STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Call for Papers:

Sampling methods and estimation

Research article

Other articles:

Multivariate Statistical Analysis 2013, Łódź. Conference Papers

Conference Report

Volume 15, Number 2, Spring 2014

Editorial Office **ISSN 1234-7655**

Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Beata Witek: Secretary
b.witek@stat.gov.pl. Phone number 00 48 22 — 608 33 66
Rajmund Litkowiec: Technical Assistant

Address for correspondence

GUS, al. Niepodległości 208, 00-925 Warsaw, POLAND, Tel./fax:00 48 22 — 825 03 95

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 175–178

FROM THE EDITOR

At the outset of this issue, I would like to turn attention of the reader to the important change in the composition of the Editorial Board of the *Statistics in Transition new series*. First, let me take this opportunity to thank those who are stepping down after serving on the Editorial Board for the past seven years for all their help, inspiration and encouragement which I personally and other members of the Editorial Office have been obtaining from them over that period. And to express my, as well as of **Prof. Janusz Witkowski,** the President of the Central Statistical Office of Poland and of **Prof. Czeslaw Domanski**, the President of the Polish Statistical Association, gratefulness for their hitherto contributions and for readiness to continue to collaborate with us in a slightly different function, as members of the Associate Editors, to: **Prof. Prof. Walenty Ostasiewicz, Tomasz Panek, Jan Paradysz, Miroslaw Szreder,** and **Mr. Wladyslaw W. Lagodzinski.**

At the same time, I would like to welcome new members of the Board who were invited to serve in this function, starting with this issue. Given a vital role the body plays for the overall quality of the journal, and for its image and prestige, I would like to express - also on behalf of **Prof. Janusz Witkowski** and of **Prof. Czeslaw Domanski**, who are jointly co-chairing the new Board - our sincere appreciation for accepting this invitation to Prof. Prof.: **Sir Anthony B. Atkinson, Graham Kalton, Malay Ghosh, Miroslaw Krzysko,** and **Janusz L. Wywial**. We all look forward to keeping the journal on the ambitious track of its growing significance and usability for the community of statisticians, producers and users of statistics, worldwide.

Another thing the importance of which I would like to stress here is the announced earlier intention to prepare one of the future issues of the journal - actually, the first issue of the next year (i.e. Winter 2015) - as a thematic collection of papers devoted to subjective well-being as an object of survey research in both national and international contexts. With the invaluable aid of **Graham Kalton**, who has kindly agreed to act as a Guest (co)Editor of the planned issue, and helps us with the challenging task of arranging for such a collection of papers, we hope to be able to complete the needed input by the end of this year - see the ***Call for Papers*** below.

Along the line of our efforts to have the *Statistics in Transition new series* covered by monitoring systems of international indexation bases, we are pleased to inform our partners and collaborators that (in addition to the systems which are already monitoring our journal) currently the *SiTns* is under consideration for being included in the system of Central and Eastern European Online Library (CEEOL).

As regards the contents of this issue, there are three articles devoted to *sampling methods and estimation,* one *research* paper and six articles based on papers presented at the conference *Multivariate Statistical Analysis* held in Lodz last year 2013. They are briefly characterized below.

Adulhakeem A. H. Eideh's paper ***On the Use of Sampling Weights and Sample Distribution When Estimating Regression Models Under Informative Sampling*** shows that the use of sampling weights when estimating regression models with survey data and the use of sample distribution in fitting regression models with survey data proposed in the literature are coincide methods, dealing with essentially the same statistical problem. Discussion of these two methods leads to conclusion that only difference between them lays in estimating the informativeness parameter (γ_2). Author hopes that his investigation will contribute to further theoretical and empirical research in these areas.

Kumari Priyanka and **Richa Mittal** discuss the problem of estimation of population median at current occasion in two-occasion successive sampling in ***Effective Rotation Patterns for Median Estimation in Successive Sampling***. They propose best linear unbiased estimators by utilizing additional auxiliary information, readily available on both the occasions. Asymptotic variances of the proposed estimators are derived and the optimum replacement policies are discussed. The behaviours of the proposed estimators are analyzed on the basis of data from natural populations. Simulation studies have been carried out to measure the precision of the proposed estimators. Authors believe that the proposed estimators may be useful for survey practitioners.

G. N. Singh and **D. Majhi** also use the information on two-auxiliary variables in order to propose in the paper on ***Some Chain-type Exponential Estimators of Population Mean in Two-Phase Sampling*** three different exponential chain-type estimators of population mean of study variable in two-phase (double) sampling. Properties of the proposed estimators have been studied and their performances are examined with respect to several well known chain-type estimators. Empirical studies are carried out to support the theoretical results. The proposed estimators show to be preferable over alternative estimators for the population which satisfies some conditions (derived in the text); therefore, they may be recommended for their practical applications.

Tomasz Gorecki, Miroslaw Krzysko, Lukasz Waszak, Waldemar Wolynski discuss some ***Methods of Reducing Dimension for Functional Data.*** They start with classical data analysis with objects being characterized by many features observed at one point of time and typically presented graphically in order to see their configuration, eliminate outlying observations, observe relationships between them, or to classify them. Authors propose a new method of constructing principal components for multivariate functional data, and illustrate its application for data from environmental studies. Their research has shown that the use of a multivariate functional principal components analysis leads to desired results, though the performance of the algorithm needs to be further evaluated on both real and artificial data sets.

The series of articles based on aforementioned conference papers is opened by **Lukasz Feldman's, Radoslaw Pietrzyk's** and **Pawel Rokita's** article ***Multiobjective Optimization of Financing Household Goals with Multiple Investment Programs.*** They propose a technique of facilitating life-long financial planning for a household by finding the optimal match between systematic investment products and multiple financial goals of different realization terms and magnitudes. As this is a multi-criteria optimization, they consider several objectives, such as (i) compliance between the expected term structure of cumulated net cash flow throughout the life cycle of the household with its life-length risk aversion and bequest motive; (ii) financial liquidity in all periods under expected values of all stochastic factors; (iii) minimization of net cash flow volatility; and (iv) minimization of costs of the investment plan combination. The result is a set of systematic-investment programs with accompanying information which programs are destined to cover particular financial goal. As a result, an optimization procedure is proposed based on an original goal function (adjusted to the proposed household financial plan model).

Alina Jedrzejczak employs the Gini index decomposition procedures to analyze ***Income Inequality and Income Stratification in Poland***. Starting with an overview of several methods of decomposing Gini, selected approaches to the analysis of income distribution were used to show the extent to which the inequality in different subpopulations contributes to the overall income inequality in Poland. And to what extent members of the subpopulation groups (of households) form distinct segments or strata. Particular use was made of the Dagum procedure of Gini decomposition since it is based on the concept of economic distance between distributions and relative economic affluence and accounts for different variances and asymmetries of income distributions in subpopulations, and gives an important contribution to the understanding of the overlapping term. Also decomposition proposed by Yitzhaki and Lerman is discussed as it encompasses the stratification problem due to linking social stratification with inequality. The households were divided by economic regions using the Eurostat classification units NUTS 1 as well as by family type defined by the number of children.

Grazyna Trzpiot's paper ***Application of Coherent Distortion Risk Measures*** is devoted to solving the problem of portfolio selection. It presents an extension of the well-known optimization framework for Conditional Value-at-Risk (CVaR)-based portfolio selection problems to optimization over a more general class of risk measure known as the class of Coherent Distortion Risk Measure (CDRM). CDRM class of risk measures is the intersection of Coherent Risk Measure (CRM) and Distortion Risk Measure (DRM). CDRM includes many well-known risk measures. In conclusion, the use of the discussed procedure to the development of a CDRM-based portfolio optimization framework is being offered.

In the next paper, ***Selected Tests Comparing the Accuracy of Inflation Rate Forecasts Constructed by Different Methods*** by **Agnieszka Przybylska-Mazur,** the problem of forecasts of macroeconomic variables - including the forecasts of inflation rate - is discussed in the context of projection of future situation in the economy. Knowledge of effective forecasts allows making optimal business, financial and investment decisions. Author applies selected tests to the evaluation of the accuracy of inflation rate forecasts determined by different methods. A general conclusion after employing different procedures to the problem of projection states that the differences in values result from the change in the assumptions about the projections in the different reports.

Malgorzata Markowska, Marek Sobolewski, Andrzej Sokolowski, Danuta Strahl present ***Tests for Connection Between Clustering of Polish Counties and Province Structure*** based on Sokolowski et. al. idea of statistical tests which allow to check the influence of geographical or administrative units of upper level onto clustering results of lower level units. They use so called “active border” notion for the borders between counties and also between provinces. The number and length of active boarders are used in the proposed test statistics, the distribution of which depends on the actual geographic division of a given country. Table for test critical values and the approximation functions are provided. According to the authors, the proposed test can be useful in testing the relations between administrative levels in Poland with respect to economic as well as to public administration and quality of life phenomena.

Bronislaw Ceranka's and **Malgorzata Graczyk's** paper ***On Certain A-Optimal Biased Spring Balance Weighing Designs*** is focused on the estimation of unknown measurements of p objects in the experiment conducted in accordance with the model of the spring balance weighing design. The weighing design is called biased if the first column of the design matrix has elements equal to one only. The A-optimal design is a design in which the trace of the inverse of information matrix is minimal. The main result is the broadening of the class of experimental designs so that we are able to determine the regular A-optimal design. Authors provide the lowest bound of the covariance matrix of errors and they give new construction methods of the regular A-optimal spring balance weighing design based on the incidence matrices of the balanced incomplete block designs. An example illustrates the procedure at work.

The issue is concluded with information on the conference on ***Coherence Policy and the Development of Cross-border Areas Along the European Union’s External Border*** (27-28 June 2014, Krasiczyn-Arlamow, Poland)

Wlodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 179

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

> Manuscript should be submitted electronically to the Editor:
> sit@stat.gov.pl., followed by a hard copy addressed to
> Prof. Wlodzimierz Okrasa,
> GUS / Central Statistical Office
> Al. Niepodległości 208, R. 287, 00-925 Warsaw, Poland

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

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://pts.stat.gov.pl/en/journals/statistics-in-transition/

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 181

CALL FOR PAPERS:
THE MEASUREMENT OF SUBJECTIVE WELL-BEING IN SURVEY RESEARCH

Subjective well-being has been a subject of research for many years, but interest in the subject has grown markedly in the past decade, including among official statisticians. The early efforts to measure subjective well-being in surveys were made by behavioural (psychometrics), economic (econometrics) and social scientists in academic settings. The subject is now recognized as having important policy applications, thus leading to the demand for the measurement of subjective well-being in both national and international official surveys. Significant problems related to survey context exist in developing effective measures of subjective well-being, and these problems are particularly acute when the measures are to be used for international or cross-cultural comparisons. In recognition of the importance of measuring subjective well-being in surveys and the challenges this presents, *Statistics in Transition* plans to publish a collection of papers on statistically relevant aspects of research on subjective well-being in its first issue next year (Vol. 16, No. 1). The aim is to provide an insightful overview of the theoretical, methodological, and practical issues involved.

Researchers are invited to submit papers on any aspect of the subject including, but not limited to, the following:

1. The conceptualization of subjective well-being and its multidimensional nature.
2. The operationalization and the measurement of subjective well-being, including the trade-offs involved (e.g., the number of items used in its measurement *vs*. the added response burden, the concern about validity and other methodological matters *vs.* a tendency to build upon existing methods and surveys).
3. The data collection methods used (e.g., if and how respondents' reports of their well-being are mode-dependent).
4. The uses made of the well-being measures and the analytic framework employed.
5. The interpretation of results, particularly for international comparisons.
6. Experiences in the use of subjective well-being measures in surveys.

In order to allow time for each submitted paper to go through a double-blind peer review process, papers should be submitted by December 30, 2014. For technicalities concerning editorial requirements and the submission procedure, please consult our 'Guidelines' link:

http://pts.stat.gov.pl/en/journals/statistics-in-transition/

Graham Kalton, Guest Editor
Christopher Mackie, Guest Editor
Wlodzimierz Okrasa, Editor

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 183–196

ON THE USE OF SAMPLING WEIGHTS AND SAMPLE DISTRIBUTION WHEN ESTIMATING REGRESSION MODELS UNDER INFORMATIVE SAMPLING

Adulhakeem A. H. Eideh[1]

ABSTRACT

In this paper we show that the use of sampling weights when estimating regression models with survey data discussed by Magee, Robb and Burbidge (1998), and the use of sample distribution in fitting regression models with survey data proposed by Pfeffermann and Sverchkov (1999) are coincide methods dealing with the same statistical problem.

Key words: sample likelihood, first order inclusion probability, two-step maximum likelihood method.

1. Introduction

Some recent work has considered the definition of the sample distribution under informative sampling. When the sample selection probabilities depend on the values of the model response variable, even after conditioning on auxiliary variables, the sampling mechanism becomes informative and the selection effects need to be accounted for in the inference process. Pfeffermann, Krieger and Rinott (1998) propose a general method of inference on the population distribution (model) under informative sampling that consists of approximating the parametric distribution of the sample measurements. The sample distribution is defined as the distribution of the sample measurements given the selected sample. Under informative sampling, this distribution is different from the corresponding population distribution, although for several examples the two distributions are shown to be in the same family and only differ in some or all the parameters. The authors discuss and illustrate a general approach of approximating the marginal sample distribution for a given population distributions and first order sample selection probabilities. For more discussion on analysis of complex survey data, see Chambers and Skinner (2003), Skinner, Holt, and Smith (1989), Skinner (1994), Magee, Robb and Burbidge (1998),

[1] Department of Mathematics, College of Science and Technology, Al-Quds University, Abu-Dies campus, Palestine, P.O. Box 20002, Jerusalem. E-mail: msabdul@science.alquds.edu.

Eideh (2003, 2007, 2008, 2009, 2010, 2011, 2012a, 2012b), Eideh and Nathan (2006, 2009), Pfeffermann, Krieger and Rinott (1998), Pfeffermann and Sverchkov (1999, 2003), and Sverchkov and Pfeffermann (2004).

In this paper we will show that the use of sampling weights when estimating regression models with survey data discussed by Magee, Robb and Burbidge (1998), and the use of sample distribution in estimating regression models with survey data discussed by Pfeffermann, Krieger and Rinott (1998) and Pfeffermann and Sverchkov (1999) are coincide methods dealing with same statistical problem.

The plan of this paper is as follows. In Section 2 we consider probability weighting. In Section 3 we discuss pseudo-likelihood estimation. Section 4 deals with the use of sampling weights when estimating regression models with survey data. Section 5 i ntroduces the use of sample distribution when estimating regression models with survey data. We conclude with a brief discussion in Section 6.

2. Probability weighting

Let $U = \{1,...,N\}$ denote a finite population consisting of N units. Let y be the target or study variable of interest and let y_i be the value of y for the ith population unit. At this stage the values y_i are assumed to be fixed unknown quantities. Suppose that an estimate is needed for the population total of y, $T = \sum_{i \in U} y_i$. A probability sample s is drawn from U according to a specified sampling design. The sample size is denoted by n . The sampling design induces inclusion probabilities for the different units of U . Let $\pi_i = \Pr(i \in s)$ be the first order inclusion probability of the ith population unit. The Horvitz-Thompson estimator or probability-weighted (PW) estimator of the population total of y, $T = \sum_{i \in U} y_i$ is given by:

$$\hat{T} = \sum_{i \in s} w_i y_i$$

where $w_i = 1/\pi_i$ is the sampling weight of unit $i \in U$, that is we weigh each sample observation i by the sampling weight, w_i. This estimator is design-unbiased, that is $E_D\left(\sum_{i \in s} w_i y_i\right) = T$, where E_D denotes the expectation under repeated sampling. For more discussion on pr obability weighting, see Sarndal, Swensson, and Wretman (1992).

3. Pseudo-likelihood estimation

We now consider the population values $y_1,...,y_N$ as random variables, which are independent realizations from a distribution with probability density function (pdf) $f_p(y_i \mid \theta)$, indexed by a vector of parameters θ. We n ow consider the estimation of the superpopulation parameter, θ, rather than the prediction of the (random variable) total T. Let

$$l(\theta \mid y_1,...,y_N) = \sum_{i=1}^{N} \log f_p(y_i \mid \theta)$$

be the census log-likelihood. The census maximum likelihood estimator of θ solves the population likelihood equations:

$$U(\theta) = \sum_{i=1}^{N} \frac{\partial\{\log f_p(y_i \mid \theta)\}}{\partial\theta} = 0$$

Following Binder (1983), the pseudo-maximum likelihood (PML) estimator is the solution of: $\hat{U}(\theta) = 0$, where $\hat{U}(\theta)$ is a sam ple estimator of the function $U(\theta)$. For example, the probability-weighted estimator of $U(\theta)$ is such an estimator:

$$\hat{U}_w(\theta) = \sum\nolimits_{i\in s} w_i \frac{\partial\{\log f_p(y_i \mid \theta)\}}{\partial\theta}, \text{ where } w_i = 1/\pi_i .$$

That is, when the explicit form of the population likelihood is not available, we weight instead the sample likelihood and solve the weighted equations.

4. On the use of sampling weights when estimating regression models with survey data

Magee, Robb and Burbidge (1998), from now on (MRB1998), argue that when the population regression coefficient is of interest, the use of sampling weights can be desirable in regression models with complex survey data. A two-step maximum likelihood estimator is proposed as an alternative to ordinary least square and weighted least squares.

Before dealing with the problem, and defining the sample distribution mathematically, let us introduce the following notations: f_p and $E_p(\cdot)$ denote the pdf and the mathematical expectation of the population distribution, respectively, and f_s and $E_s(\cdot)$ denote the pdf and the mathematical expectation of the sample distribution.

4.1. Population model

We now consider the population values (x_i, y_i), $i = 1,\ldots,N$ as random variables, which are independent realizations from a distribution with probability density function $f_p(x, y \mid \theta)$, indexed by a vector parameter θ.

4.2. Sampling scheme

We consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s)$, and sampling weight $w_i = 1/\pi_i$; $i = 1,...,N$. The π_i, s may depend on the population values (x, y) as well as on other factors unknown to the researches, call these factors Z. Assume that $\pi_i \sim h(\pi|x, y, \gamma)$ where γ is a parameter indexing h. Thus, we now consider the population values (x_i, y_i, π_i) , $i = 1,\ldots,N$, as random variables, which are independent realizations from a distribution with probability density function (pdf):

$$f_p(x, y, \pi|\theta, \gamma) = f_p(x, y|\theta) \times h_p(\pi|\gamma)$$

The researcher has a sample of n observations (x_i, y_i, π_i) , $i \in s$. Each $i \in U$ is included in s with probability π_i.

The parameter of interest is the regression coefficient $\mathbf{\beta} = (\beta_0, \beta_1)$:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where $E_p(u_i x_i) = 0$, $i = 1,\ldots,N$.

4.3. Two-step maximum likelihood (ML) estimators

We consider an estimator that uses structure on the population probability density function imposed by modelling the process that generates the π_i, s. Assume that:

$$f_p(x, y, \pi|\theta, \gamma) = f_p(x, y|\theta) \times h_p(\pi|\gamma)$$

can be described as:

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

where $u \underset{p}{\sim} N\left(0, \sigma_y^2\right)$, and

$$\pi^* = \ln \pi = \gamma_0 + \gamma_1 x + \gamma_2 xy + v \tag{2}$$

where $v \underset{p}{\sim} N\left(0, \sigma_{\pi^*}^2\right)$. Also, assume that u and v are independent of each other and of x .

4.4. Sample model

The probability density function of y given x_i , in the sample is given by:

$$\begin{aligned} f_s\left(y \middle| x_i\right) &= f_p\left(y \middle| x_i, i \in s\right) \\ &= \frac{f_p\left(y \middle| x_i\right) \times \Pr\left(i \in s \middle| x_i, y\right)}{\int f_p\left(y \middle| x_i\right) \times \Pr\left(i \in s \middle| x_i, y\right) dy} \end{aligned} \tag{3}$$

Under the conditions of equation (1), we have:

$$\left(y \middle| x_i, i \in s\right) \sim N\left(\beta_0 + x_i\left(\beta_1 + \gamma_2 \sigma_y^2\right), \sigma_y^2\right) \tag{4}$$

Similarly, the probability density function of π^* given $\left(x_i, y_i\right)$, in the sample is given by:

$$\begin{aligned} h_s\left(\pi^* \middle| x_i, y_i\right) &= h_p\left(\pi^* \middle| x_i, y_i, i \in s\right) \\ &= \frac{h_p\left(\pi^* \middle| x_i, y_i\right) \times \Pr\left(i \in s \middle| x_i, y_i, \pi^*\right)}{\int h_p\left(\pi^* \middle| x_i, y_i\right) \times \Pr\left(i \in s \middle| x_i, y_i, \pi^*\right) d\pi^*} \end{aligned} \tag{5}$$

From equation (2), we have:

$$\Pr\left(i \in s \middle| x_i, y_i, \pi^*\right) = \pi = \exp\left(\pi^*\right)$$

Thus, under the conditions of equation (2), we obtain:

$$\left(\pi^* \middle| x_i, y_i, i \in s\right) \sim N\left(\gamma_0^* + \gamma_1 x_i + \gamma_2 x_i y_i, \sigma_{\pi^*}^2\right) \tag{6}$$

where $\gamma_0^* = \gamma_0 + \sigma_{\pi^*}^2$.

