!}{(r-1)!(n-r)!} \sum_{u=0}^{r-1} (-1)^{u\,r_1-1} C_{u_1} \int_0^\infty \gamma x^{\gamma-1} e^{-x^\gamma u} e^{itx} e^{-x^\gamma(n-r+1)} \tag{23}$$

On expanding e^{itx} in terms of power series, making transformation $x^\gamma = z$ and performing some algebra in (20) we get

$$\phi_r(t) = \frac{n!}{(r-1)!(n-r)!} \sum_{u=0}^{r-1} (-1)^{u\,r_1-1} C_{u_1} \sum_{p=0}^{\infty} \frac{(it)^p}{p!} \frac{p/\gamma \Gamma(p/\gamma)}{(n-r+u+1)^{p/\gamma+1}} \tag{24}$$

4. Illustration

The above formulae are used to find the first two moments for a small sample of one-parameter Weibull distribution by means of order statistics. The following computations for $n = 3$ and $\gamma = 2.5$ illustrate the first and second moments corresponding to $r = 1, r = 2$ and $r = 3$,

$$E(X_1) = 0.572,\ E(X_2) = 0.874$$

$$E(X_3) = 1.216$$

and

$$E(X_1^2) = 0.387,\ E(X_2^2) = 0.831$$

$$E(X_3^2) = 1.575$$

Variance corresponding to these statistics for $r = 1, r = 2$ and $r = 3$ is,

$$V(X_1) = 0.572,\ V(X_2) = 0.874$$

$$V(X_3) = 1.216$$

The expected value of joint of order statistics for $r = 1$ and $s = 2$ is,

$$E(X,Y) = 0.728$$

In the same manner we can find out the higher order moments about origin and mean on the basis of the procedure described in this article.

5. Conclusion

The Weibull distribution is used to describe various observed failures of phenomena and widely used in survival analysis and reliability theory. Techniques introduced here to evaluate the moments of order statistics from one-parameter Weibull distribution involve survival function and will be highly useful for practical situations. Characteristic function has great ttheoretical importance and it has central importance in statistics and the expression is derived here for the order statistics from Weibull distribution. Moreover it is able to generate higher order moments of order statistics and can be used to evaluate the expression for mean, variance, covariance, skewness and kurtosis of the distribution.

REFERENCES

ARNOLD, B. C., BALAKRISHNAN, N., (1989). Relations, bounds and applications for order statistics, Springer-Verlag, lecture Notes in Statistics, No. 53.

BALAKRISHNAN, N., COHEN, A. C., (1991). Order statistics and inference: Estimation methods, Academic Press.

BALAKRISHNAN, N., KOCHERLAKOTA, (1986). On the moments of order statistics from the doubly truncated logistic distribution, Journal of Statistical Planning and Inference, 13, pp.117–129.

CHAKRABORTI, S., JARDIM, F. AND EPPRECHT, E., (2017). Higher order moments using the survival function: The alternative expectation formula, The American Statistician, pp. 1–12.

DAVID H.A., NAGARAJA H. N., (2003). Order Statistics, Third Edition, John Wiley, New York.

FELLER, W., (1966). An introduction to probability theory and its application, New York: Wiley.

HONG, L., (2012). A remark non the alternative expectation formula, The American Statistician, 66, pp. 232–233.

HONG, L., (2015). Another remark on the alternative expectation formula, The American Statistician, 69(3), pp. 232–233.

LIEBLEIN, J., (1953). On the exact evaluation of the variances and covariances of order statistics in samples from the extreme-value distribution, Annals of Mathematical Statistics, 24, pp. 282–287.

LIEBLEIN, J., (1955). On moments of order statistics from Weibull distribution. Annals of Mathematical Statistics, 24, pp. 330–333.

NADARAJAH, S., MITOV, K. V., (2003). Product moments of multivariate random vectors, Communication in Statistics: Theory and Methods, 32(1), pp. 47–60.

PRABHAKAR MURTHY, D. N., XIE, M., JLANG, R., (2004). The Weibull Model. Weily Series in Probability and Statistics.

RINNE, H., (2008). The Weibull distribution, CRC Press.

WEIBULL, W., (1939a). A statistical theory of the strength and materials, Ing. Vetenskaps Akad. Handl., 151, p. 15.

WEIBULL, W., (1939b). The Phenomenon of rupture in solids. Ing. Vetenskaps Akad. Handl., 153, p. 17.

WEIBULL, W., (1951). A statistical distribution of wide applicability. J. Appl. Mech., 18, pp. 293–297.

WEIBULL, W., (1952). Statistical design of fatigue experiments. J. Appl. Mech., 19, pp. 109–113.