The two-step maximum likelihood method can be performed as follows:

First step:

Estimate of γ_2 can be obtained from ordinary least squares (OLS) estimation of:

$$\pi_i^* = \gamma_0^* + \gamma_1 x_i + \gamma_2 x_i y_i + error_i,\ i = 1,\dots,n \tag{7}$$

Second step:

Estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ based on equation (4). A consistent estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ can be obtained from the OLS estimation, (or ML estimation, because of normality), of

$$y_i = \beta_0 + x_i \beta_1^* + error_i,\ i = 1,\dots,n$$

where $\beta_1^* = \beta_1 + \hat{\gamma}_2 \sigma_y^2$, which are given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1^* \bar{x}$$

$$\beta_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Thus,

$$\hat{\beta}_1 = \hat{\beta}_1^* - \hat{\gamma}_2 \hat{\sigma}_y^2 \tag{8}$$

5. The use of sample distribution when estimating regression models with survey data

In recent articles by Krieger and Pfeffermann (1997), Pfeffermann, Krieger, and Rinott (1998), from now on (PKR1998) and Pfeffermann and Sverchkov (1999), the authors introduced an analytic likelihood-based inference from complex survey data under informative sampling. Their basic idea is to derive the distribution of the sample data by modelling the population distribution and the conditional expectation of the first order sample inclusion probabilities. Once this sample distribution is extracted, standard likelihood-based inferential methods can be used to obtain estimates of the parameters of the population model under consideration.

The sample distribution refers to the superpopulation distribution of the sample measurements, as induced by the population model and the sample selection scheme with the selected sample of units held fixed. In order to describe the fundamental idea behind this approach, we assume full response. Let

$\mathbf{x}_i = (x_{i1},..., x_{ip})'$, $i \in U$ be the values of a vector of auxiliary variables, $x_1,..., x_p$, and $\mathbf{z} = \{z_1,..., z_N\}$ 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 $\pi_i = \Pr(i \in s)$, 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 $i \in U$. Since $\pi_1,..., \pi_N$ are defined by the realizations $(\mathbf{x}_i, y_i, \mathbf{z}_i), i = 1,..., N$, therefore, they are random realizations defined on the space of possible populations. The sample S consists of the subset of U selected at random by the sampling scheme with inclusion probabilities $\pi_1,..., \pi_N$. Denote by $\mathbf{I} = (I_1,..., I_N)'$ the N by one 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. The sample s is defined accordingly as $s = \{i \mid i \in U, I_i = 1\}$ and its complement by $c = \bar{s} = \{i \mid i \in U, I_i = 0\}$. We assume probability sampling, so that $\pi_i = \Pr(i \in s) > 0$ for all units $i \in U$.

5.1. Population model

We now consider the population values $y_1,..., y_N$ as random variables, which are independent realizations from a distribution with probability density function (pdf) $f_p(y_i \mid \theta)$, indexed by a vector of parameters θ. Assume that the population pdf depends on known values of the auxiliary variables $\mathbf{x}_i$, so that $y_i \sim f_p(y_i \mid \mathbf{x}_i, \theta)$.

5.2. Sample model

We consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s)$ be the first order inclusion probability of the ith population unit, and the sampling weights $w_i = 1/\pi_i$ is the sampling weight of unit $i \in U$. In practice, the π_i's may depend on the population values $(\mathbf{x}, \mathbf{y}, \mathbf{z})$. We express this dependence by writing:

$$\pi_i = \Pr(i \in s \mid \mathbf{x}, \mathbf{y}, \mathbf{z}) \quad \text{for all units } i \in U$$

According to Krieger and Pfeffermann (1997), the (marginal) sample pdf of y_i is defined as:

$$\begin{aligned} f_s(y_i \mid \mathbf{x}_i, \theta, \gamma) &= f_p(y_i \mid \mathbf{x}_i, \theta, \gamma, i \in \mathrm{s}) \\ &= \frac{E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma) \times f_p(y_i \mid \mathbf{x}_i, \theta)}{E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma)} \end{aligned} \tag{9}$$

where θ is the parameter indexing the population distribution, and γ is the informativeness parameter indexing:

$$E_p(\pi_i \mid \mathbf{x}_i, \theta, \gamma) = \int E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma) \times f_p(y_i \mid \mathbf{x}_i, \theta) dy_i$$

Note that $E_p(\pi_i \mid \mathbf{y}_i) = E_{\mathbf{z}_i \mid \mathbf{y}_i} E_p(\pi_i \mid \mathbf{y}_i, \mathbf{z}_i)$, so that $\mathbf{z}_i$ is integrated out in equation (9). See Eideh and Nathan (2006).

The question that arises is how we can identify and estimate $E_p(\pi_i \mid y_i, \mathbf{x}_i)$ based only on the sample data $\{y_i, \mathbf{x}_i, w_i;\ i \in s\}$. Pfeffermann and Sverchkov (1999) proved the following relationships: for vector of random variables $(y_i, \mathbf{x}_i)$, the following relationships hold:

$$E_s(w_i \mid y_i, x_i) = \{E_p(\pi_i \mid y_i, x_i)\}^{-1} \tag{10a}$$

$$E_p(y_i \mid \mathbf{x}_i) = \{E_s(w_i \mid \mathbf{x}_i)\}^{-1} E_s(w_i y_i \mid \mathbf{x}_i) \tag{10b}$$

$$E_s(w_i) = \{E_p(\pi_i)\}^{-1} \tag{10}$$

5.3. Estimation

Having derived the sample distribution, (PKR1998) proved that if the population measurements y_i are independent, then as $N \to \infty$ (with n fixed) the sample measurements are asymptotically independent, so we can apply standard inference procedures to complex survey data by using the marginal sample distribution for each unit. Based on the sample data $\{y_i, \mathbf{x}_i, w_i;\ i \in s\}$, (PKR1998) proposed a two-step estimation method.

Step one:

Estimate the informativeness parameters γ using equation (10a), using regression analysis. Denote the resulting estimate of γ by $\tilde{\gamma}$.

Step two:

Substitute $\tilde{\gamma}$ in the sample log-likelihood function, and then maximize the resulting sample log-likelihood function with respect to the population parameters, θ:

$$\begin{aligned} l_{rs}(\theta,\tilde{\gamma}) &= l_{srs}(\theta) - \sum_{i=1}^{n} \log E_p(\pi_i \mid \mathbf{x}_i, \theta, \tilde{\gamma}) \\ &= l_{srs}(\theta) + \sum_{i=1}^{n} \log E_s(w_i \mid \mathbf{x}_i, \theta, \tilde{\gamma}) \end{aligned} \qquad (11)$$

where $l_{rs}(\theta,\tilde{\gamma})$ is the sample log-likelihood after substituting $\tilde{\gamma}$ in the sample log-likelihood function, and where

$$l_{srs}(\theta) = \sum_{i=1}^{n} \log\{f_p(y_i \mid \mathbf{x}_i, \theta)\}$$

is the classical log-likelihood obtained by ignoring the sample design.

5.4. Illustration

5.4.1. Population model

Assume the following population model:

$$y_i = \beta_0 + \beta_1 x_i + u_i \qquad (12)$$

where $u_i \underset{p}{\sim} N(0,\sigma_y^2)$ and $E_p(u_i x_i) = 0$, so that $y_i \underset{p}{\sim} N(\beta_0 + \beta_1 x_i, \sigma_y^2)$, $i = 1,\ldots,N$.

Now assume that:

$$E_p(\pi_i \mid \mathbf{x}_i, y_i, \gamma) = \exp(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i),\ i = 1,\ldots,N \qquad (13)$$

We interpret this exponential inclusion probability model approximation (13) in the spirit of probability proportional to size sampling scheme as follows. Let the size measure be:

$$d_i = \exp(\gamma_{0a} + \gamma_1 x_i + \gamma_2 x_i y_i + v_i)$$

where $E_p(v_i) = 0$ and $V_p(v_i) = \sigma_{\pi^*}^2$.

Let

$$\pi_i = \frac{n d_i}{T_d},\ T_d = \sum\nolimits_{i=1}^{N} d_i$$

Assume N is large enough so that the difference between $N\bar{d}$ and $E\left(N\bar{d}\right)=N\mu_d$ can be ignored, so that $\pi_i = nd_i/T_d \cong nd_i/N\mu_d$, or $\pi_i \propto d_i$. Furthermore, since

$$\begin{aligned}\pi_i &= \frac{nd_i}{T_d} \cong \frac{nd_i}{N\mu_d} \\ &= \frac{n}{N\mu_d}\exp\left(\gamma_{0a}+\gamma_1 x_i+\gamma_2 x_i y_i+v_i\right) \\ &= \exp\left(\gamma_0+\gamma_1 x_i+\gamma_2 x_i y_i+v_i\right)\end{aligned}$$

where $\gamma_0=\gamma_{0a}+\ln\left(\frac{n}{N\mu_d}\right)$, therefore

$$\pi_i^* = \ln\pi_i = \gamma_0+\gamma_1 x_i+\gamma_2 x_i y_i+v_i \qquad (14)$$

where $E_p\left(v_i\right)=0$ and $V_p\left(v_i\right)=\sigma^2_{\pi^*}$.

Under these assumptions and using Taylor series approximation, we can show that:

$$E_p\left(\pi_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}\right)=\exp\left(\gamma_0+\gamma_1 x_i+\gamma_2 x_i y_i\right)$$

where $\boldsymbol{\gamma}=\left(\gamma_0,\gamma_1,\gamma_2\right)$.

Comment 1: See the similarity between (14) and (2).

5.4.2. Sample model

By substituting (12) and (13) in (9), we have:

$$y_i \underset{s}{\sim} N\left(\beta_0+x_i\left(\beta_1+\gamma_2\sigma_y^2\right),\sigma_y^2\right) \qquad (15)$$

Comment 2: Note that (4) and (15) are similar.

5.4.3. Two-step estimation

First step:

Estimate the informativeness parameter γ_2 using (13) and (10) as follows:

$$E_s(w_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}) = \exp(-(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i)) \tag{16}$$

Using Taylor series approximation, we have $E \ln Y \cong \ln E(Y)$, so that

$$\begin{aligned} \ln E_s(w_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}) &= E_s(\ln w_i \mid \mathbf{x}_i, y_i, \boldsymbol{\gamma}) \\ &= -(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i) \end{aligned} \tag{17}$$

Hence,

$$\ln w_i = -\ln \pi_i = -(\gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i)$$

or

$$\ln \pi_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i y_i + error_i,\ i \in s \tag{18}$$

Therefore, estimation of γ_2, denoted by $\tilde{\gamma}_2$, can be obtained from OLS estimation of (18), or you can use nonlinear regression model.

Second step:

Estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ can be obtained by using OLS estimation (or ML estimation method, because of linearity) of the following regression model:

$$y_i = \beta_0 + x_i \beta_1^* + error_i\ i = 1, \ldots, n \tag{19}$$

where $\beta_1^* = \beta_1 + \tilde{\gamma}_2 \sigma_y^2$, which are given by:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1^* \bar{x}$$

$$\beta_1^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

So that,

$$\hat{\beta}_1 = \hat{\beta}_1^* - \hat{\gamma}_2 \hat{\sigma}_y^2$$

which are similar to (8).

6. Conclusion

In this paper we investigated two methods on the use of sampling weights when fitting regression model to survey data under informative probability sampling design. We showed that the only difference between the method proposed by MRB1998 and the method proposed by PKR1998 is in estimating the informativeness parameter γ_2. In MRB1998 method the estimator of γ_2 is

ML estimator, while in PKR1998 method the estimator of γ_2 is only the OLS. The intercepts γ_0^* and γ_0 have the same functional form.

The MRB1998 consider the estimator that uses more structure on the population density imposed by modelling the process generating the first order inclusion probabilities, and in their paper they consider only one model, see equation (2); while PKR1998 incorporate the sampling weights via the conditional expectation of first order inclusion probabilities given the response variable, and they consider only two models. Subsequently, Eideh (2003) proposed logit and probit models. In this paper we justified that the models that generate the first order inclusion probabilities are similar, see equations 7, 14 and 18.

Furthermore, in the last decade survey statisticians have been using the sample distribution for analysis of survey data under informative probability sampling design in several applications, in particular: prediction of finite population total under single stage sampling and two-stage sampling; fitting multilevel modelling; fitting time series models; s mall area estimation; estimating generalized linear models. A lso, they have proposed tests of informativeness of sampling design and the test of ignorability of nonresponse in surveys, which is not the case for MRB1998. Hence, the use of sample distribution in analysis of survey data applies. Consequently, it is suggested to use the PKR1998.

We hope that this investigation will encourage further theoretical, empirical and practical research in these directions.

Acknowledgements

The author is grateful to the referees for their valuable comments.

REFERENCES

BINDER, D. A., (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279–292.

CHAMBERS, R., SKINNER, C., (2003). Analysis of Survey Data. New York: John Wiley.

EIDEH, A. H., (2003). Estimation for Longitudinal Survey Data under Informative Sampling, PhD Thesis, Department of Statistics, Hebrew University of Jerusalem.

EIDEH, A. H., (2007). A Correction Note on Small Area Estimation. International Statistical Review. 75, 122–123.

EIDEH A. H., (2008). Estimation and Prediction of Random Effects Models for Longitudinal Survey Data under Informative Sampling. Statistics in Transition – New Series. Volume 9, Number 3, December 2008, pp. 485–502.

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, pp. 18–29.

EIDEH, A. H., (2010). Analytic Inference of Complex Survey Data under Informative Probability Sampling. Proceedings of the Tenth Islamic Countries Conference on Statistical Sciences (ICCS-X), Volume I. The Islamic Countries Society of Statistical Sciences, Lahore: Pakistan, (2010). Edited by: Zeinab Amin and Ali S. Hadi. The American University in Cairo: pp. 507–536.

EIDEH, A. H., (2011). Informative Sampling on Two Occasions: Estimation and Prediction. Pakistan Journal of Statistics and Operation Research (PJSOR). Pak.j.stat.oper.res. VII No.2, 2011, pp. 283–303.

EIDEH, A. H., (2012a). Fitting Variance Components Model and Fixed Effects Model for One-Way Analysis of Variance to Complex Survey Data. Communications in Statistics – Theory and Methods, 41, pp. 3278–3300.

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

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. 3052–3069. [Corrigendum, 137 (2007), p 628].

EIDEH, A. H., NATHAN, G., (2009). Two-Stage Informative Cluster Sampling with application in Small Area Estimation. Journal of Statistical Planning and Inference, 139, pp. 3088–3101.

KRIEGER, A. M, PFEFFERMANN, D., (1997). Testing of distribution functions from complex sample surveys. Journal of Official Statistics. 13: pp. 123–142.

MAGEE, L., ROBB, A. L., BURBIDGE, J. B., (1998). On the use of sampling weights when estimating regression models with survey data. Journal of Econometrics 84, 251–271.

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

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

PFEFFERMANN, D., SVERCHKOV, M., (2003). Fitting Generalized Linear Models under Informative Probability Sampling. In: Analysis of Survey Data. (Eds. R. Chambers and C. J. Skinner). New York: Wiley, pp. 175–195.

SARNDAL, C-E., SWENSSON, B., WRETMAN, J., (1992). Model assisted survey sampling, New York: Springer.

SKINNER, C. J., (1994). Sample models and weights. American Statistical Association Proceedings of the Section on Survey Research Methods, 133–142.

SKINNER, C. J., HOLT, D., SMITH, T. M. F (Eds.), (1989). Analysis of Complex Surveys, New York: Wiley.

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

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 197–220

EFFECTIVE ROTATION PATTERNS FOR MEDIAN ESTIMATION IN SUCCESSIVE SAMPLING

Kumari Priyanka[1], **Richa Mittal**[2]

ABSTRACT

The present work deals with the problem of estimation of population median at current occasion in two-occasion successive sampling. Best linear unbiased estimators have been proposed by utilizing additional auxiliary information, readily available on both the occasions. Asymptotic variances of the proposed estimators are derived and the optimum replacement policies are discussed. The behaviours of the proposed estimators are analyzed on the basis of data from natural populations. Simulation studies have been carried out to measure the precision of the proposed estimators.

Key words: population median, successive sampling, auxiliary information, optimum replacement policy.

1. Introduction

When the value of the study character of a finite population is subject to change (dynamically) over time, a survey carried out on a single occasion will provide information about the characteristics of the surveyed population for the given occasion only and will not give any information on the nature of change of the characteristic over different occasions and the average value of the characteristic over all occasions or the most recent occasion. To meet these requirements, sampling is done on successive occasions that provide a strong tool for generating the reliable estimates at different occasions. The problem of sampling on two successive occasions was first considered by Jessen (1942), and later this idea was extended by Patterson (1950), Narain (1953), Eckler (1955), Gordon (1983), Arnab and Okafor (1992), Feng and Zou (1997), Singh and Singh (2001), Singh and Priyanka (2008), Singh et al. (2012), Bandyopadhyay and Singh (2014), and many others.

[1] Department of Mathematics, Shivaji College (University of Delhi), New Delhi – 110027, India. E-mail: priyanka.ism@gmail.com.

[2] Department of Mathematics, Shivaji College (University of Delhi), New Delhi – 110027, India. E-mail: sovereignricha@gmail.com.

All the abovestudies were concerned with the estimation of population mean or variance on two or more occasions.

There are many problems of practical interest which involves variables with extreme values that strongly influence the value of the mean. In such situations the study variable is having highly skewed distributions. For example, the study of environmental issues, the study of social evil such as abortions, the study of income, expenditure, etc. In these situations, the mean may offer results which are not representative enough because the mean moves with the direction of the asymmetry. The median, on the other hand, is unaffected by extreme values.

Most of the studies related to medians have been developed by assuming simple random sampling or its ramification in stratified random sampling (Gross (1980), Sedransk and Meyer (1978), Smith and Sedransk (1983) consider only the variable of interest without making explicit use of auxiliary variables. Some of the researchers, namely Chambers and Dunstan (1986), Kuk and Mak (1989), Rao et al. (1990), Rueda et al.(1998), Khoshnevisan et al. (2002), Singh and Solanki (2013) etc., make use of auxiliary variables to estimate the population median).

It is to be mentioned that a large number of estimators for estimating the population mean at current occasion have been proposed by various authors, however only a few efforts (namely Martinez-Miranda et al. (2005), Singh et al. (2007), Rueda et al. (2008) and Gupta et al. (2008)) have been made to estimate the population median on the current occasion in two occasions successive sampling. It is well known that the use of auxiliary information at the estimation stage can typically increase the precision of estimates of a parameter. To the best of our knowledge, no effort has been made to use additional auxiliary information readily available on both the occasions to estimate population median at current occasion in two-occasion successive sampling.

Motivated with the above arguments and utilizing the information on an additional auxiliary variable, readily available on both the occasions, the best linear unbiased estimators for estimating the population median on c urrent occasion in two-occasion successive sampling have been proposed. It has been assumed that the additional auxiliary variable is stable over the two-occasions.

The paper is spread over ten sections. Sample structure and notations have been discussed in section 2. I n section 3 the proposed estimator has been formulated. Properties of proposed estimators including variances are derived under section 4. Minimum variance of the proposed estimator is derived in section 5. Practicability of the proposed estimator is also discussed. In section 6 optimum replacement policies are discussed. Section 7 contains comparison of the proposed estimator with the natural sample median estimator when there is no matching from the previous occasion and the estimator when no a dditional auxiliary information has been used. Practicability of the estimator Δ is also discussed. In section 8 simulation studies have been carried out to investigate the performance of the proposed estimators. The results obtained as a result of empirical and simulation studies have been elaborated in section 9. Finally, the conclusion of the entire work has been presented in section 10.

2. Sample structures and notations

Let $U = (U_1, U_2, ---, U_N)$ be the finite population of N units, which has been sampled over two occasions. It is assumed that the size of the population remains unchanged but values of the unit change over two occasions. Let the character under study be denoted by x (y) on the first (second) occasion respectively. It is further assumed that information on a n auxiliary variable z (with known population median) is available on both the occasions. A simple random sample (without replacement) of n units is taken on the first occasion. A random sub-sample of $m = n\lambda$ units is retained (matched) for use on the second occasion. Now, at the current occasion a simple random sample (without replacement) of $u = (n - m) = n\mu$ units is drawn afresh from the remaining (N - n) units of the population so that the sample size on the second occasion is also n. λ and μ, ($\lambda + \mu = 1$) are the fractions of matched and fresh samples respectively at the second (current) occasion. The following notations are considered for further use:

M_x, M_y, M_z : Population median of x, y and z, respectively.

$\hat{M}_{x(n)}, \hat{M}_{x(m)}, \hat{M}_{y(m)}, \hat{M}_{y(u)}, \hat{M}_{z(n)}, \hat{M}_{z(m)}, \hat{M}_{z(u)}$: Sample median of the respective variables of the sample sizes shown in suffices.

$\rho_{yx}, \rho_{xz}, \rho_{yz}$: The Correlation coefficient between the variables shown in suffices.

3. Formulation of estimator

To estimate the population median M_y on the current (second) occasion, the minimum variance linear unbiased estimator of M_y under SRSWOR sampling scheme have been proposed and is given as

$$T = \left\{\alpha_1\hat{M}_{y(u)} + \alpha_2\hat{M}_{y(m)}\right\} + \left\{\alpha_3\hat{M}_{x(m)} + \alpha_4\hat{M}_{x(n)}\right\} + \left\{\alpha_5\hat{M}_{z(u)} + \alpha_6\hat{M}_{z(m)} + \alpha_7\hat{M}_{z(n)} + \alpha_8 M_z\right\} \tag{1}$$

where $\alpha_i \left(i = 1,\ 2,\ ---,\ 8\right)$ are constants to be determined so that

(i) The estimator T becomes unbiased for M_y and

(ii) The variance of T attains a minimum

For unbiasedness, the following conditions must hold:

$\left(\alpha_1 + \alpha_2\right) = 1$, $\left(\alpha_3 + \alpha_4\right) = 0$ and $\left(\alpha_5 + \alpha_6 + \alpha_7 + \alpha_8\right) = 0$.

Substituting $\alpha_1 = \phi_1, \alpha_3 = \beta_1$ and $\alpha_8 = -(\alpha_5 + \alpha_6 + \alpha_7)$ in equation (1), the estimator *T* takes the following form:

$$T = \left\{\phi_1 \hat{M}_{y(u)} + (1 - \phi)\hat{M}_{y(m)}\right\} + \beta_1\left\{\hat{M}_{x(m)} - \hat{M}_{x(n)}\right\} + \left\{\alpha_5\left(\hat{M}_{z(u)} - M_z\right)\right.$$

$$\left. + \alpha_6\left(\hat{M}_{z(m)} - M_Z\right) + \alpha_7\left(\hat{M}_{z(n)} - M_Z\right)\right\}$$

$$= \phi_1\left\{\hat{M}_{y(u)} + k_1\left(\hat{M}_{z(u)} - M_Z\right)\right\} + (1 - \phi_1)\left\{\hat{M}_{y(m)} + k_2\left(\hat{M}_{x(m)} - \hat{M}_{x(n)}\right) + \right.$$

$$\left. k_3\left(\hat{M}_{z(m)} - M_Z\right) + k_4\left(\hat{M}_{z(n)} - M_Z\right)\right\}$$

$$T = \phi_1 T_1 + (1 - \phi_1)\ T_2 \tag{2}$$

where $T_1 = \hat{M}_{y(u)} + k_1\ \left(\hat{M}_{z(u)} - M_Z\right)$ is based on the sample of size *u* drawn afresh at current occasion and the estimator

$$T_2 = \left\{\hat{M}_{y(m)} + k_2\left(\hat{M}_{x(m)} - \hat{M}_{x(n)}\right) + k_3\left(\hat{M}_{z(m)} - M_Z\right) + k_4\left(\hat{M}_{z(n)} - M_Z\right)\right\}$$

is based on the sample of size *m* matched form previous occasion.

$k_1 = \dfrac{\alpha_5}{\phi_1}$, $k_2 = \dfrac{\beta_1}{1 - \phi_1}$, $k_3 = \dfrac{\alpha_6}{1 - \phi_1}$, $k_4 = \dfrac{\alpha_7}{1 - \phi_1}$ and ϕ_1 are the unknown constants to be determined so as to minimize the variance of estimator *T*.

Remark 3.1. For estimating the median on each occasion, the estimator T_1 is suitable, which implies that more belief on T_1 could be shown by choosing ϕ_1 as 1 (or close to 1), while for estimating the change from one occasion to the next, the estimator T_2 could be more useful so ϕ_1 be chosen as 0 (or close to 0). For asserting both the problems simultaneously, the suitable (optimum) choice of ϕ_1 is required.

4. Properties of the estimator *T*

The properties of the proposed estimator *T* are derived under the following assumptions:

(i) Population size is sufficiently large *(i.e. N→∞),* therefore finite population corrections are ignored.

(ii) As *N→∞,* the distribution of bivariate variable *(a, b)* where *a* and *b* $\in \{x, y, z\}$ and *a ≠ b* approaches a continuous distribution with marginal densities $f_a(\cdot)$ and $f_b(\cdot)$ for *a* and *b* respectively, see Kuk and Mak (1989).

(iii) The marginal densities $f_x(\cdot)$, $f_y(\cdot)$ and $f_z(\cdot)$ are positive.

(iv) The sample medians $\hat{M}_{x(n)}$, $\hat{M}_{x(m)}$, $\hat{M}_{y(m)}$, $\hat{M}_{y(u)}$, $\hat{M}_{z(n)}$, $\hat{M}_{z(m)}$ and $\hat{M}_{z(u)}$ are consistent and asymptotically normal (see Gross (1980)).

(v) Following Kuk and Mak (1989), let P_{ab} be the proportion of elements in the population such that $a \le M_a$ and $b \le M_b$ where *a* and $b \in \{x, y, z\}$ and $a \neq b.$

(vi) The following large sample approximations are assumed:

$\hat{M}_{y(u)} = M_y(1+e_0)$, $\hat{M}_{y(m)} = M_y(1+e_1)$, $\hat{M}_{x(m)} = M_x(1+e_2)$, $\hat{M}_{x(n)} = M_x(1+e_3)$, $\hat{M}_{z(u)} = M_z(1+e_4)$, $\hat{M}_{z(m)} = M_z(1+e_5)$ and $\hat{M}_{z(n)} = M_z(1+e_6)$ such that $|e_i| < 1$ $\forall\, i = 0, 1, 2, 3, 4, 5, 6.$

The values of various related expectations can be seen in Allen et al. (2002) and Singh (2003). Under the above transformations, the estimators T_1 and T_2 take the following forms:

$$T_1 = M_y(1+e_0) + k_1 M_z e_4 \tag{3}$$

$$T_2 = M_y(1+e_1) + k_2 M_x(e_2 - e_3) + M_z(k_3 e_5 + k_4 e_6) \tag{4}$$

Thus we have the following theorems:

Theorem 4.1. *T* is unbiased estimator of M_y.

Proof: Since T_1 and T_2 are difference and difference-type estimators, respectively, they are unbiased for M_y. The combined estimator *T* is a convex linear combination of T_1 and T_2, hence it is also an unbiased estimator of M_y.

Theorem 4.2. Ignoring the finite population corrections, the variance of *T* is

$$V(T) = \phi_1^2\, V(T_1) + (1 - \phi_1)^2 V(T_2) \tag{5}$$

where
$$V(T_1) = \frac{1}{u}\xi_1 \tag{6}$$

and
$$V(T_2) = \frac{1}{m}\xi_2 + \left(\frac{1}{m} - \frac{1}{n}\right)\xi_3 + \frac{1}{n}\xi_4 \tag{7}$$

$$\xi_1 = A_1 + k_1^2 A_2 + 2k_1 A_3, \xi_2 = A_1 + k_3^2 A_2 + 2k_3 A_3,$$
$$\xi_3 = k_2^2 A_4 + 2k_2 A_5 + 2k_2 k_3 A_6, \xi_4 = k_4^2 A_2 + 2k_4 A_3 + 2k_3 k_4 A_2,$$
$$A_1 = \frac{1}{4}\left\{ f_y\left(M_y\right)\right\}^{-2}, A_2 = \frac{1}{4}\left\{ f_z\left(M_z\right)\right\}^{-2},$$
$$A_3 = \left(P_{yz} - 0\cdot 25\right)\left\{ f_y\left(M_y\right)\right\}^{-1}\left\{ f_z\left(M_z\right)\right\}^{-1}, A_4 = \frac{1}{4}\left\{ f_x\left(M_x\right)\right\}^{-2},$$
$$A_5 = \left(P_{yx} - 0\cdot 25\right)\left\{ f_y\left(M_y\right)\right\}^{-1}\left\{ f_x\left(M_x\right)\right\}^{-1} \text{ and}$$
$$A_6 = \left(P_{xz} - 0\cdot 25\right)\left\{ f_x\left(M_x\right)\right\}^{-1}\left\{ f_z\left(M_z\right)\right\}^{-1}.$$

Proof: The variance of T is given by

$$V(T) = E\left(T - M_y\right)^2 = E\left[\phi_1\left(T_1 - M_y\right) + \left(1 - \phi_1\right)\left(T_2 - M_y\right)\right]^2$$
$$= \phi_1^2 V\left(T_1\right) + \left(1 - \phi_1\right)^2 V\left(T_2\right) + \phi_1\left(1 - \phi_1\right) \operatorname{cov}\left(T_1, T_2\right) \tag{8}$$

where $V\left(T_1\right) = E\left(T_1 - M_y\right)^2$ and $V\left(T_2\right) = E\left(T_2 - M_y\right)^2$.

As T_1 and T_2 are based on two independent samples of sizes *u* and *m* respectively, hence cov $\left(T_1, T_2\right) = 0$.

Now, substituting the expressions of T_1 and T_2 from equations (3) and (4) in equation (8), taking expectations and ignoring finite population corrections, we have the expression for variance of *T* as in equation (5).

5. Minimum variance of the estimator *T*

Since the variance of the estimator *T* in equation (5) is the function of unknown constants k_1, k_2, k_3, k_4 and ϕ_1, therefore it is minimized with respect to k_1, k_2, k_3, k_4 and ϕ_1 and subsequently the optimum values of k_1, k_2, k_3, k_4 and ϕ_1 are obtained as

$$k_1^* = \frac{-A_3}{A_2} \tag{9}$$

$$k_2^* = \frac{A_3 A_4 A_6 - A_2 A_4 A_5}{A_4\left(A_2 A_4 - A_6^2\right)} \tag{10}$$

$$k_3^* = \frac{-A_3 A_4 + A_5 A_6}{\left(A_2 A_4 - A_6^2\right)} \tag{11}$$

$$k_4^* = \frac{A_3 A_6^2 - A_2 A_5 A_6}{A_2\left(A_2 A_4 - A_6^2\right)} \tag{12}$$

$$\phi_{1opt\cdot} = \frac{V\left(T_2\right)}{V\left(T_1\right) + V\left(T_2\right)} \tag{13}$$

Using the optimum values of $k_i\text{'}s\ \left(i = 1,\ 2,\ 3,\ 4\right)$ in equation (6) and (7), we get the optimum variances of T_1 and T_2 as

$$V\left(T_1\right)_{opt\cdot} = \frac{1}{u} A_7 \tag{14}$$

$$V\left(T_2\right)_{opt\cdot} = \frac{1}{m} A_8 + \left(\frac{1}{m} - \frac{1}{n}\right) A_9 + \frac{1}{n} A_{10} \tag{15}$$

where $A_7 = A_1 + k_1^{*^2} A_2 + 2k_1^* A_3,\ A_8 = A_1 + k_3^{*^2} A_2 + 2k_3^* A_3$

$$A_9 = k_2^{*^2} A_4 + 2k_2^* A_5 + 2k_2^* k_3^* A_6 \quad \text{and}$$

$$A_{10} = k_4^{*^2} A_2 + 2k_4^* A_3 + 2k_3^* k_4^* A_2 .$$

Further, substituting the values of $V\left(T_1\right)_{opt\cdot}$ and $V\left(T_2\right)_{opt\cdot}$ from equations (14) and (15) in equation (13), we get the optimum values of $\phi_{1opt\cdot}$ with respect to $k_i^*\text{'}s\ \left(i = 1,\ 2,\ 3,\ 4\right)$ as

$$\phi_{1opt\cdot}^* = \frac{V\left(T_2\right)_{opt\cdot}}{V\left(T_1\right)_{opt\cdot} + V\left(T_2\right)_{opt\cdot}} \tag{16}$$

Again substituting the value of $\phi_{1opt\cdot}^*$ from equation (16) in equation (5), we get the optimum variance of T as

$$\mathrm{V}\left(\mathrm{T}\right)_{\mathrm{opt}\cdot} = \frac{\mathrm{V}\left(\mathrm{T}_1\right)_{\mathrm{opt}\cdot} \mathrm{V}\left(\mathrm{T}_2\right)_{\mathrm{opt}\cdot}}{\mathrm{V}\left(\mathrm{T}_1\right)_{\mathrm{opt}\cdot} + \mathrm{V}\left(\mathrm{T}_2\right)_{\mathrm{opt}\cdot}} \tag{17}$$

Further, substituting the value from (14) and (15) in equation (16) and (17), we get the simplified values of $\phi_{1opt\cdot}^*$ and $V\left(T\right)_{opt\cdot}$ as

$$\phi_{1opt\cdot}^* = \frac{\mu\left(A_{11} + \mu A_{12}\right)}{\mu^2 A_{12} + \mu^2 A_{13} + A_7} \tag{18}$$

$$V\left(T\right)_{opt\cdot} = \frac{1}{n} \frac{A_7\left(A_{11} + \mu A_{12}\right)}{\left(\mu^2 A_{12} + \mu A_{13} + A_7\right)} \tag{19}$$

where $A_{11} = A_8 + A_{10},\ A_{12} = A_9 - A_{10},\ A_{13} = A_{11} - A_7$ and μ is the fraction of fresh sample at current occasion for the estimator *T*.

5.1. Estimator *T* in practice

The main difficulty in using the proposed estimator *T* defined in equation (2) is the availability ofk_i's $(\mathrm{i}=1, 2, 3, 4)$ as the optimum values of k_i's $(\mathrm{i}=1, 2, 3, 4)$ depend on the population parameters $P_{yx}, P_{yz}, P_{xz}, f_y(M_y), f_x(M_x)$ and $f_z(M_z)$. If these parameters are known, the proposed estimator can be easily implemented. Otherwise, which is the most often situation in practice, the unknown population parameters are replaced by their respective sample estimates. The population proportions P_{yx}, P_{yz} and P_{xz} are replaced by the sample estimates $\hat{P}_{yx}, \hat{P}_{yz}$ and $\hat{P}_{xz}$ respectively, and the marginal densities $f_y(M_y), f_x(M_x)$ and $f_z(M_z)$ can be substituted by their kernel estimator or nearest neighbour density estimator or generalized nearest neighbour density estimator related to the kernel estimator (Silverman (1986)). Here, the marginal densities $f_y(M_y)$, $f_x(M_x)$ and $f_z(M_z)$ are replaced by $\hat{f}_y(\hat{M}_{y(m)}), \hat{f}_x(\hat{M}_{x(n)})$ and $\hat{f}_z(\hat{M}_{z(n)})$ respectively, which are obtained by the method of generalized nearest neighbour density estimation related to the kernel estimator.

Remark 5.1.1. To estimate $f_x(M_x)$ by the generalized nearest neighbour density estimator related to the kernel estimator, the following procedure has been adopted:

Choose an integer $h \approx n^{1/2}$ and define the distance $d(x_1, x_2)$ between two points on the line to be$|x_1 - x_2|$.

For $\hat{M}_{x(n)}$ define $d_1(\hat{M}_{x(n)}) \le d_2(\hat{M}_{x(n)}) \le --- \le d_n(\hat{M}_{x(n)})$ to be the distances, arranged in ascending order, from $\hat{M}_{x(n)}$ to the points of the sample.

The generalized nearest neighbour density estimate is defined by

$$\hat{f}(\hat{M}_{x(n)}) = \frac{1}{nd_h(\hat{M}_{x(n)})} \sum_{i=1}^{n} K\left(\frac{\hat{M}_{x(n)} - x_i}{d_h(\hat{M}_{x(n)})}\right) \tag{20}$$

where the kernel function *K*, satisfies the condition $\int_{-\infty}^{\infty} K(x)\, dx = 1$.

Here, the kernel function is chosen as Gaussian Kernel given by $K(x) = \frac{1}{2\pi} e^{-\left(\frac{1}{2}x^2\right)}$.

Similarly, the estimate of $f_y(M_y)$ and $f_z(M_z)$ can be obtained.

Remark 5.1.2. For estimating $f_y\left(M_y\right)$, P_{yz} and P_{yx} we have two independent samples of sizes *u* and *m* respectively at current occasion. So, either of the two can be used, but in general for good sampling design in successive sampling $u \le m$. So, in the present work $f_y\left(M_y\right)$, P_{yz} and P_{yx} are estimated from the sample of size *m*, matched from the first occasion.

Therefore, under the above substitutions of the unknown population parameters by their respective sample estimates, the estimator *T* takes the following form:

$$T^* = \psi_1 T_1^* + \left(1 - \psi_1\right) T_2^* \tag{21}$$

where
$$T_1^* = \hat{M}_{y(u)} + k_1^{**}\left(\hat{M}_{z(u)} - M_Z\right) \tag{22}$$

and

$$T_2^* = \left\{\hat{M}_{y(m)} + k_2^{**}\left(\hat{M}_{x(m)} - \hat{M}_{x(n)}\right) + k_3^{**}\left(\hat{M}_{z(m)} - M_Z\right) + k_4^{**}\left(\hat{M}_{z(n)} - M_Z\right)\right\} \tag{23}$$

$$k_1^{**} = \frac{-A_3^*}{A_2^*},\quad k_2^{**} = \frac{A_3^* A_4^* A_6^* - A_2^* A_4^* A_5^*}{A_4^*\left(A_2^* A_4^* - A_6^{*2}\right)},\ k_3^{**} = \frac{-A_3^* A_4^* + A_5^* A_6^*}{\left(A_2^* A_4^* - A_6^{*2}\right)},$$

$$k_4^{**} = \frac{A_3^* A_6^{*2} - A_2^* A_5^* A_6^*}{A_2^*\left(A_2^* A_4^* - A_6^{*2}\right)},\ A_1^* = \frac{1}{4}\left\{\hat{f}_y\left(\hat{M}_{y(m)}\right)\right\}^{-2},\qquad A_2^* = \frac{1}{4}\left\{\hat{f}_z\left(\hat{M}_{z(n)}\right)\right\}^{-2},$$

$$A_3^* = \left(\hat{P}_{yz} - 0\cdot 25\right)\left\{\hat{f}_y\left(\hat{M}_{y(m)}\right)\right\}^{-1}\left\{\hat{f}_z\left(\hat{M}_{z(n)}\right)\right\}^{-1},\qquad A_4^* = \frac{1}{4}\left\{\hat{f}_x\left(\hat{M}_{x(n)}\right)\right\}^{-2},$$

$$A_5^* = \left(\hat{P}_{yx} - 0\cdot 25\right)\left\{\hat{f}_y\left(\hat{M}_{y(m)}\right)\right\}^{-1}\left\{\hat{f}_x\left(\hat{M}_{x(n)}\right)\right\}^{-1} \text{ and}$$

$$A_6^* = \left(\hat{P}_{xz} - 0\cdot 25\right)\left\{\hat{f}_x\left(\hat{M}_{x(n)}\right)\right\}^{-1}\left\{\hat{f}_z\left(\hat{M}_{z(n)}\right)\right\}^{-1}.$$

ψ_1 is an unknown constant to be determined so as to minimize the mean square error of the estimator T^*.

Remark 5.1.3. The proposed estimator *T* is a difference-type estimator therefore after replacing the unknown population parameters by their respective sample estimates it becomes a regression-type estimator. Hence, up to the first order of approximations the estimator T^* will be equally precise to that of the estimator *T* (see Singh and Priyanka (2008)). Therefore, similar conclusions are applicable for T^* as that of *T*.

6. Optimum replacement policy

To determine the optimum value of μ (fraction of a sample to be taken afresh at second occasion) so that M_y may be estimated with maximum precision, we minimize $V\left(T\right)_{opt\cdot}$ in equation (19) with respect to μ and hence we get the optimum value of μ as

$$\mu_{opt.^*} = \frac{-S_2 \ \pm \ \sqrt{S_2^2 \ - \ S_1 S_3}}{S_1} = \mu_0 \text{ (say)} \tag{24}$$

where $S_1 = A_{12}^2$, $S_2 = A_{11}A_{12}$ and $S_3 = A_{11}A_{13} \ - \ A_7 A_{12}$.