STATISTICS IN TRANSITION new series, March 2020
Vol. 21, No. 1 pp. 177–178,

About the Authors

Al-Jararha J. is an Associate Professor at the Department of Statistics, Faculty of Science, Yarmouk University, Irbid, Jordan. Graduated from Colorado State University; Fort Collins, Colorado, USA. The main areas of interest are survey sampling and linear models. His PhD dissertation has been published as a book.

Awe Olushina Olawale is an Assistant Professor of Applied Statistics in the Department of Mathematical Sciences, Anchor University Lagos. His research interests include multivariate statistical analysis, statistical computing, Biostatistics, predictive modelling, analysis of multivariate time series data and Bayesian econometrics. Professor Awe has published over 40 research papers in international/national journals and conferences. He is an active member of many scientific and professional bodies. He has been a visiting scholar and statistical collaborator at the Department of Statistics, Virginia Tech, USA. A fellow of the African Scientific Institute, USA, and an affiliate member of the African Academy of Sciences, Nairobi Kenya, he is presently the LISA 2020 Ambassador to Africa and a visiting research scholar/data scientist at the Institute of Mathematics and Statistics, Federal University of Bahia (UFBA), Salvador, Brazil.

Błażej Mirosław currently holds the position of Director of the Macroeconomic Studies and Finance Department, Statistics Poland. His main area of interest includes: macroeconomic analysis, financial market, business cycle and quantitative methods in statistics and economy. Between 2006 and 2010 he acted as Alternate Member of the EU Economic Policy Committee, representing the Ministry of Finance. He is an author and co-author of publications in economics and physics.

Eideh Abdulhakeem is an Associate Professor of statistical theory for official and survey statistics and survey Methodology at the Department of Mathematics, Al-Quds University, Palestine. His main research areas are analytic inference from complex surveys under informative sampling and nonignorable nonresponse, with applications in small area estimation, time series models and longitudinal survey data. He has delivered around 10 short courses for statisticians in official statistics in the Mediterranean Countries through Eurostat. He has published around 20 articles in international statistical journals and several papers in international conferences. He served as a consultant for ONS, UNESCWA, and UNICEF. Dr. Eideh is a member of the ISI, IASS, and IAOS. In the year 2015 he received the best paper in the field of sampling awarded by the Indian Society of Agricultural Statistics, New Delhi.

Dehnel Grażyna is an Associate Professor at the Department of Statistics, Poznań University of Economics and Business. Her main research domain is small area estimation, survey sampling, short-term and structural business statistics. She is also interested in outlier robust regression applied on business data, business demography and data integration.

Kotlewski Dariusz is an Assistant Professor at the Collegium of Business Administration, Warsaw School of Economics (SGH). Simultaneously he holds the position of Consultant in the Department of Macroeconomic Studies and Finance, Statistics Poland. His main area of interest include: macroeconomics, particularly the economic growth theory, including productivity and growth accounting and growth decompositions of different kinds; international economics, particularly the international trade theory, including new economic geography; analyses of spatial distribution of economic growth; electric power economics. He is the author or co-author of over 20 publications – research papers in international/national journals and books.

Rossa Agnieszka is an Associate Professor at the Department of Demography, Faculty of Economics and Sociology, University of Lodz. Her main areas of interest include demographic forecasting, event history analysis and multivariate statistical analysis. Currently, she is a member of the Committee on Demographic Studies of The Polish Academy of Science.

Tailor Rajesh is a Reader in School of Studies in Statistics, Vikram University, Ujjain (M.P.), India. He has served National Council of Educational Research and Training (NCERT), New Delhi as a lecturer. He has published 83 research papers in national and international journals. He has presented research papers in over 30 conferences and delivered invited talks in over 20 workshops and training programs. Nine students have received their PhD degrees in his supervision.

Yadav Rohini received her PhD in Sample Surveys from IIT-ISM Dhanbad, India in 2012. Currently, she is working as an Assistant Professor in the Department of Statistics, Amity University, Noida, India. She has published 15 research papers in international journals of repute. She has presented several research papers in national and international conferences. She is a referee of several journals.

Wawrowski Łukasz is an Assistant Professor at the Department of Statistics, Poznań University of Economics and Business. His main areas of interest include poverty measurement, small area estimation, data visualization and multivariate data analysis.

Zaman Tolga is an Assistant Professor at the Department of Statistics in Cankiri Karatekin University, Cankiri, Turkey. His research interests are sampling theory, resampling methods, robust statistics, and statistical inference. He has published over 50 research papers in international/national journals and conferences.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, Key words (in bold) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).