From equation (24) it is obvious that the real value of $\mu_{opt\cdot}$ exists if $S_2^2 - S_1 S_3 \geq 0$. For certain situation, there might be two values of $\mu_{opt\cdot}$ satisfying the above condition, hence to choose a value of $\mu_{opt\cdot}$, it should be remembered that $0 \leq \mu_{opt\cdot} \leq 1$. All other values of $\mu_{opt\cdot}$ are inadmissible. In case both the values of $\mu_{opt\cdot}$ are admissible, we choose the minimum of these two as μ_0. Substituting the value of $\mu_{opt\cdot}$ from equation (24) in (19) we have

$$V\left(T\right)_{opt.^*} = \frac{1}{n}\frac{A_7\left(A_{11} + \ \mu_0 A_{12}\right)}{\left(\mu_0^2 A_{12} + \ \mu_0 A_{13} + \ A_7\right)} \tag{25}$$

where $V\left(T\right)_{opt.^*}$ is the optimum value of T with respect μ.

7. Efficiency comparison

To study the performance of the estimator T, the percent relative efficiencies of T with respect to (i) $\hat{M}_{y(n)}$, the natural estimator of M_y, when there is no matching, and (ii) the estimator Δ, when no additional auxiliary information is used at any occasion, have been computed for two natural population data. The estimator Δ is defined under the same circumstances as the estimator T, but in the absence of information on additional auxiliary variable z on both the occasions is proposed as

$$\Delta = \left\{\delta_1 \hat{M}_{y(u)} \ + \ \delta_2 \hat{M}_{y(m)}\right\} \ + \ \left\{\delta_3 \hat{M}_{x(m)} \ + \ \delta_4 \hat{M}_{x(n)}\right\} \tag{26}$$

where $\delta_i \left(i = 1, \ 2, \ 3, \ 4\right)$ are constants to be determined so that

(i) The estimator Δ becomes unbiased for M_y and

(ii) The variance of Δ attains the minimum.

For unbiasedness, the following conditions must hold:
$\left(\delta_1+\delta_2\right)=1$ and $\left(\delta_3+\delta_4\right)=0$.

Substituting $\delta_1=\phi_2$ and $\delta_3=\beta_2$ in equation (26), the estimator Δ takes the following form:

$$\Delta=\left\{\phi_2\hat{M}_{y(u)}\ +\ \left(1\ -\ \phi_2\right)\hat{M}_{y(m)}\right\}\ +\ \beta_2\left(\hat{M}_{x(m)}\ -\ \hat{M}_{x(n)}\right)$$

$$=\ \phi_2\hat{M}_{y(u)}+\ \left(1\ -\ \phi_2\right)\left\{\hat{M}_{y(m)}\ +\ k_5\left(\hat{M}_{x(m)}-\ \hat{M}_{x(n)}\right)\right\}$$

$$\Delta\ =\ \phi_2\Delta_1+\left(1\ -\ \phi_2\right)\Delta_2 \tag{27}$$

where the estimator $\Delta_1=\hat{M}_{y(u)}$ is based on the fresh sample of size *u* and the estimator $\Delta_2=\left\{\hat{M}_{y(m)}+k_5\left(\hat{M}_{x(m)}-\hat{M}_{x(n)}\right)\right\}$ is based on the matched sample of size *m*, $k_5=\frac{\beta_2}{\left(1-\phi_2\right)}$ and ϕ_2 are the unknown constants to be determined so as to minimize the variance of estimator Δ. Following the methods discussed in Sections 4, 5 and 6, the optimum value of k_5, $\mu_{1opt\cdot}$ (optimum value of fraction of the fresh sample for the estimator Δ), variance of $\hat{M}_{y(n)}$ and optimum variance of Δ ignoring the finite population corrections are given by

$$k_5^*=\frac{-A_5}{A_4} \tag{28}$$

$$\mu_{1opt.^*}=\frac{\text{-}A_1\ \pm\ \sqrt{A_1\left(A_1+\ A_{14}\right)}}{A_{14}}=\ \mu^*\left(say\right) \tag{29}$$

$$V\left(\hat{M}_{y(n)}\right)=\frac{1}{n}A_1 \tag{30}$$

$$V\left(\Delta\right)_{opt\cdot^*}=\frac{1}{n}\frac{A_1\left(\ A_1+\ \mu^*A_{14}\right)}{\left(\ \mu^{*2}A_{14}+\ A_1\right)} \tag{31}$$

where $A_{14}=\frac{-A_5^2}{A_4}$.

The optimum values of μ, μ_1 and percent relative efficiencies E_1 and E_2 of the estimator *T* with respect to the estimator $\hat{M}_{y(n)}$ and Δ are computed for two natural populations and results are shown in Tabe-2, where

$$E_1=\frac{V\left(\hat{M}_{y(n)}\right)}{V\left(T\right)_{opt^*\cdot}}\times 100 \text{ and } E_2=\frac{V\left(\Delta\right)_{opt^*\cdot}}{V\left(T\right)_{opt^*\cdot}}\times 100$$

7.1. Estimator Δ in practice

The main difficulty in using the proposed estimator Δ defined in equation (27) is the availability of k_5, as the optimum values of k_5 depends on the population parameters $P_{yx}, f_y\left(M_y\right)$ and $f_x\left(M_x\right)$. If these parameters are known, the estimator Δ can easily be implemented, otherwise the unknown population parameters are replaced by their respective sample estimates as discussed in subsection 5.1. Hence, in this scenario the estimator Δ takes the following form:

$$\Delta^* = \psi_2 \Delta_1 + \left(1 - \psi_2\right)\Delta_2^* \tag{32}$$

where $\Delta_2^* = \left\{\hat{M}_{y(m)} + k_5^{**}\left(\hat{M}_{x(m)} - \hat{M}_{x(n)}\right)\right\}$, $k_5^{**} = \frac{-A_5^*}{A_4^*}$ and ψ_2 is the unknown constants to be determined so as to minimize the mean square error of the estimator Δ^*.

Remark 7.1.1. Since Δ^* is a regression-type estimator corresponding to the difference-type estimator Δ, hence up to the first order of approximations similar conclusions are applicable to Δ^* as that of Δ (See Singh and Priyanka (2008)).

Remark 7.1.2. For simulation study the proposed estimators T^* and Δ^* are considered instead of the proposed estimators *T* and Δ, respectively.

8. Monte Carlo Simulation

Empirical validation can be carried out by Monte Carlo Simulation. Real life situations of completely known two finite populations have been considered.

Population Source: [Free access to data by Statistical Abstracts of the United States]

The first population comprise *N = 51* states of the United States. Let y_i represent the number of abortions during 2007 in the i^{th} state of the US, x_i be the number of abortions during 2005 in the i^{th} state of the U,S and z_i denote the number of abortions during 2004 in the i^{th} state of the US. The data are presented in Figure 1.

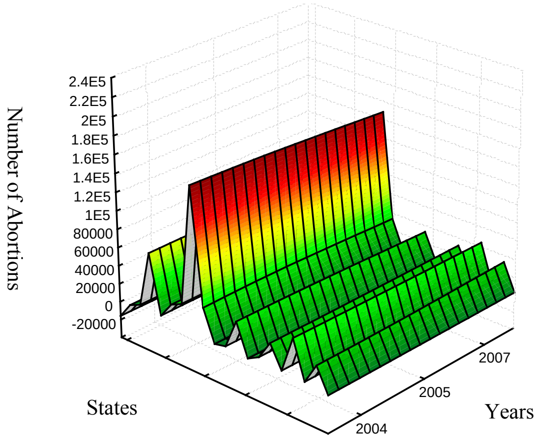

Figure 1. Number of abortions during 2004, 20 05 and 2007 versus different states of the US

Similarly, the second population consists of $N=41$ corn producing states of the United States. We assume y_i the production of corn (in million bushels) during 2009 in the i^{th} state of the US, x_i be the production of corn (in million bushels) during 2008 in the i^{th} state of the US and z_i denote the production of corn (in million bushels) during 2007 in the i^{th} state of the US. The data are represented by means of graph in Figure 2.

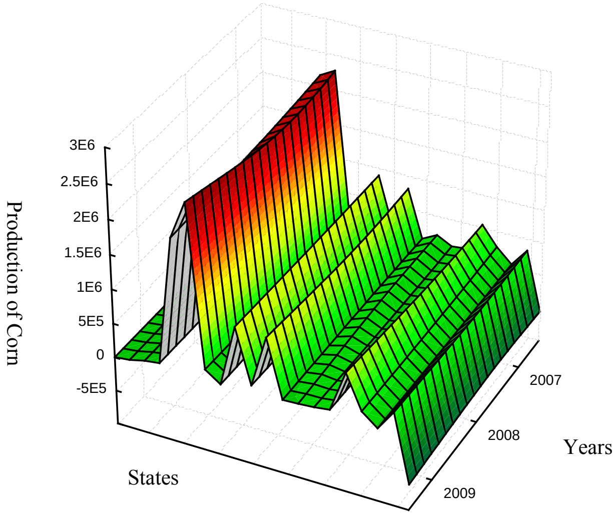

Figure 2. Production of corn during 2007, 2008 and 2009 versus different states of the US

The graphs in Figure1 and Figure 2 show that the number of abortions and the production of corn in different states are skewed towards right. One reason of skewness for the population-I may be the distribution of population in different states, that is the states having larger population are expected to have larger number of abortion cases. Similarly, for population-II the states having larger area for farming are expected to have larger production of corn. Thus, skewness of data indicates that the use of median may be a better measure of central location than mean in these situations.

For performing the Monte Carlo Simulation in the considered population-I, 5000 samples of *n=20* states were selected using simple random sampling without replacement in the year 2005. The sample medians $\hat{M}_{x(n)|k}$ and $\hat{M}_{z(n)|k}$, k =1, 2,---,5000 were computed and the parameters $f_x(M_x)$, $f_z(M_z)$ and P_{xz} were estimated by the method given in Remark 5.1.1. From each one of the selected samples, *m=17* states were retained and new *u=3* states were selected out of *N – n =51 – 20 = 31* states using simple random sampling without replacement in the year 2007. From the m units retained in the sample at the current occasion, the sample medians $\hat{M}_{x(m)|k}$, $\hat{M}_{y(m)|k}$ and $\hat{M}_{z(m)|k}$, *k = 1, 2,- - -,5000* were computed and the parameters $f_y(M_y)$, P_{yz} and P_{xz} were estimated. From the new unmatched units selected on the current occasion the sample medians $\hat{M}_{y(u)|k}$ and $\hat{M}_{z(u)|k}$, *k = 1, 2,- - -,5000* were computed. The parameters ψ_1 and ψ_2 are selected between 0.1 and 0.9 with a step of 0.1.

The percent relative efficiencies of the proposed estimator T^* with respect to $\hat{M}_{y(n)}$ and Δ^* are respectively given by:

$$E_{1sim} = \frac{\sum_{k=1}^{5000}\left[\hat{M}_{y(n)|k} - M_y\right]^2}{\sum_{k=1}^{5000}\left[T_k^* - M_y\right]^2}\times 100 \quad \text{and} \quad E_{2sim} = \frac{\sum_{k=1}^{5000}\left[\Delta_k^* - M_y\right]^2}{\sum_{k=1}^{5000}\left[T_k^* - M_y\right]^2}\times 100$$

For better analysis, this simulation experiments were repeated for different choices of μ.

Similar steps are also followed for Population-II. The simulation results in Table 3, Table 4 and Table 5 show the comparison of the proposed estimator T^* with respect to the estimators $\hat{M}_{y(n)}$ and Δ^*, respectively. For convenience the different choices of μ are considered as different sets for the considered Population-I and Population-II, which are shown below:

Sets	Population-I	Population-II
I	*n =20; μ = 0.15 (m =17, u =3)*	*n=15; μ = 0.13 (m =13, u =2)*
II	*n =20; μ = 0.25 (m = 15, u =5)*	*n=15; μ = 0.20 (m =12, u =3)*
III	*n =20; μ = 0.35 (m = 13, u =7)*	*n=15; μ = 0.30 (m = 10, u =5)*
IV	*n =20; μ = 0.50 (m = 10, u =10)*	*n=15; μ = 0.40 (m = 9, u =6)*

Table 1. Descriptive statistics for Population-I and Population-II

	Population-I			**Population-II**		
	Abortions 2004 (z)	**Abortions 2005 (x)**	**Abortions 2007 (y)**	**Production of Corn in 2007 (z)**	**Production of Corn in 2008 (x)**	**Production of Corn in 2009 (y)**
Mean	23963.14	23651.76	23697.65	317997	294918.2	319313.7
Median	11010.00	10410.00	9600.00	83740	66650	79730
Standard Deviation	38894.81	38487.71	39354.65	565641.6	530483.7	563103.3
Kurtosis	12.02669	12.39229	14.42803	6.838888	6.492807	6.036604
Skewness	3.275197	3.310767	3.527683	2.638611	2.595704	2.499771
Minimum	80	70	90	2997	2475	2635
Maximum	208180	208430	223180	2376900	2188800	2420600
Count	51	51	51	41	41	41

Table 2. Comparison of the proposed estimator *T* (at optimal conditions) with respect to the estimators $\hat{M}_{y(n)}$ and Δ (at optimal conditions)

	Population - I	**Population-II**
μ_0	0.5411	0.6669
μ^*	0.6800	0.7642
E_1	1407.5	1401.3
E_2	1034.9	916.80

Table 3. Monte Carlo Simulation results when the proposed estimator T^* is compared to $\hat{M}_{y(n)}$ for Population-I and Population-II

	Population-I				**Population-II**			
Set	**I**	**II**	**III**	**IV**	**I**	**II**	**III**	**IV**
ψ_1 ↓	E_{1sim}	E_{1sim}	E_{1sim}	E_{1sim}	E_{1sim}	E_{1sim}	E_{1sim}	E_{1sim}
0.1	338.42	285.75	294.74	191.46	762.21	747.03	127.19	321.48
0.2	330.71	291.82	320.22	238.4	860.29	644.25	140.93	364.51
0.3	315.85	288.81	333.44	254.30	971.34	536.15	154.84	397.27
0.4	282.71	288.70	326.08	276.75	1097.6	427.33	166.51	420.99
0.5	248.64	268.90	322.70	295.47	1219.7	340.46	172.53	413.40
0.6	210.41	249.90	299.55	301.46	1377.0	262.76	175.98	413.49
0.7	178.81	220.94	269.87	304.12	1529.3	206.40	172.93	398.24
0.8	152.05	194.11	245.61	297.46	1707.7	166.72	166.51	369.96
0.9	127.19	168.82	216.58	289.94	1855.9	136.86	161.50	336.32

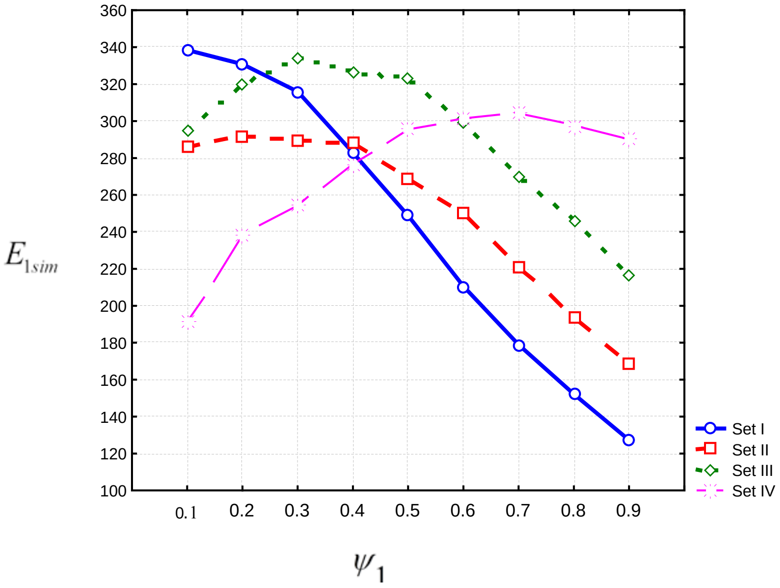

Figure 3. PRE of the estimator T^* with respect to $\hat{M}_{y(n)}$ for Population-I

Table 4. Monte Carlo Simulation results for Population-I when the proposed estimator T^* is compared to Δ^*

$\psi_1 \downarrow$	$\psi_2 \rightarrow$		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	E_{2sim}	I	329.1	470.4	707.2	1017.2	1590.3	2211.0	2869.2	4255.0	5490.3
		II	269.4	272.6	291.4	424.8	681.0	752.7	1023.3	1511.8	1790.9
		III	285.6	233.2	273.0	320.1	430.9	624.4	770.1	1126.7	1353.6
		IV	205.2	188.5	168.7	168.4	198.1	230.3	318.0	419.5	559.2
0.2	E_{2sim}	I	340.3	456.3	714.2	1078.2	1685.3	2268.1	3064.6	4227.3	5437.1
		II	285.8	282.7	312.6	461.3	678.1	824.9	1150.8	1600.8	2034.9
		III	295.9	251.1	279.7	344.3	457.5	636.8	831.4	1126.8	1428.8
		IV	242.3	199.2	177.2	182.9	222.9	269.7	351.5	483.4	631.6
0.3	E_{2sim}	I	325.9	440.9	688.6	1071.6	1547.1	2158.4	2979.3	4060.1	5145.1
		II	288.6	285.4	336.3	475.3	677.2	839.5	1187.6	1643.4	1983.4
		III	298.7	264.8	287.5	358.9	456.2	642.1	852.9	1159.3	1466.2
		IV	261.4	216.4	192.2	198.1	247.3	294.9	391.5	529.6	681.6
0.4	E_{2sim}	I	298.2	411.3	624.7	967.3	1430.2	1975.9	2648.7	3594.8	4721.6
		II	284.9	282.3	329.8	454.1	659.4	842.4	1152.1	1600.3	1946.5
		III	289.6	265.6	284.4	341.2	460.3	635.6	857.8	1142.6	1440.9
		IV	279.6	231.6	204.9	212.9	263.5	314.2	419.5	559.7	739.3
0.5	E_{2sim}	I	262.6	358.2	548.2	883.8	1247.1	1709.9	2238.4	3128.2	4213.1
		II	266.7	263.7	312.7	430.3	620.7	789.8	1072.8	1468.6	1775.0
		III	274.8	251.4	270.1	327.9	442.0	616.1	820.8	1111.1	1404.6
		IV	296.9	246.8	219.2	222.8	273.9	331.8	440.8	586.7	765.7
0.6	E_{2sim}	I	230.1	310.8	463.6	754.2	1078.0	1509.3	2016.2	2669.3	3583.8
		II	248.8	244.8	283.3	403.9	565.8	730.9	1004.8	1336.5	1673.8
		III	249.3	238.5	253.4	314.6	412.2	574.3	775.3	1016.9	1336.2
		IV	303.9	256.0	226.1	231.7	283.7	343.1	456.8	600.3	783.1
0.7	E_{2sim}	I	194.5	257.1	396.7	625.2	920.4	1275.6	1753.0	2249.7	2955.3
		II	226.0	216.7	252.9	352.7	512.4	656.3	907.6	1182.0	1473.9
		III	226.1	214.6	226.1	285.9	382.3	532.1	706.8	898.9	1208.2
		IV	305.8	258.3	227.1	235.5	284.2	346.9	459.8	599.8	788.4
0.8	E_{2sim}	I	159.8	221.7	341.1	523.4	757.4	1095.9	1515.0	1960.0	2478.9
		II	193.4	190.9	228.7	320.2	438.1	580.6	825.6	1037.5	1328.2
		III	201.6	194.7	205.2	265.1	347.7	481.8	628.9	800.2	1082.0
		IV	299.9	256.9	223.5	233.7	283.7	341.6	453.7	589.5	772.5
0.9	E_{2sim}	I	136.5	186.4	289.7	440.6	635.9	939.3	1269.8	1663.2	2125.0
		II	172.9	165.9	202.6	288.7	373.1	514.3	709.8	894.3	1160.4
		III	182.2	167.1	185.0	234.8	309.8	418.6	552.9	722.3	930.8
		IV	293.8	245.8	216.8	225.3	272.8	329.7	438.3	574.2	742.7

Table 5. Monte Carlo Simulation results for Population-II when the proposed estimator T^* is compared to Δ^*

$\psi_1 \downarrow$	$\psi_2 \rightarrow$		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	E_{2sim}	I	1126.40	2860.5	5849.0	9978.9	14402.0	22607.0	30230.0	40853.0	46469.0
		II	961.19	1757.9	3077.6	5323.8	7930.8	11637.0	14805.0	20847.0	26905.0
		III	274.83	264.72	298.76	362.76	515.77	742.68	1006.7	1174.6	1320.8
		IV	448.87	445.82	537.81	641.19	1000.5	1320.8	1757.2	2256.2	3038.8
0.2	E_{2sim}	I	873.59	2198.3	4489.6	7729.9	11800.0	17466.0	22954.0	31590.0	3644.3
		II	831.99	1472.2	2545.2	4305.6	6678.7	9960.1	13156.0	17250.0	23024.0
		III	302.79	284.98	314.11	406.01	562.11	821.52	995.42	1259.0	1522.1
		IV	495.59	481.24	567.79	708.65	1010.5	1426.0	1852.1	2354.0	3098.0
0.3	E_{2sim}	I	621.89	1594.20	3184.1	5627.4	8573.0	12582.0	16513.0	22385.0	27277.0
		II	682.77	1169.0	2044.1	3405.3	5386.4	7770.3	10373.0	13378.0	17978.0
		III	328.74	312.90	338.97	448.28	617.43	89.51	1079.6	1333.3	1719.8
		IV	528.81	521.64	667.01	761.28	1069.9	1502.1	1953.7	2645.4	3251.4
0.4	E_{2sim}	I	441.33	1136.90	2342.9	4039.8	6230.6	8970.8	11971.0	16010.0	20221.0
		II	540.36	905.32	1585.1	2637.0	4066.8	5938.0	8098.8	10354.0	13708.0
		III	349.27	334.32	366.96	469.80	658.16	909.27	1131.5	1455.1	1817.1
		IV	557.80	535.90	625.09	792.63	1111.7	1534.2	2022.3	2703.7	3360.2
0.5	E_{2sim}	I	325.32	829.35	1693.8	2954.8	4550.0	6503.2	8647.7	11725.0	14875.0
		II	423.09	685.55	1205.1	2062.0	3128.3	4491.7	6008.1	7843.8	10477.0
		III	358.42	347.77	382.11	498.04	683.40	938.99	1172.6	1524.7	1908.0
		IV	552.30	537.56	627.89	796.60	1104.7	1536.0	2036.20	2690.1	3371.6
0.6	E_{2sim}	I	247.94	628.85	1282.4	2233.8	3406.2	4921.7	6612.4	8869.5	11284.0
		II	326.45	531.46	954.37	1614.8	2416.2	3449.1	4720.8	6152.4	8021.9
		III	369.80	356.29	390.36	507.65	697.08	953.09	1193.9	1553.5	1966.7
		IV	545.08	519.34	607.57	778.51	1081.1	1486.7	1976.3	2607.6	3256.7
0.7	E_{2sim}	I	191.82	481.70	989.78	1738.2	2659.8	3832.4	5161.5	6844.7	8705.7
		II	256.24	421.16	747.44	1246.6	1864.4	2796.1	3789.1	4836.2	6404.1
		III	368.09	357.34	391.04	507.07	692.18	943.99	1198.0	1548.7	1972.1
		IV	523.74	448.94	569.41	738.38	1020.9	1405.1	1886.9	2452.8	3067.3
0.8	E_{2sim}	I	154.29	383.89	790.48	1385.5	2112.4	3041.20	4114.9	5376.9	6949.5
		II	206.36	335.56	604.62	1004.1	1507.5	2283.7	3062.3	3868.2	5119.9
		III	361.45	347.49	391.04	490.64	667.61	915.93	1161.0	1510.2	1915.8
		IV	488.89	463.14	526.20	689.27	941.81	1304.0	1735.1	2254.4	2837.2
0.9	E_{2sim}	I	124.89	310.43	635.21	1100.2	1714.1	2458.4	3302.5	4362.3	5601.2
		II	169.07	271.88	498.12	826.69	1245.4	1855.6	2493.5	3169.4	4211.6
		III	346.69	330.68	379.63	469.72	629.28	869.77	1114.2	1438.0	1843.1
		IV	445.87	413.45	477.73	615.16	848.82	1179.9	1569.1	2032.7	2622.9

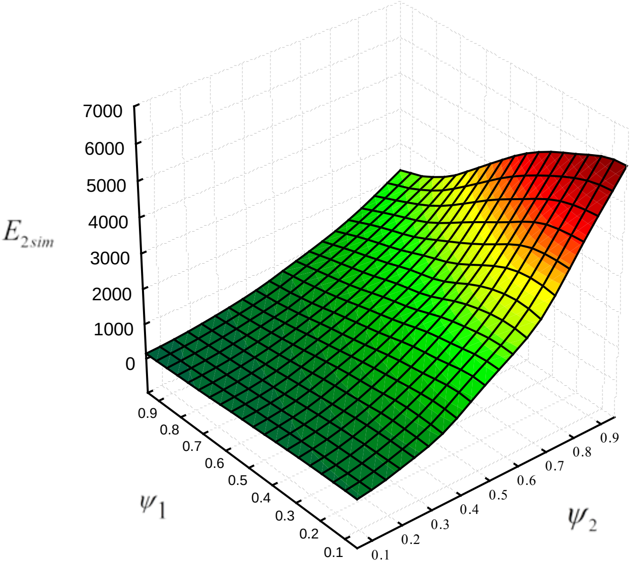

Figure 4. PRE of estimator T^* with respect to Δ^* for set-I for Population-I

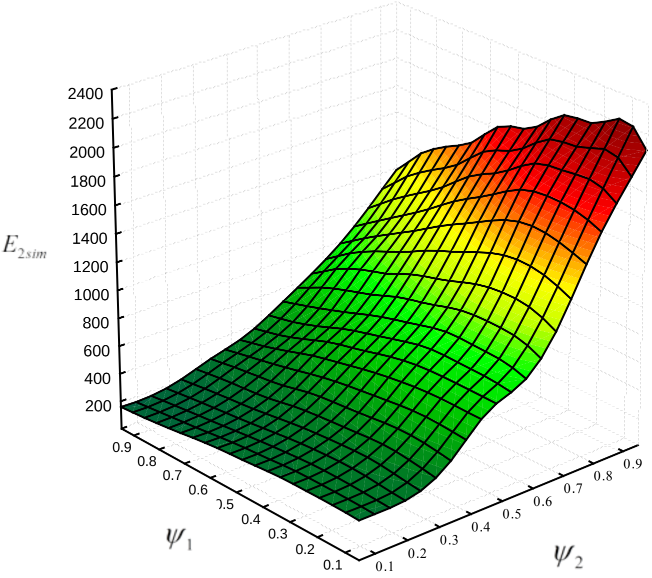

Figure 5. PRE of estimator T^* with respect to ψ_1 for set-II for Population-I

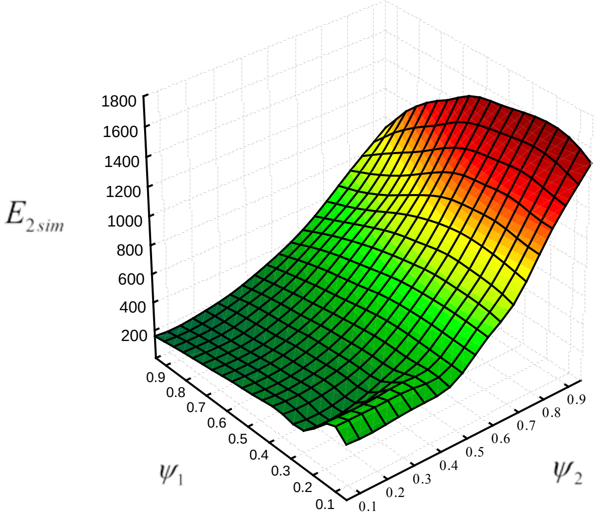

Figure 6. PRE of estimator T^* with respect to ψ_1 for set-III for Population-I

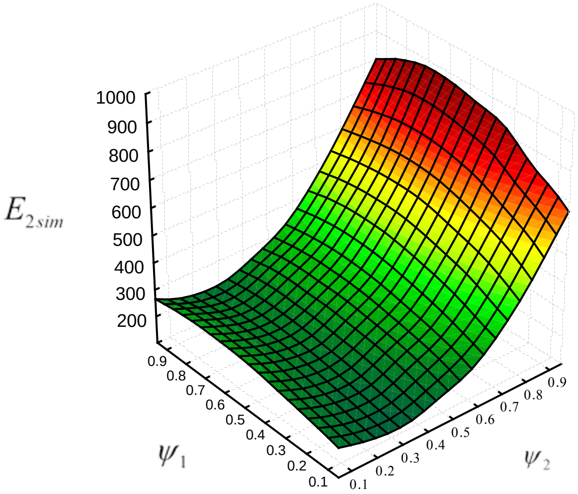

Figure 7. PRE of estimator T^* with respect to ψ_1 for set-IV for Population-I

9. Analysis of empirical and simulation results

1. From table 2 it is visible that the optimum values of μ (fraction of a fresh sample to be drawn at current occasion) exist and this value for the estimator T is less than that of the estimator Δ for both the considered populations. This indicates that the use of additional auxiliary information at both the occasion reduces the cost of the survey.

2. Appreciable gain is observed in terms of precision indicating the proposed estimator T (at optimal condition) preferable over the estimators $\hat{M}_{y(n)}$ and Δ (at optimal condition). This result justifies the use of additional auxiliary information at both the occasions in two-occasion successive sampling.

3. The following conclusion may be observed from Table 3 and Figure 3:

   (i) For Set-I of Population-I, the value of E_{1sim} decreases as the value of ψ_1 increases. This result is expected as for Set-I the value of μ is very low, however for Set-I of Population-II E_{1sim} increases with the increasing value of ψ_1.

   (ii) For Set-II, III and IV of the Population-I, the value of E_{1sim} first increases and then starts decreasing with the increasing value of ψ_1, however no specific pattern is observed for set II, III and IV of Population-II.

   (iii) For all the considered combinations appreciable gain in precision is observed when the proposed estimator is compared with the sample median estimator. Hence, the use of additional auxiliary information at both the occasions is highly justified.

4. The following points may be noted from Table 4, Table 5 and Figures 4, 5, 6 and 7:

   (i) For fixed value of ψ_1 and ψ_2, the value of E_{2sim} decreases with the increasing value of μ, except for few combinations of ψ_1 and ψ_2 for Population-I, however no specific pattern is observed for Population-II.

   (ii) For fixed value of ψ_1 and μ and increasing value of ψ_2, the value of E_{2sim} also increases, except for few combinations.

   (iii) For fixed value of ψ_2, and lower value of μ, the value of E_{2sim} decreases with increasing value of ψ_1, however for higher value of μ, the value of E_{2sim} increases with the increasing value of Ψ_1, except for few combinations.

   (iv) Tremendous gain in precision is obtained for all the considered cases.

10. Conclusion

From the analysis of empirical and simulation results it can be concluded that the proposed estimator T compares favourably in terms of efficiency with the standard sample median estimator, where there is no matching from previous occasion. The estimator T also proves to be much better than the estimator Δ, when no additional auxiliary information is used at any occasion. Therefore, the use of additional auxiliary information at both the occasions in two occasion successive sampling for estimating population median at current occasion is highly rewarding in terms of precision and reducing the total cost of survey. Hence, the proposed estimators may be recommended for further use by survey practitioners.

Acknowledgements

Authors are thankful to the anonymous referees for their precious and constructive suggestions regarding improvement of the article. Authors appreciate the free access and acknowledge the use of data from Statistical Abstracts of the United States, 2012 that is available on the Internet. Authors are also thankful to UGC, New Delhi, India for providing the financial assistance to carry out the present work.

REFERENCES

ARNAB, R., OKAFOR, F. C., (1992). A note on double sampling over two occasions. Pakistan JNL of statistics 8, 9–18.

BANDYOPADHYAY, A., SINGH, G. N., (2014). On the use of two auxiliary variables to improve the precision of estimate in two-occasion successive sampling. International Journal of Mathematics and Statistics15(1): 73–88.

CHAMBERS, R. L., DUNSTAN, R., (1986). Estimating distribution functions from survey data. Biometrika 73, 597–604.

ECKLER, A. R., (1955). Rotation Sampling. Ann. Math. Statist.: 664–685.

FENG, S., ZOU, G., (1997). Sample rotation method with auxiliary variable. Commun. Statist. Theo-Meth.26: 6, 1497–1509.

GORDON, L., (1983). Successive sampling in finite populations. The Annals of statistics 11(2): 702–706.

GROSS, S. T., (1980). Median estimation in sample surveys. Proc. Surv. Res. Meth. Sect. Amer. Statist. Assoc.: 181–184.

GUPTA, S., SHABBIR, J., AHMAD, S., (2008). Estimation of median in two phase sampling using two auxiliary variables. Communications in Statistics - Theory & Methods 37(11): 1815–1822.

JESSEN, R. J., (1942). Statistical investigation of a sample survey for obtaining farm facts. Iowa Agricultural Experiment Station Road Bulletin No. 304, Ames: 1–104.

KHOSHNEVISAN, M., SAXENA, S., SINGH, H. P., SINGH, S., SMARANDACHE, F., (2002). Randomness and optimal estimation in data sampling. American Research Press, Second Edition, Rehoboth.

KUK, A. Y. C., MAK, T. K., (1989). Median estimation in presence of auxiliary information. J. R. Statit. Soc. B, 51: 261–269.

MARTINEZ-MIRANDA, M. D., RUEDA-GARCIA, M., ARCOS-CEBRIAN, A., ROMAN-MONTOYA, Y., GONZAEZ-AGUILERA, S., (2005). Quintile estimation under successive sampling. Computational Statistics, 20:385–399.

NARAIN, R. D., (1953). On the recurrence formula in sampling on successive occasions. Journal of the Indian Society of Agricultural Statistics 5: 96–99.

PATTERSON, H. D., (1950). Sampling on successive occasions with partial replacement of units. Jour. Royal Statist. Assoc., Ser. B, 12: 241–255.

RAO, J. N. K., KOVAR, J. G., MANTEL, H. J., (1990). On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika, 77: 2, 365–375.

RUEDA, M. D. M., ARCOS, A., ARTES, E., (1998). Quantile interval estimation in finite population using a multivariate ratio estimator. Metrika 47: 203–213.

RUEDA, M. D. M., MUNOZ, J. F., (2008). Successive sampling to estimate quantiles with P-Auxiliary Variables. Quality and Quantity, 42:427–443.

SEDRANSK, J., MEYER, J., (1978). Confidence intervals for quantiles of a finite population: Simple random and stratified simple random sampling. J. R. Statist. Soc., B, 40: 239–252.

SILVERMAN, B. W., (1986). Density Estimation for Statistics and Data Analysis, Chapman and Hall, London.

SINGH, G. N., SINGH, V. K., (2001). On the use of auxiliary information in successive sampling. Jour. Ind. Soc. Agri. Statist. 54(1): 1–12.

SINGH, G. N., PRIYANKA, K., (2008). Search of good rotation patterns to improve the precision of estimates at current occasion. Communications in Statistics (Theory and Methods) 37(3), 337–348.

SINGH, G. N., PRASAD, S., MAJHI, D., (2012). Best Linear Unbiased Estimators of Population Variance in Successive Sampling. Model Assisted Statistics and Applications, 7, 169–178.

SINGH, H. P., TAILOR, R., SINGH, S., JONG-MIN KIM, (2007). Quintile estimation in successive sampling, Journal of the Korean Statistical Society, 36: 4, 543–556.

SINGH, H. P., SOLANKI, R. S., (2013). Some Classes of estimators for population median using auxiliary information. Communications in Statistics - Theory & Methods, 42, (23), 4222–4238.

SINGH, S., (2003). Advanced Sampling Theory with Applications; How Michael 'selected' Amy. (Vol. 1 and 2) pp. 1–1247, Kluwer Academic Publishers, The Netherlands.

SMITH, P., SEDRANSK, J., (1983). Lower bounds for confidence coefficients for confidence intervals for finite population quantiles. Communications in Statistics - Theory & Methods, 12: 1329–1344.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 221–230

SOME CHAIN-TYPE EXPONENTIAL ESTIMATORS OF POPULATION MEAN IN TWO-PHASE SAMPLING

G. N. Singh[1], D. Majhi[2]

ABSTRACT

Using the information on two-auxiliary variables, three different exponential chain-type estimators of population mean of study variable have been proposed in two-phase (double) sampling. Properties of the proposed estimators have been studied and their performances are examined with respect to several well known chain-type estimators. Empirical studies are carried out to support the theoretical results.

Key words: two-phase, auxiliary information, bias, mean square error.
Mathematics subject classification: 62D05

1. Introduction

Ratio, product and regression methods of estimation require the knowledge of population mean of the auxiliary variable. If population mean of the auxiliary variable is not known, it is customary to move towards two-phase sampling scheme, which provides a cost effective estimate of the unknown population mean of auxiliary variable in first-phase sample. Utilizing the information on known population mean of another auxiliary variable in first-phase sample, Chand (1975) introduced chain-type ratio estimator of population mean of study variable. His work was further extended by Kiregyera (1980, 1984), Mukherjee *et al.* (1987), Srivastava *et al.* (1989), Upadhyaya *et al.* (1990), Singh and Singh (1991), Singh *et al.* (1994), Singh and Upadhyaya (1995), Upadhyaya and Singh (2001), Singh (2001), Pradhan (2005), Gupta and Shabbir (2008) and Singh *et al.* (2011) among others. Motivated with the above works, the aim of the present research is to propose some different structures of chain-type estimators in two-phase sampling which may estimate the population mean in a more precise way in comparison with the contemporary estimators of similar kind.

[1] Department of Applied Mathematics, Indian School of Mines, Dhanbad-826004, India. E-mail: gnsingh_ism@yahoo.com.
[2] Department of Applied Mathematics, Indian School of Mines, Dhanbad-826004, India.

2. Two-phase sampling set-up

Consider a finite population U of size N indexed by triplet characters (y, x, z). We wish to estimate the population mean $\bar{Y}$ of study variable y in the presence of two auxiliary variables x and z. Let x and z be called first and second auxiliary variables respectively such that y is highly correlated with x while in comparison with x it is remotely correlated with z $\left(\text{i.e. } \rho_{yx} > \rho_{yz}\right)$.When the population mean $\bar{X}$ of x is unknown but information on z is available on all the units of the population, we use the following two-phase sampling scheme.

Let us now consider a two-phase sampling where in the first phase a large (preliminary) sample $s'\left(s' \subset U\right)$ of fixed size n' is drawn following SRSWOR to observe two auxiliary variables x and z to estimate $\bar{X}$, while in the second phase a sub-sample $s \subset s'$ of fixed size n is drawn by SRSWOR to observe the characteristic y under study.

3. Estimators based on one auxiliary variable

Ratio and regression estimators in two-phase sampling are the traditional estimators utilizing the information on one auxiliary variable and are reproduced below along with their respective mean square errors up to $o\left(n^{-1}\right)$.

$$\bar{y}_{rd} = \frac{\bar{y}}{\bar{x}}\bar{x}' \tag{1}$$

$$M\left(\bar{y}_{rd}\right) = \bar{Y}^2\left[f_1C_y^2 + f_3\left(C_x^2 - 2\rho_{yx}C_yC_x\right)\right] \tag{2}$$

$$\bar{y}_{lrd} = \bar{y} + b_{yx}\left(n\right)\left(\bar{x}' - \bar{x}\right) \tag{3}$$

$$M\left(\bar{y}_{lrd}\right) = S_y^2\left[f_1\left(1-\rho_{yx}^2\right) + f_2\rho_{yx}^2\right] \tag{4}$$

where $b_{yx}\left(n\right)$ is the sample regression coefficient of y on x calculated from the data based on s and

$\bar{y} = \frac{1}{n}\sum_{i\in s} y_i$, $\bar{x} = \frac{1}{n}\sum_{i\in s} x_i$ and $\bar{x}' = \frac{1}{n'}\sum_{i\in s'} x_i$, $f_1 = \left(\frac{1}{n} - \frac{1}{N}\right)$, $f_2 = \left(\frac{1}{n'} - \frac{1}{N}\right)$, $f_3 = \left(f_1 - f_2\right) = \left(\frac{1}{n} - \frac{1}{n'}\right)$

$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2$, $S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$, $C_x = \frac{S_x}{\bar{X}}$, $C_y = \frac{S_y}{\bar{Y}}$, and ρ_{yx} be the correlation coefficient between the variables y and x. $\bar{X}$ and $\bar{Y}$ are the population means of the variables x and y, respectively.

4. Estimators based on two auxiliary variables

Chand (1975) introduced a chain-type ratio estimator under two-phase sampling using two auxiliary variables x and z when the population mean $\bar{X}$ of x is unknown but information on z is available on all the units of the population, which is given as

$$\bar{y}_{rc}=\frac{\bar{y}}{\bar{x}}\frac{\bar{x}'}{\bar{z}'}\bar{Z} \tag{5}$$

The mean square error of the estimator $\bar{y}_{rc}$ up to $o\left(n^{-1}\right)$ is derived as

$$M\left(\bar{y}_{rc}\right)=\bar{Y}^2\left[f_1C_y^2+f_3\left(C_x^2-2\rho_{yx}C_yC_x\right)+f_2\left(C_z^2-2\rho_{yz}C_yC_z\right)\right] \tag{6}$$

where $\bar{Z}$ is the population mean of the variable z, $\bar{z}'=\frac{1}{n'}\sum_{i\in s'}z_i$, $C_z=\frac{S_z}{\bar{Z}}$, $S_z^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(z_i-\bar{Z}\right)^2$ and ρ_{yz} be correlation coefficient between variables y and z.

Kiregyera (1980, 1984) extended the work of Chand (1975) and suggested chain-type ratio to regression, regression to ratio and regression to regression estimators of population mean of study variable y in two-phase sampling which utilized the information on two auxiliary variables. The suggested estimators are given below along with their respective mean square errors up to $o\left(n^{-1}\right)$.

$$\bar{y}_{k1}=\frac{\bar{y}}{\bar{x}}\left[\bar{x}'+b_{xz}\left(n'\right)\left(\bar{Z}-\bar{z}'\right)\right] \tag{7}$$

$$M\left(\bar{y}_{k1}\right)==\bar{Y}^2\left[f_3\left(C_x^2+C_y^2-2\rho_{yx}C_yC_x\right)+f_2C_y^2+f_2\rho_{xz}C_x\left(\rho_{xz}C_x-2\rho_{yz}C_y\right)\right] \tag{8}$$

$$\bar{y}_{k2}=\bar{y}+b_{yx}(n)\left(\bar{x}'_{rd}-\bar{x}\right);\ \bar{x}'_{rd}=\frac{\bar{x}'}{\bar{z}'}\bar{Z} \tag{9}$$

$$M\left(\bar{y}_{k2}\right)=\bar{Y}^2C_y^2\left[f_1\left(1-\rho_{yx}^2\right)+f_2\left(\rho_{yx}^2+\rho_{yx}^2\frac{C_z^2}{C_x^2}-2\rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right)\right] \tag{10}$$

$$\bar{y}_{k3}=\bar{y}+b_{yx}(n)\left(\bar{x}'_{ld}-\bar{x}\right);\bar{x}'_{ld}=\left[\bar{x}'+b_{xz}\left(n'\right)\left(\bar{Z}-\bar{z}'\right)\right] \tag{11}$$

$$M\left(\bar{y}_{k3}\right)=\bar{Y}^2C_y^2\left[f_3\left(1-\rho_{yx}^2\right)+f_2\left(1+\rho_{xz}^2\rho_{yx}^2-2\rho_{yx}\rho_{yz}\rho_{xz}\right)\right] \tag{12}$$

where $b_{xz}\left(n'\right)$ is the sample regression coefficient of the variable x on z calculated from the data based on s' and ρ_{xz} be correlation coefficient between variables x and z.

5. Proposed estimators

The intelligible use of auxiliary information at estimation stage is a fascinating act in sample surveys. In presence of auxiliary information Bahl and Tuteja (1991) suggested an exponential structure to estimate the population mean of study variable and their work was extended by Singh and Vishwakarma (2007) in two-phase sampling scheme. Motivated with the work related to the proposition of chain-type estimators in two-phase sampling set-up, and looking on the nice behaviours of the exponential type estimators, we suggest below three different chain-types exponential estimators of population mean $\overline{Y}$ of the study variable y. The suggested estimators are given as

$$T_1=\frac{\overline{y}}{\overline{x}}\overline{x}'\exp\left(\frac{\overline{Z}-\overline{z}'}{\overline{Z}+\overline{z}'}\right) \tag{13}$$

$$T_2=\overline{y}+b_{yx}\left(n\right)\left\{\overline{x}'\exp\left(\frac{\overline{Z}-\overline{z}'}{\overline{Z}+\overline{z}'}\right)-\overline{x}\right\} \tag{14}$$

and

$$T_3=\overline{y}\exp\left(\frac{\overline{x}'-\overline{x}}{\overline{x}'+\overline{x}}\right)\left(\frac{\overline{Z}}{\overline{z}'}\right) \tag{15}$$

6. Properties of the estimators $T_i\left(i=1,2,3\right)$

Theorem 6.1. Biases of the estimators $T_i\left(i=1,2,3\right)$ defined in equations (13), (14) and (15) up to $o\left(n^{-1}\right)$ are obtained as

$$B\left(T_1\right)=\overline{Y}\left[f_3\left(C_x^2-\rho_{yx}C_yC_x\right)+\frac{f_2}{2}\left(\frac{3}{4}C_z^2-\rho_{yz}C_yC_z\right)\right] \tag{16}$$

$$B\left(T_2\right)=\beta_{yx}\left[f_3\left(\frac{\mu_{300}}{\mu_{200}}-\frac{\mu_{210}}{\mu_{110}}\right)+\frac{f_2}{2}\left(\frac{3}{4}\frac{\overline{X}}{\overline{Z}^2}\mu_{002}-\frac{1}{\overline{Z}}\mu_{101}+\frac{\overline{X}}{\overline{Z}}\frac{\mu_{201}}{\mu_{200}}-\frac{\overline{X}}{\overline{Z}}\frac{\mu_{111}}{\mu_{110}}\right)\right] \tag{17}$$

and

$$B\left(T_3\right)=\bar{Y}\left[\frac{f_3}{2}\left(\frac{3}{4}C_x^2-\rho_{yx}C_yC_x\right)+f_2\left(C_z^2-\rho_{yz}C_yC_z\right)\right] \tag{18}$$

where $\mu_{rst}=E\left[\left(x_i-\bar{X}\right)^r\left(y_i-\bar{Y}\right)^s\left(z_i-\bar{Z}\right)^t\right];\left(r,s,t\right)\geq 0$ are integers.

Theorem 6.2. Mean square errors of the estimators $T_i\left(i=1,2,3\right)$ defined in equations (13), (14) and (15) up to $o\left(n^{-1}\right)$ are derived as

$$M\left(T_1\right)=\bar{Y}^2\left[f_1C_y^2+f_3\left(C_x^2-2\rho_{yx}C_yC_x\right)+\frac{f_2}{4}\left(C_z^2-4\rho_{yz}C_yC_z\right)\right] \tag{19}$$

$$M\left(T_2\right)=\bar{Y}^2C_y^2\left[f_3\left(1-\rho_{yx}^2\right)+f_2\left(1+\frac{\rho_{yx}^2}{4}\frac{C_z^2}{C_x^2}-\rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right)\right] \tag{20}$$

and

$$M\left(T_3\right)=\bar{Y}^2\left[f_1C_y^2+\frac{f_3}{4}\left(C_x^2-4\rho_{yx}C_yC_x\right)+f_2\left(C_z^2-2\rho_{yz}C_yC_z\right)\right] \tag{21}$$

7. Comparison of the estimators

In this section we compare the proposed estimators $T_i\left(i=1,2,3\right)$ with to respect to the estimators $\bar{y}_{rd},\bar{y}_{lrd},\bar{y}_{rc},\bar{y}_{k1},\bar{y}_{k2}$ and $\bar{y}_{k3}$. Preference zones of the estimators T_i are explored and shown below:

(i) $T_i\left(i=1,2,3\right)$ are better than $\bar{y}_{rd}$ if $M\left(T_i\right)\leq M\left(\bar{y}_{rd}\right)$, which gives

$$\rho_{yz}\frac{C_y}{C_z}\geq\frac{1}{4}\quad\left(\text{for } i=1\right) \tag{22}$$

$$\frac{4\left(C_x-\rho_{yx}C_y\right)^2}{C_y^2\left(\rho_{yx}^2\frac{C_z^2}{C_x^2}-4\rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right)}\geq\frac{f_2}{f_3}\quad\left(\text{for } i=2\right) \tag{23}$$

$$\frac{\left(3C_x^2-4\rho_{yx}C_yC_x\right)}{4\left(C_z^2-2\rho_{yz}C_yC_z\right)}\geq\frac{f_2}{f_3}\quad\left(\text{for } i=3\right) \tag{24}$$

(ii) $T_i\left(i=1,2,3\right)$ are preferable over $\bar{y}_{lrd}$ if $M\left(T_i\right)\leq M\left(\bar{y}_{lrd}\right)$, which gives

$$\frac{4\left(C_x-\rho_{yx}C_y\right)^2}{\left(4\rho_{yz}C_yC_z-C_z^2\right)}\leq\frac{f_2}{f_3}\left(\text{for}\, i=1\right) \tag{25}$$

$$\left(\frac{\rho_{yz}}{\rho_{yx}}\right)\frac{C_x}{C_z}\geq\frac{1}{4}\left(\text{for}\, i=2\right) \tag{26}$$

$$\frac{\left(C_x-2\rho_{yx}C_y\right)^2}{4\left(2\rho_{yz}C_yC_z-C_z^2\right)}\leq\frac{f_2}{f_3}\left(\text{for}\ i=3\right) \tag{27}$$

(iii) $T_i\left(i=1,2,3\right)$ will dominate $\bar{y}_{rc}$ if $M\left(T_i\right)\leq M\left(\bar{y}_{rc}\right)$ and subsequently we get the conditions

$$\rho_{yz}\frac{C_y}{C_z}\leq\frac{3}{4}\left(\text{for}\, i=1\right) \tag{28}$$

$$\frac{\left(C_x-\rho_{yx}C_y\right)^2}{\left\{\frac{C_y^2}{4}\left(\frac{\rho_{yx}C_z}{C_x}-2\rho_{yz}\right)^2-\left(C_z-\rho_{yz}C_y\right)^2\right\}}\geq\frac{f_2}{f_3}\left(\text{for}\, i=2\right) \tag{29}$$

$$\rho_{yx}\frac{C_y}{C_x}\leq\frac{3}{4}\left(\text{for}\, i=3\right) \tag{30}$$

(iv) $T_i\left(i=1,2,3\right)$ are more efficient than $\bar{y}_{k1}$ if $M\left(T_i\right)\leq M\left(\bar{y}_{k1}\right)$, which gives

$$\left|\frac{\left(\rho_{yz}C_y-\rho_{xz}C_x\right)}{\left(C_z-2\rho_{yz}C_y\right)}\right|\geq\frac{1}{2}\left(\text{for}\, i=1\right) \tag{31}$$

$$\frac{\left(C_x-\rho_{yx}C_y\right)^2}{\left\{C_y^2\left(\frac{\rho_{yx}^2C_z^2}{4C_x^2}-\rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right)^2-\left(\rho_{xz}^2C_x^2-2\rho_{xz}\rho_{yz}C_yC_x\right)^2\right\}}\geq\frac{f_2}{f_3}\left(\text{for}\, i=2\right) \tag{32}$$

$$\frac{\left(4\rho_{yx}C_yC_x-3C_x^2\right)}{4\left\{\left(\rho_{xz}C_x-\rho_{yz}C_y\right)^2-\left(C_z-\rho_{yz}C_y\right)^2\right\}}\leq\frac{f_2}{f_3}\left(\text{for}\, i=3\right) \tag{33}$$

(v) $T_i\left(i=1,2,3\right)$ are preferable over $\overline{y}_{k2}$ if $M\left(T_i\right)\le M\left(\overline{y}_{k2}\right)$, which gives

$$\frac{\left(C_x-\rho_{yx}C_y\right)^2}{\left[\left\{C_y\left(\rho_{yx}\frac{C_z}{C_x}-\rho_{yz}\right)\right\}^2-\left(\frac{C_z}{2}-\rho_{yz}C_y\right)^2\right]}\le\frac{f_2}{f_3}\ \left(\text{for } i=1\right) \tag{34}$$

$$\frac{\rho_{yz}C_x}{\rho_{yx}C_z}\le\frac{3}{4}\ \left(\text{for } i=2\right) \tag{35}$$

$$\frac{\left(C_x-2\rho_{yx}C_y\right)^2}{4\left\{\left(\rho_{yx}\frac{C_z}{C_x}\right)\left(\rho_{yx}\frac{C_z}{C_x}-2\rho_{yz}\right)+\left(2\rho_{yz}C_yC_z-C_z^2\right)\right\}}\le\frac{f_2}{f_3}\ \left(\text{for } i=3\right) \tag{36}$$

(vi) $T_i\left(i=1,2,3\right)$ will dominate $\overline{y}_{k3}$ if $M\left(T_i\right)\le M\left(\overline{y}_{k3}\right)$ and subsequently we get the conditions

$$\frac{\left(C_x-\rho_{yx}C_y\right)^2}{\left[\left\{C_y\left(\rho_{yx}\rho_{xz}-\rho_{yz}\right)\right\}^2-\left\{\frac{1}{2}\left(C_z-\rho_{yz}C_y\right)\right\}^2\right]}\le\frac{f_2}{f_3}\ \left(\text{for } i=1\right) \tag{37}$$

$$\left[\rho_{yx}\rho_{xz}^2-\rho_{yz}\rho_{xz}\right]\ge\left[\frac{1}{4}\left(\frac{C_z}{C_x}\right)\left(\rho_{yx}-4\rho_{yz}\right)\right]\left(\text{for } i=2\right) \tag{38}$$

$$\frac{\left(C_x-2\rho_{yx}C_y\right)^2}{4\left\{\left(\rho_{yx}\rho_{xz}\right)\left(\rho_{yx}\rho_{xz}-2\rho_{yz}\right)+\left(2\rho_{yz}C_yC_z-C_z^2\right)\right\}}\le\frac{f_2}{f_3}\ \left(\text{for } i=3\right) \tag{39}$$

8. Empirical studies

To examine the performance of the proposed estimators $T_i\left(i=1,2,3\right)$, we have computed the percent relative efficiencies of T_i with respect to $\overline{y}_{rd},\overline{y}_{lrd},\overline{y}_{rc},\overline{y}_{k1}$ based on two natural populations which almost satisfy the conditions shown in equations (22)-(33) and presented in Table 1. The percent relative efficiencies of the estimators T_i with respect to an estimator δ are defined as $\text{PRE}=\left[\frac{\text{MSE}\left(\delta\right)}{\text{MSE}\left(T_i\right)}\right]\text{X}100$; $\left(i=1,2,3\right)$.

Population Source-I: Cochran (1977)

y: Number of ‘placebo’ children.

x: Number of paralytic polio cases in the placebo group.

z: Number of paralytic polio cases in the ‘not inoculated’ group

$N = 34$, $n = 10$, $n' = 15$, $\bar{Y} = 4.92$, $\bar{X} = 2.59$, $\bar{Z} = 2.91$, $C_y^2 = 1.0248$, $C_x^2 = 1.5175$, $C_z^2 = 1.1492$, $\rho_{yx} = 0.7326$, $\rho_{yz} = 0.6430$ and $\rho_{xz} = 0.6837$.

Population Source-II: Fisher (1936)

Consisting of measurements on three variables, namely sepal width (y), sepal length (x) and petal length (z) for 50 Iris flowers (versicolor) such that:

$N = 50$, $n = 10$, $n' = 20$, $\bar{Y} = 2.770$, $C_y^2 = 0.012566$, $C_x^2 = 0.007343$, $C_z^2 = 0.011924$, $\rho_{yx} = 0.5605$, $\rho_{yz} = 0.5259$ and $\rho_{xz} = 0.7540$.

Table 1. The percent relative efficiencies (PRE) based on population I and population II of the proposed estimators $T_i\,(i=1,2,3)$ with respect to the estimators $\bar{y}_{rd}, \bar{y}_{lrd}, \bar{y}_{rc}, \bar{y}_{k1}$.

Estimators	PRE FOR POPULATION I			
	$\bar{y}_{rd}$	$\bar{y}_{lrd}$	$\bar{y}_{rc}$	$\bar{y}_{k1}$
T_1	132.7310	115.5850	113.0849	101.8650
T_2	146.8405	127.8718	125.1060	112.6934
T_3	136.7430	119.0787	116.5031	104.9441
	PRE FOR POPULATION II			
T_1	114.1981	110.6203	110.2922	100.0537
T_2	116.7494	113.0917	112.7563	102.2890
T_3	104.3392	101.0703	100.7705	91.4159

9. Conclusions

It is visible in Table 1 that the proposed estimators $T_i\,(i=1,2,3)$ are preferable over the estimators $\bar{y}_{rd}, \bar{y}_{lrd}, \bar{y}_{rc}$ and $\bar{y}_{k1}$ except the estimator T_3 which is being dominated by the estimator $\bar{y}_{k1}$ in population II. The proposed estimators will also be preferable over the estimators $\overline{y}_{k2}$ and $\overline{y}_{k3}$ for the population which satisfies the conditions derived in equations (34)-(39). Hence, looking on the dominance nature of the proposed estimators, they may be recommended for their practical applications.

Acknowledgments

The authors are thankful to the reviewers for their valuable and inspiring suggestions. The authors are also thankful to the UGC, New Delhi and the Indian School of Mines, Dhanbad for providing the financial assistance and necessary infrastructure to carry out the present research work.

REFERENCES

BAHL, S., TUTEJA, R. K., (1991). Ratio and product type exponential estimator. Information and optimization sciences, 12, 159–163.

CHAND, L., (1975). Some ratio-type estimators based on two or more auxiliary variables. Ph. D. dissertation, Iowa State University, Ames, Iowa.

COCHRAN,W. G., (1977). Sampling techniques. New-York: JohnWiley and Sons.

FISHER, R. A., (1936). The use of multiple measurements in taxonomic problems. Ann. Eugenics, 7,179–188.

GUPTA, S., SHABBIR, J., (2008). On improvement in estimating the population mean in simple random sampling. Journal of Applied Statistics, 35 (5), 559–566.

KIREGYERA, B., (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. Metrika, 27, 217–223.

KIREGYERA, B., (1984) Regression-type estimators using the two auxiliary variables and the model of double sampling from finite population. Metrika, 31(3-4), 215–226.

MUKHERJEE, R., RAO, T. J., VIJAYAN, K., (1987). Regression type estimators using multiple auxiliary information. Australian & New Zealand Journal of Statistics, 29, 244–254.

PRADHAN, B. K., (2005). A chain regression estimator in two-phase sampling using multi-auxiliary information. Bulletin of the malaysian mathematical sciences society, 28(1), 81–86.

SINGH, H. P., VISHWAKARMA, G. K., (2007). Modified exponential ratio and product estimators for finite population mean in double sampling. Austrian journal of statistics, 36, 3, 217–225.

SINGH, G. N., (2001). On the use of transformed auxiliary variable in the estimation of population mean in two-phase sampling. Statistics in Transition, 5(3), 405–416.

SINGH, G. N., UPADHYAYA, L. N., (1995). A class of modified chain type estimators using two auxiliary variables in two-phase sampling. Metron, LIII, 117–125.

SINGH, V. K., SINGH, G. N., (1991). Chain type regression estimators with two auxiliary variables under double sampling scheme. Metron, Vol. XLIX, No. 1–4, 279–289.

SINGH, V. K., SINGH, H. P., SINGH, H. P., (1994). A general class of chain estimators for ratio and product of two means of a finite population. Communications in Statistics-Theory and Methods, 23, 1341–1355.

SINGH, R., CHAUHAN, P., SMARANDACHE, F., (2011). Improvement in estimating population mean using two-auxiliary variables in two-phase sampling. Italian Journal of Pure and Applied Mathematics, 28, 135–142.

SRIVASTAVA, S. R., SRIVASTAVA, S. P., KHARE, B. B., (1989). Chain ratio type estimators for ratio of two population means using auxiliary characters. Communications in Statistics-Theory and Methods, 18, 3917–3926.

UPADHYAYA, L. N., KUSHWAHA, K. S., SINGH, H. P., (1990) A modified chain ratio-type estimator in two-phase sampling using multi-auxiliary information. Metron, 48, 381–393.

UPADHYAYA, L. N., SINGH, G. N., (2001). Chain type estimators using transformed auxiliary variable in two-phase sampling. Advances in modeling and analysis, 38, (1–2), (1–10).

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 231–242

METHODS OF REDUCING DIMENSION FOR FUNCTIONAL DATA

Tomasz Górecki[1], **Mirosław Krzyśko**[2], **Łukasz Waszak**[3], **Waldemar Wołyński**[4]

ABSTRACT

In classical data analysis, objects are characterized by many features observed at one point of time. We would like to present them graphically, to see their configuration, eliminate outlying observations, observe relationships between them or to classify them. In recent years methods for representing data by functions have received much attention. In this paper we discuss a new method of constructing principal components for multivariate functional data. We illustrate our method with data from environmental studies.

Key words: multivariate functional data, functional data analysis, principal component analysis, multivariate principal component analysis.

1. Introduction

The idea of principal component analysis (PCA) is to reduce the dimensionality of a data set consisting of a large number of correlated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming them to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables.

In recent years methods for representing data by functions or curves have received much attention. Such data are known in the literature as functional data (Ramsay and Silverman, 2005). Examples of functional data can be found in various application domains, such as medicine, economics, meteorology and

[1] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: tomasz.gorecki@amu.edu.pl.

[2] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: mkrzysko@amu.edu.pl.

[3] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: lwaszak@amu.edu.pl.

[4] Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poland. E-mail: wolynski@amu.edu.pl.

many others. In previous papers on functional data analysis, objects are characterized by only one feature observed at many time points (see Ramsay and Silverman (2005)). In many applications there is a need to use statistical methods for objects characterized by many features observed at many time points (double multivariate data). In this case, such data are called multivariate functional data. A pioneering theoretical work was that of Besse (1979), where random variables take values in a general Hilbert space. Saporta (1981) presents an analysis of multivariate functional data from the point of view of factorial methods (principal components and canonical analysis). Finally, Jacques and Preda (2014) proposed principal component analysis for multivariate functional data (MFPCA) applied to the methods of cluster analysis. In this paper we propose another method of construction of principal components for multivariate functional data, along with an in-depth interpretation of these variables.

2. Classical principal component analysis (PCA)

Suppose we observe a p-dimensional random vector $\boldsymbol{X} = (X_1, X_2, ..., X_p)' \in \mathbb{R}^p$. We further assume that $\mathrm{E}(\boldsymbol{X}) = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

In the first step we seek a variable U_1 in the form

$$U_1 =< \boldsymbol{u}_1, \boldsymbol{X} >= \boldsymbol{u}_1'\boldsymbol{X} = \sum_{i=1}^{p} u_{1i}X_i,$$

having maximum variance for all $\boldsymbol{u} \in \mathbb{R}^p$ such that $< \boldsymbol{u}, \boldsymbol{u} >= 1$.

Let

$$\lambda_1 = \sup_{\boldsymbol{u}\in\mathbb{R}^p} \mathrm{Var}(< \boldsymbol{u}, \boldsymbol{X} >) = \mathrm{Var}(< \boldsymbol{u}_1, \boldsymbol{X} >) = \boldsymbol{u}_1'\Sigma\boldsymbol{u}_1,$$

where $< \boldsymbol{u}_1, \boldsymbol{u}_1 >= \boldsymbol{u}_1'\boldsymbol{u}_1 = 1$.

The random variable U_1 will be called the first principal component, and the vector $\boldsymbol{u}_1$ will be called the vector of weights of the first principal component.

In the next step we seek a v ariable $U_2 =< \boldsymbol{u}_2, \boldsymbol{X} >= \boldsymbol{u}_2'\boldsymbol{X}$ which is not correlated with the first principal component U_1 and which has maximum variance. We continue this process until we obtain p new variables $U_1, U_2, \ldots, U_p$ (principal components).

In general, the kth principal component $U_k =< \boldsymbol{u}_k, \boldsymbol{X} >= \boldsymbol{u}_k'\boldsymbol{X}$ satisfies the conditions:

$$\lambda_k = \sup_{\boldsymbol{u}\in\mathbb{R}^p} \mathrm{Var}(< \boldsymbol{u}, \boldsymbol{X} >) = \mathrm{Var}(< \boldsymbol{u}_k, \boldsymbol{X} >) = \boldsymbol{u}_k'\Sigma\boldsymbol{u}_k,$$

$$< \boldsymbol{u}_{\kappa_1}, \boldsymbol{u}_{\kappa_2} >= \delta_{\kappa_1\kappa_2}, \qquad \kappa_1, \kappa_2 = 1, ..., k.$$

The expression $(\lambda_k, \boldsymbol{u}_k)$ will be called the kth principal system of the variable $\boldsymbol{X}$ (Jolliffe (2002)).

It can be shown that $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p$ and $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_p$ are the eigenvalues and corresponding eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$.

In practice this matrix is unknown, and must be estimated from the sample. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ be realizations of the vector $\boldsymbol{X}$.

Then

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{x}\boldsymbol{x}'.$$

Moreover, let $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \geqslant \cdots \geqslant \hat{\lambda}_p$ and $\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{u}}_2, \ldots, \hat{\boldsymbol{u}}_p$ be eigenvalues and corresponding eigenvectors of the matrix $\hat{\boldsymbol{\Sigma}}$.

Then $(\hat{\lambda}_k, \hat{\boldsymbol{u}}_k)$ is called the *k*th principal system of the sample of the vector $\boldsymbol{X}$.

The coordinates of the projection of the *i*th realization $\boldsymbol{x}_i$ of the vector $\boldsymbol{X}$ on the *k*th principal component are equal to:

$$\hat{U}_{ik} =< \hat{\boldsymbol{u}}_k, \boldsymbol{x}_i >= \hat{\boldsymbol{u}}_k'\boldsymbol{x}_i,$$

for $i = 1, 2, ..., n, k = 1, 2, ..., p$. Finally, the coordinates of the projection of the i th realization $\boldsymbol{x}_i$ of the vector $\boldsymbol{X}$ on the plane of the first two principal components from the sample are equal to $(\hat{\boldsymbol{u}}_1'\boldsymbol{x}_i, \hat{\boldsymbol{u}}_2'\boldsymbol{x}_i), i = 1, 2, ..., n$.

3. Multivariate functional principal component analysis (MFPCA)

The functional case of PCA (FPCA) is a more informative way of looking at the variability structure in the variance-covariance function for one-dimensional functional data (Górecki and Krzyśko (2012)). In this section we present PCA for multivariate functional data (MFPCA) (Jacques and Preda (2014)).

Suppose that we are observing a p-dimensional stochastic process $\boldsymbol{X}(t) = (X_1(t), X_2(t), ..., X_p(t))'$, with continuous parameter $t \in I$. We will further assume that $\mathrm{E}(\boldsymbol{X}(t)) = \boldsymbol{0}$ and $\boldsymbol{X}(t) \in L_2^p(I)$, where $L_2(I)$ is a Hilbert space of square integrable functions on the interval I equipped with the following inner product:

$$< \boldsymbol{u}(t), \boldsymbol{v}(t) >= \int_I \boldsymbol{u}'(t)\boldsymbol{v}(t)dt.$$

Moreover, assume that the *k*th component of the process $\boldsymbol{X}(t)$ can be represented by a finite number of orthonormal basis functions $\{\varphi_b\}$

$$X_k(t) = \sum_{b=0}^{B_k} c_{kb}\varphi_b(t), t \in I, k = 1, 2, ..., p,$$

where c_{kb} are random variables such that $\mathrm{E}(c_{kb}) = 0$, $\mathrm{Var}(c_{kb}) < \infty$, $k = 1, 2, ..., p, b = 0, ..., B_k$.

Let

$$\boldsymbol{c} = (c_{10}, ..., c_{1B_1}, ..., c_{p0}, ..., c_{pB_p})',$$

$$\boldsymbol{\Phi}(t)=\begin{bmatrix}\boldsymbol{\varphi}_1'(t) & \mathbf{0} & \dots & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\varphi}_2'(t) & \dots & \mathbf{0}\\ \dots & \dots & \dots & \dots\\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\varphi}_p'(t)\end{bmatrix}, \tag{1}$$

where $\boldsymbol{\varphi}_k(t)=(\varphi_0(t),...,\varphi_{B_k}(t))', k=1,...,p.$

Then, the process $\boldsymbol{X}(t)$ can be represented as

$$\boldsymbol{X}(t)=\boldsymbol{\Phi}(t)\boldsymbol{c},\ t\in I,\ \mathrm{E}(\boldsymbol{c})=\mathbf{0},\ \mathrm{Var}(\boldsymbol{c})=\Sigma_{\boldsymbol{c}}.$$

We are interested to find the inner product

$$U=<\boldsymbol{u}(t),\boldsymbol{X}(t)>=\int_I\boldsymbol{u}'(t)\boldsymbol{X}(t)dt$$

having maximal variance for all $\boldsymbol{u}(t)\in L_2^p(I)$ such that $<\boldsymbol{u}(t),\boldsymbol{u}(t)>=1$. It may be assumed that the vector weight function $\boldsymbol{u}(t)$ and the process $\boldsymbol{X}(t)$ are in the same space, i.e. the function $\boldsymbol{u}(t)$ can be written in the form:

$$\boldsymbol{u}(t)=\boldsymbol{\Phi}(t)\boldsymbol{u},$$

where $\boldsymbol{u}\in\mathbb{R}^{K+p}$, $K=B_1+...+B_p$. Then

$$<\boldsymbol{u}(t),\boldsymbol{X}(t)>=<\boldsymbol{\Phi}(t)\boldsymbol{u},\boldsymbol{\Phi}(t)\boldsymbol{c}>=\boldsymbol{u}'<\boldsymbol{\Phi}(t),\boldsymbol{\Phi}(t)>\boldsymbol{c}=\boldsymbol{u}'\boldsymbol{c}$$

and

$$\mathrm{E}(<\boldsymbol{u}(t),\boldsymbol{X}(t)>)=\boldsymbol{u}'E(\boldsymbol{c})=\boldsymbol{u}'\mathbf{0}=0,$$

$$\mathrm{Var}(<\boldsymbol{u}(t),\boldsymbol{X}(t)>)=\boldsymbol{u}'E(\boldsymbol{c}\boldsymbol{c}')\boldsymbol{u}=\boldsymbol{u}'\boldsymbol{\Sigma}_{\boldsymbol{c}}\boldsymbol{u}.$$

Let

$$\lambda_1=\sup_{\boldsymbol{u}(t)\in L_2^p(I)}\mathrm{Var}(<\boldsymbol{u}(t),\boldsymbol{X}(t)>)=\mathrm{Var}(<\boldsymbol{u}_1(t),\boldsymbol{X}(t)>)=\boldsymbol{u}_1'\boldsymbol{\Sigma}_{\boldsymbol{c}}\boldsymbol{u}_1,$$

where $<\boldsymbol{u}_1(t),\boldsymbol{u}_1(t)>=\boldsymbol{u}_1'\boldsymbol{u}_1=1$.

The inner product $U_1=<\boldsymbol{u}_1(t),\boldsymbol{X}(t)>=\boldsymbol{u}_1'\boldsymbol{c}$ will be called the first principal component, and the vector function $\boldsymbol{u}_1(t)$ will be called the first vector weight function. Subsequently we look for the second principal component $U_2=<\boldsymbol{u}_2(t),\boldsymbol{X}(t)>=\boldsymbol{u}_2'\boldsymbol{c}$, maximizing $\mathrm{Var}(<\boldsymbol{u}(t),\boldsymbol{X}(t)>)=\boldsymbol{u}'\boldsymbol{\Sigma}_{\boldsymbol{c}}\boldsymbol{u}$, such that $<\boldsymbol{u}_2(t),\boldsymbol{u}_2(t)>=\boldsymbol{u}_2'\boldsymbol{u}_2=1$, and not correlated with the first functional principal component U_1,i.e. subject to the restriction $<\boldsymbol{u}_1(t),\boldsymbol{u}_2(t)>=\boldsymbol{u}_1'\boldsymbol{u}_2=0$.

In general, the kth functional principal component $U_k=<\boldsymbol{u}_k(t),\boldsymbol{X}(t)>=\boldsymbol{u}_k'\boldsymbol{c}$ satisfies the conditions:

$$\lambda_k=\sup_{\boldsymbol{u}(t)\in L_2^p(I)}\mathrm{Var}(<\boldsymbol{u}(t),\boldsymbol{X}(t)>)=\mathrm{Var}(<\boldsymbol{u}_k(t),\boldsymbol{X}(t)>)=\boldsymbol{u}_k'\boldsymbol{\Sigma}_{\boldsymbol{c}}\boldsymbol{u}_k,$$

$$<\boldsymbol{u}_{\kappa_1}(t),\boldsymbol{u}_{\kappa_2}(t)>=\delta_{\kappa_1\kappa_2},\qquad \kappa_1,\kappa_2=1,...,k.$$

The expression $(\lambda_k, \boldsymbol{u}_k(t))$ will be called the *k*th principal system of the process $\boldsymbol{X}(t)$.

Now, let us consider the principal components of the random vector $\boldsymbol{c}$. The k th principal component $U_k^* =< \boldsymbol{u}_k, \boldsymbol{c} >$ of this vector satisfies the conditions:

$$\gamma_k = \sup_{\boldsymbol{u}\in\mathbb{R}^{K+p}} Var(< \boldsymbol{u}, \boldsymbol{c} >) = \sup_{\boldsymbol{u}\in\mathbb{R}^{K+p}} \boldsymbol{u}' Var(\boldsymbol{c})\boldsymbol{u} = \sup_{\boldsymbol{u}\in\mathbb{R}^{K+p}} \boldsymbol{u}'\boldsymbol{\Sigma}_{\boldsymbol{c}}\boldsymbol{u},$$

$$\boldsymbol{u}'_{\kappa_1}\boldsymbol{u}_{\kappa_2} = \delta_{\kappa_1\kappa_2},$$

where $\kappa_1, \kappa_2 = 1, ..., k$, $K = B_1 + ... + B_p$. The expression $(\gamma_k, \boldsymbol{u}_k)$ will be called the *k*th principal system of the vector $\boldsymbol{c}$.

Determining the *k*th principal system of the vector $\boldsymbol{c}$ is equivalent to solving for the eigenvalue and corresponding eigenvectors of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{c}}$ of that vector, standardized so that $\boldsymbol{u}'_{\kappa_1}\boldsymbol{u}_{\kappa_2} = \delta_{\kappa_1\kappa_2}$.

From the above considerations, we have the following theorem:

Theorem. The *k*th principal system $(\lambda_k, \boldsymbol{u}_k(t))$ of the stochastic process $\boldsymbol{X}(t)$ is related to the *k*th principal system $(\gamma_k, \boldsymbol{u}_k)$ of the random vector $\boldsymbol{c}$ by the equations:

$$\lambda_k = \gamma_k, \qquad \boldsymbol{u}_k(t) = \boldsymbol{\Phi}(t)\boldsymbol{u}_k, \qquad t \in I,$$

where $k = 1, ..., K + p$, $K = B_1 + B_2 + \cdots + B_p$.

Principal component analysis for random vectors $\boldsymbol{c}$ is based on the matrix $\boldsymbol{\Sigma}_{\boldsymbol{c}}$. In practice this matrix is unknown. We estimate it on the basis of n independent realizations $\boldsymbol{x}_1(t), \boldsymbol{x}_2(t), ..., \boldsymbol{x}_n(t)$ of the random process $\boldsymbol{X}(t)$.

Typically data are recorded at discrete moments in time. The process of transformation of discrete data to functional data is performed for each variable $X_1, X_2, \ldots X_p$ separately.

Let x_{kj} denote an observed value of feature X_k, $k = 1, 2, \ldots p$ at the *j*th time point t_j, where $j = 1, 2, ..., J$. Then our data consist of pJ pairs of (t_j, x_{kj}). This discrete data can be smoothed by continuous functions $x_k(t)$, where $t \in I$ (Ramsay and Silverman (2005)). Let I be a compact set such that $t_j \in I$, for $j = 1, ..., J$. Let us assume that the function $x_k(t)$ has the following representation

$$x_k(t) = \sum_{b=0}^{B_k} c_{kb}\varphi_b(t),\ t \in I,\ k = 1, \ldots, p, \tag{2}$$

where $\{\varphi_b\}$ are orthonormal basis functions, and $c_{k0}, c_{k1}, \ldots, c_{kB_k}$ are the coefficients.

Let $\boldsymbol{x}_k = (x_{k1}, x_{k2}, \ldots, x_{kJ})'$, $\boldsymbol{c}_k = (c_{k0}, c_{k1}, \ldots, c_{kB_k})'$ and $\boldsymbol{\Phi}_k(t)$ be a matrix of dimension $J \times (B_k + 1)$ containing the values $\varphi_b(t_j)$, $b = 0, 1, ..., B$, $j = 1, 2, ..., J$, $k = 1, \ldots, p$. The coefficient $\boldsymbol{c}_k$ in (2) is estimated by the least squares method, that is, so as to minimize the function:

$$S(\boldsymbol{c}_k) = (\boldsymbol{x}_k - \boldsymbol{\Phi}_k(t)\boldsymbol{c}_k)' (\boldsymbol{x}_k - \boldsymbol{\Phi}_k(t)\boldsymbol{c}_k),\ k = 1, \ldots, p.$$

Differentiating $S(\boldsymbol{c}_k)$ with respect to the vector $\boldsymbol{c}_k$, we obtain the least squares method estimator

$$\hat{\boldsymbol{c}_k} = \left(\boldsymbol{\Phi}_k'(t)\boldsymbol{\Phi}_k(t)\right)^{-1}\boldsymbol{\Phi}_k'(t)\boldsymbol{x}_k,\ k = 1,\dots,p.$$

The degree of smoothness of the function $x_k(t)$ depends on the value B_k (a small value of B_k causes more smoothing of the curves). The optimum value for B_k may be selected using the Bayesian information criterion BIC (see Shmueli (2010)).

Let us assume that there are n independent pairs of values (t_j, x_{kij}), $k = 1,\dots,p$, $i = 1,...,n$, $j = 1,...,J$. These discrete data are smoothed to continuous functions in the following form:

$$x_{ki}(t) = \sum_{b=0}^{B_{ki}} \hat{c}_{kib}\varphi_b(t),\ k = 1,\dots,p,\ i = 1,...,n,\ t \in I.$$

Among all the $B_{k1}, B_{k2},..., B_{kn}$ one common value of B_k is chosen, as the modal value of the numbers $B_{k1}, B_{k2},..., B_{kn}$, and we assume that each function $x_{ki}(t)$ has the form

$$x_{ki}(t) = \sum_{b=0}^{B_k} \hat{c}_{kib}\varphi_b(t),\ k = 1,\dots,p,\ i = 1,...,n,\ t \in I.$$

The data $\{x_{k1}(t),...,x_{kn}(t)\}$ are called functional data (see Ramsay and Silverman (2005)).

Finally, each of n independent realizations $\boldsymbol{x}_1(t), \boldsymbol{x}_2(t),....,\boldsymbol{x}_n(t)$ has the form $\boldsymbol{x}_i(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{c}}_i$ where $\boldsymbol{\Phi}(t)$ is given by (1) and the vectors $\hat{\boldsymbol{c}}_i = (\hat{c}_{10},...,\hat{c}_{1B_1},...,\hat{c}_{p0},...,\hat{c}_{pB_p})'$ are centred, $i = 1, 2,..., n$.

Let $\hat{\boldsymbol{C}} = (\hat{\boldsymbol{c}}_1, \hat{\boldsymbol{c}}_2,...,\hat{\boldsymbol{c}}_n)$. Then

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{c}} = \frac{1}{n}\hat{\boldsymbol{C}}\hat{\boldsymbol{C}}'.$$

Let $\hat{\gamma}_1 \geqslant \hat{\gamma}_2 \geqslant ... \geqslant \hat{\gamma}_s$ be non-zero eigenvalues of the matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{c}}$, and $\hat{\boldsymbol{u}}_1, \hat{\boldsymbol{u}}_2,..., \hat{\boldsymbol{u}}_s$ the corresponding eigenvectors, where s =rank($\hat{\boldsymbol{\Sigma}}_{\boldsymbol{c}}$).

Moreover, the kth principal system of the random process $\boldsymbol{X}(t)$ determined from the sample has the following form:

$$(\hat{\lambda}_k = \hat{\gamma}_k, \hat{\boldsymbol{u}}_k(t) = \boldsymbol{\Phi}(t)\hat{\boldsymbol{u}}_k), \qquad k = 1,..., s.$$

The coordinates of the projection of the ith realization $\boldsymbol{x}_i(t)$ of the process $\boldsymbol{X}(t)$ on the kth functional principal component are equal to:

$$\hat{U}_{ik} =< \hat{\boldsymbol{u}}_k(t), \boldsymbol{x}_i(t) >=< \boldsymbol{\Phi}(t)\hat{\boldsymbol{u}}_k, \boldsymbol{\Phi}(t)\hat{\boldsymbol{c}}_i >= \hat{\boldsymbol{u}}_k' < \boldsymbol{\Phi}(t), \boldsymbol{\Phi}(t) > \hat{\boldsymbol{c}}_i = \hat{\boldsymbol{u}}_k'\hat{\boldsymbol{c}}_i,$$

for $i = 1, 2,..., n$, $k = 1, 2,..., s$. Finally, the coordinates of the projection of the i th realization $\boldsymbol{x}_i(t)$ of the process $\boldsymbol{X}(t)$ on the plane of the first two functional principal components from the sample are equal to $(\hat{\boldsymbol{u}}_1'\hat{\boldsymbol{c}}_i, \hat{\boldsymbol{u}}_2'\hat{\boldsymbol{c}}_i)$, $i = 1, 2,..., n$.

4. Example

Data relating to environmental protection were obtained from Professor W. Ratajczak of the Spatial Econometry Group at the Geographical and Geological Sciences Faculty of Adam Mickiewicz University, Poznań. The analysis relates to the 16 Polish provinces ($n = 16$). On the graphs, the provinces are denoted by numbers as given in Table 1.

Table 1. Designations of provinces

1	ŁÓDZKIE
2	MAZOWIECKIE
3	MAŁOPOLSKIE
4	ŚLĄSKIE
5	LUBELSKIE
6	PODKARPACKIE
7	PODLASKIE
8	ŚWIĘTOKRZYSKIE
9	LUBUSKIE
10	WIELKOPOLSKIE
11	ZACHODNIOPOMORSKIE
12	DOLNOŚLĄSKIE
13	OPOLSKIE
14	KUJAWSKO-POMORSKIE
15	POMORSKIE
16	WARMIŃSKO-MAZURSKIE

The analyzed data cover a period of 10 years, from 2002 to 2011 ($J = 10$). Each province was characterized by a group of 6 features ($p = 6$):

1. Gaseous pollutant emissions [t/km^2]
2. Dust pollutant emissions [kg/km^2]
3. Solid waste produced [t/km^2]
4. Total liquid waste [dam^3/1000 residents]
5. Industrial liquid waste [dam^3/1000 residents]
6. Household and industrial water consumption [dam^3/1000 residents]

The classical method of principal component analysis (PCA) permits only separate analysis for each year of observation. Tables 2–5 contain the weights and the percentage contributions for the first and second principal component.

Table 2. Weights (eigenvectors) of the first principal component (analysis for a fixed time)

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1	0.6181	0.6480	0.6404	0.6535	0.6985	0.7076	0.7564	0.7514	0.7591	0.7748
2	0.4230	0.3964	0.3612	0.3140	0.2773	0.3042	0.2200	0.1962	0.2017	0.1884
3	0.6609	0.6486	0.6762	0.6869	0.6579	0.6363	0.6144	0.6283	0.6180	0.6022
4	0.0013	0.0010	0.0007	0.0008	0.0008	0.0006	0.0008	0.0009	0.0006	0.0005
5	-0.0333	-0.0326	-0.0316	-0.0331	-0.0320	-0.0289	-0.0297	-0.0303	-0.0215	-0.0257
6	-0.0347	-0.0338	-0.0344	-0.0370	-0.0372	-0.0327	-0.0339	-0.0358	-0.0274	-0.0300

Table 3. Percentage contribution of the original variables in the structure of the first principal component (analysis for a fixed time)

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1	38.2048	41.9904	41.0112	42.7062	48.7902	50.0698	57.2141	56.4602	57.6233	60.0315
2	17.8929	15.7133	13.0465	9.8596	7.6895	9.2538	4.8400	3.8494	4.0683	3.5495
3	43.6789	42.0682	45.7246	47.1832	43.2832	40.4878	37.7487	39.4761	38.1924	36.2645
4	0.0002	0.0001	0.0000	0.0001	0.0001	0.0000	0.0001	0.0001	0.0000	0.0000
5	0.1109	0.1063	0.0999	0.1096	0.1024	0.0835	0.0882	0.0918	0.0462	0.0660
6	0.1204	0.1142	0.1183	0.1369	0.1384	0.1069	0.1149	0.1282	0.0751	0.0900

Table 4. Weights (eigenvectors) of the second principal component (analysis for a fixed time)

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1	0.7697	0.7405	0.6493	0.7145	0.3636	0.4864	0.4985	0.4124	0.3529	0.3770
2	-0.2033	-0.1625	0.0214	0.0060	0.0548	0.0398	0.0368	0.0161	-0.0100	-0.0103
3	-0.5764	-0.6277	-0.5937	-0.6673	-0.3448	-0.5116	-0.5798	-0.4388	-0.3824	-0.4280
4	0.0013	0.0022	0.0024	0.0023	0.0018	0.0008	0.0010	0.0005	0.0022	0.0018
5	0.1235	0.1194	0.3322	0.1459	0.5975	0.4978	0.4560	0.5652	0.6065	0.5830
6	0.1368	0.1304	0.3393	0.1511	0.6237	0.5022	0.4539	0.5636	0.6010	0.5785

Table 5. Percentage contribution of the original variables in the structure of the second principal component (analysis for a fixed time)

	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
1	59.2438	54.8340	42.1590	51.0510	13.2205	23.6585	24.8502	17.0074	12.4538	14.2129
2	4.1331	2.6406	0.0458	0.0036	0.3003	0.1584	0.1354	0.0259	0.0100	0.0106
3	33.2237	39.4007	35.2480	44.5289	11.8887	26.1735	33.6168	19.2545	14.6230	18.3184
4	0.0002	0.0005	0.0006	0.0005	0.0003	0.0001	0.0001	0.0000	0.0005	0.0003
5	1.5252	1.4256	11.0357	2.1287	35.7006	24.7805	20.7936	31.9451	36.7842	33.9889
6	1.8714	1.7004	11.5124	2.2831	38.9002	25.2205	20.6025	31.7645	36.1201	33.4662

The relative position of the 16 provinces (in 2002 and 2011) in the system of the first two principal components is shown in Figure 1.

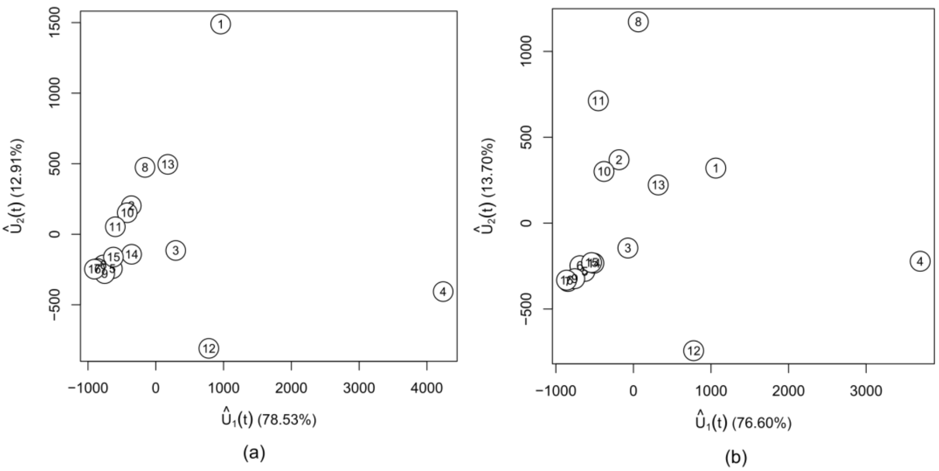

Figure 1. Projection of the six-dimensional vectors representing the 16 provinces on the plane of the first two principal components, (a) year 2002, (b) year 2011

The functional principal components method enables combined analysis of the data for the whole of the studied period of time. The data were transformed to functional data by the method described in Section 3. The calculations were performed using the Fourier basis. The time interval [0,*T*]=[0,10] was divided into moments of time in the following way: t_1=0.5(2002), t_2=1.5(2003),…, t_{10}=9.5(2011). Moreover, in view of the small number of time periods (*J*=10), for each variable the maximum number of basis components was taken, equal to $B_1 = \cdots = B_{10} = 9$.

Tables 6–7 show the coefficients of the weight functions for the first and second functional principal components.

Table 6. Coefficients of weight functions for the first functional principal component

	$\hat{u}_0$	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$	$\hat{u}_4$	$\hat{u}_5$	$\hat{u}_6$	$\hat{u}_7$	$\hat{u}_8$	Area
1	0.6947	-0.0007	-0.0179	-0.0056	0.0189	-0.0220	-0.0104	-0.0064	-0.0046	2.1968
2	0.2927	0.0794	0.0094	0.0401	0.0138	0.0114	-0.0101	0.0299	0.0011	0.9256
3	0.6443	0.0551	-0.0088	0.0086	0.0129	-0.0010	-0.0074	0.0147	-0.0002	2.0375
4	0.0008	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000	0.0025
5	-0.0317	0.0001	0.0009	-0.0005	-0.0009	0.0003	0.0003	0.0009	-0.0004	0.1002
6	-0.0355	0.0013	0.0011	0.0006	-0.0008	0.0003	0.0004	0.0013	-0.0004	0.1123

Table 7. Coefficients of weight functions for the second functional principal component

	$\hat{u}_0$	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$	$\hat{u}_4$	$\hat{u}_5$	$\hat{u}_6$	$\hat{u}_7$	$\hat{u}_8$	Area
1	0.5771	-0.0005	0.0680	0.0392	0.0193	0.0165	-0.0105	-0.0091	-0.0114	1.8250
2	-0.0122	-0.0227	-0.0289	-0.0087	-0.0092	-0.0140	0.0010	-0.0157	0.0005	0.1021
3	-0.5636	-0.0314	0.0210	-0.0042	-0.0070	0.0018	0.0017	-0.0078	0.0034	1.7823
4	0.0017	0.0004	0.0003	-0.0001	-0.0002	-0.0002	-0.0003	-0.0001	-0.0002	0.0054
5	0.4080	-0.0210	0.0120	-0.0081	0.0096	-0.0065	-0.0155	-0.0117	0.0055	1.2902
6	0.4120	-0.0164	0.0081	-0.0069	0.0122	-0.0041	-0.0163	-0.0115	0.0072	1.3029

At a given time point *t*, the greater is the absolute value of a component of the vector weight, the greater is the contribution in the structure of the given functional principal component, from the process $X(t)$ corresponding to that component. The total contribution of a particular primary process $X_i(t)$ in the structure of a particular functional principal component is equal to the area under the module weighting function corresponding to this process. These contributions for the six components of the vector process $\boldsymbol{X}(t)$, and the first and second functional principal components are given in Tables 6–7.

Figure 2 s hows the six weight functions for the first and second functional principal components.

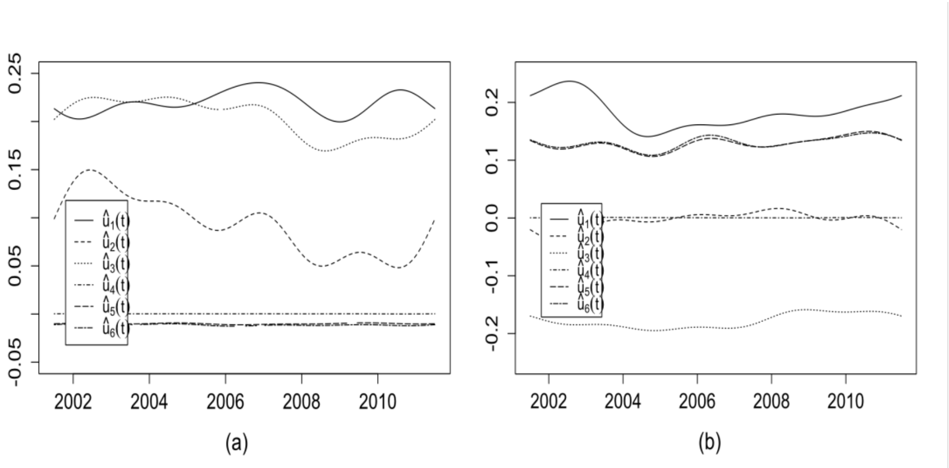

Figure 2. Weight functions for the first (a) and second (b) functional principal component (MFPCA)

The relative positions of the 16 p rovinces in the system of the first two functional principal components are shown in Figure 3. T he system of the first two functional principal components retains 90.33% of the total variation.

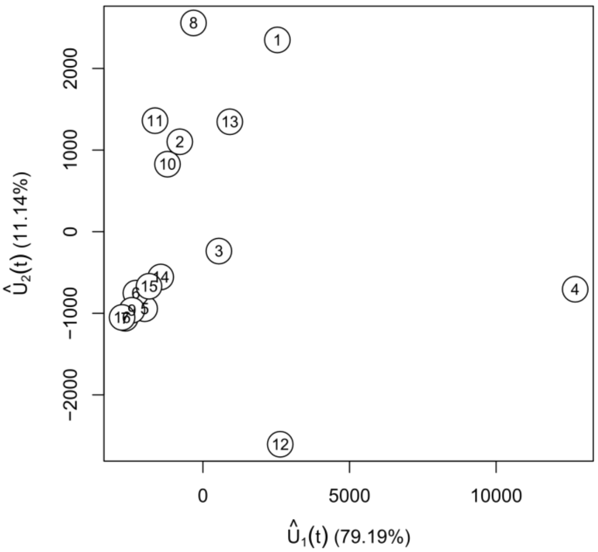

Figure 3. Projection of multidimensional functional data representing the 16 provinces on the plane of the first two functional principal components

5. Conclusions

This paper introduces and analyzes a new method of constructing principal components for multivariate functional data. This method was applied to environmental multivariate time series concerning the Polish provinces. Our research has shown, on this example, that the use of a m ultivariate functional principal components analysis gives good results. Of course, the performance of the algorithm needs to be further evaluated on additional real and artificial data sets. In a similar way, we can extend similar methods like functional discriminant coordinates (Górecki et al. (2014)) and canonical correlation analysis (Krzyśko, Waszak (2013)) to multivariate case. This is the direction of our future research.

REFERENCES

BESSE, P., (1979). Etude descriptive d'un processus. Ph.D. thesis, Université Paul Sabatier.

FISHER, R. A., (1936). The use of multiple measurements in taxonomic problem, Annals of Eugenics 7, 179–188.

GÓRECKI, T., KRZYŚKO, M., (2012). Functional Principal Components Analysis, Data analysis methods and its applications, C.H. Beck 71–87.

GÓRECKI, T., KRZYŚKO, M., WASZAK, Ł., (2014). Functional Discriminant Coordinates, Communication in Statistics – Theory and Methods 43(5), 1013–1025.

JACQUES, J., PREDA, C., (2014). Model-based clustering for multivariate functional data, Computational Statistics & Data Analysis 71, 92–106.

JOLLIFFE, I. T., (2002). Principal Component Analysis, Second Edition, Springer.

KRZYŚKO, M., WASZAK, Ł., (2013). Canonical correlation analysis for functional data, Biometrical Letters 50(2), 95–105.

RAMSAY, J. O., SILVERMAN, B. W., (2005). Functional Data Analysis, Second Edition, Springer.

SAPORTA, G., (1981). Methodes exploratoires d'analyse de données temporelles. Cahiers du Buro. Ph.D. thesis.

SHMUELI, G., (2010). To explain or to predict? Statistical Science 25(3), 289–310.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 243–268

MULTIOBJECTIVE OPTIMIZATION OF FINANCING HOUSEHOLD GOALS WITH MULTIPLE INVESTMENT PROGRAMS

Łukasz Feldman[1], **Radosław Pietrzyk**[2], **Paweł Rokita**[3]

ABSTRACT

This article proposes a technique of facilitating life-long financial planning for a household by finding the optimal match between systematic investment products and multiple financial goals of different realization terms and magnitudes. This is a multi-criteria optimization. One of the objectives is compliance between the expected term structure of cumulated net cash flow throughout the life cycle of the household with its life-length risk aversion and bequest motive. The second is financial liquidity in all periods under expected values of all stochastic factors. The third is minimization of net cash flow volatility. The fourth is minimization of costs of the investment plan combination. The result is a set of systematic-investment programs with accompanying information which programs are destined to cover which financial goal. Payoffs of one program may be used to cover more than one goal and the order may be other than sequential. An original goal function, constructed to suit conditions and assumptions of the proposed household financial plan model, is presented as an optimization procedure.

Key words: multiobjective optimization, personal finance, asset selection, intertemporal choice.

1. Introduction

Based on Modigliani and Brumberg (1954), Ando and Modigliani (1957) and Yaari (1965) life-cycle consumption, as well as on the dynamic asset allocation models by Merton (1969, 1971) and Richard (1975), a vast literature on lifetime financial planning for individuals has been developed so far. A common concept that underlies modern personal finance models is expressing intertemporal choice

[1] Wroclaw University of Economics, ul. Komandorska 118/120, 53-345 Wroclaw, Poland. E-mail: lukasz.feldman@ue.wroc.pl.

[2] Wroclaw University of Economics, ul. Komandorska 118/120, 53-345 Wroclaw, Poland. E-mail: radoslaw.pietrzyk@ue.wroc.pl.

[3] Wroclaw University of Economics, ul. Komandorska 118/120, 53-345 Wroclaw, Poland. E-mail: pawel.rokita@ue.wroc.pl.

situation in terms of expected discounted utility. Following Yaari (1965), the goal function to be maximized was expected discounted utility of consumption, where consumption was expressed as a consumption rate; utilities were weighted with conditional probability of survival of an individual, preferences did not change over time and were independent from period to period (time separable preferences). There was one argument of the utility function (consumption) and one utility function. This model was then developed and augmented in many directions. Bodie, Merton and Samuelson (1992) presented a model providing optimization of both consumption and investment decisions. Amongst other findings, they proposed to use consumption of leisure time (or, put it differently – the amount of work per unit of time) as an additional decision variable. They also showed the importance of human capital and its risk in consumption and investment decisions by individuals. On this ground a significant branch of personal finance models originated. Other assumptions of the original Yaari (1965) and Merton (1969, 1971) constructs were relaxed. The models allowed for habit formation (relaxing the assumption that preferences are independent in time), multiple risky assets (Bodie, at al., 2004), or optimization of retirement time (Sundaresan and Zapatero, 1997). Bodie (2007) presented a brief outline of the basic analytical framework including the most significant recent findings. Other propositions of further development include: using stochastic force of mortality in survival process (Huang, Milevsky, Salisbury, 2012), taking into account maximum psychological planning horizon (Carbone, Infante 2012) or behavioural biases – the concept of using hyperbolic discounting is included here (Ainslie, 1975, 1991; Kirby and Herrnstein, 1995). Geyer, Hanke and Weissensteiner (2009) presented a model allowing for stochastic labour income and investment opportunities. Scholz and Seshadri (2012) proposed to treat health as a type of asset and "production of health" as a particular form of investment. A lot of work has been also done in the area of retirement capital deployment in the retirement phase of the life cycle (Huang and Milevsky 2011; Milevsky and Huang, 2011; Gong and Webb, 2008; Dus, Maurer and Mitchell, 2004).

Despite many-sided and rapid development of the discipline, there are some important practical aspects of lifetime financial planning that have not been elaborated well yet. The aforementioned models concentrate on decisions made by individuals, whereas in personal finance a typical decision making entity is the household.

In this article a model of two-person household is used. A single decision maker is treated just as a special case.

The analysis of household consumption cannot neglect interconnections between persons. Even under the assumption that individual survival processes are independent, neither cash flows nor assets and liabilities assigned to household members are independent. For instance, cumulated investment in pension-plan products may be inherited by one spouse if that other dies before her (his) retirement age. The model assumes that after retirement date life annuity is bought, which, in turn, cannot be inherited. Thus, cumulated investment of one person depends on whether that other is alive, and if not – on time of death (in relation to retirement age). There are more such interdependences between financial categories building up c onsumption of the household. Moreover, some of the quantities are of cumulative nature. Also, the main household financial situation indicator used here, namely – the cumulated net cash flow (cumulated surplus), is a process of this kind. This makes conditional probabilities of being alive far insufficient for calculation of expected discounted utility. The whole history of the process needs to be taken into consideration instead. However, because the number of two-person survival process trajectories grows fast with the number of future periods spanned by the plan, some simplifications are needed. This is what was not necessary in the models discussed before. A proposal of simplification, which, moreover, has a very natural and practical interpretation, is presented here as one of the inputs.

Another specificity of the household, as opposed to single person, lies in the nature of risk connected with lifetime uncertainty. For a single person, only unexpected longevity might have adverse financial consequences. Thus, lifetime risk was regarded identical with longevity risk. For the household, also early death of one member (particularly the one who earns more) may threat financial liquidity.

One more difference is in retirement planning. It is not necessary (though most secure) that retirement income of each household member covers to the full extent fixed costs of the household and the part of variable costs that may be assigned to this person – 2 x full retirement as defined by Feldman, Pietrzyk and Rokita (2014b). The possibility of other retirement schemes (full-partial or even 2 x partial) gives a bigger range of feasible combinations of (1) the proportion of means allotted for consumption and investments, and (2) proportion of common investment assigned to particular persons.

Like in vast part of the literature, it is assumed here that financial goals set by the household are not subject to automated optimization. A satisfactory offset between the most desired and feasible bundle of goals (taking into account time structure of goals and their size) is approached recursively by means of external adjustments – if previous version of goal settings turns out to be unattainable given other constraints. The decision which goals should be rescheduled, reduced in size or abandoned is always made by household members. This approach is adopted because it would be a very hard task to define hierarchy of more than two goals, preferences of which may be not separable, nor transitive.

As far as the intended result of optimization is concerned, there is a difference between the majority of models discussed in the literature and the proposition presented here. The typical approach is focused on smoothing consumption, whereas in this research the aim is to obtain such term structure of cumulated surplus which best suits lifetime risk aversion and bequest motive of the household. Dependent on risk aversion level, financial plans differ just in shapes of expected cumulated surplus trajectory. The shape indicates which retirement profile will be realized (i.e., whether it will be 2 x full, full-partial or 2 x partial retirement). Concentrating on the shape of the cumulated net cash flow trajectory instead of consumption is a consequence of the way the household financial situation is modelled. The expected trajectory of the cumulated net cash flow is a fingerprint of each particular financial plan.

In addition to the aforementioned two-person household approach, intuitive and easily applicable definition of risk aversion measures and simplification of the optimization problem by limiting the number of survival scenarios, also the value function construction and its application may be ranked amongst original inputs of this research. The value function evaluates utility of the term structure of cumulated surplus, taking also into account consumption. It may be used as a goal function of the optimization procedure, but also – as it is discussed in more details in Section 4 – facilitates comparison of otherwise hardly comparable investment products. This property makes it a useful tool of finding a match between multiple investment products and household goals.

The paper is organized as follows. In Section 1 assumptions and basic components of the model are described. Household goals are discussed in Section 2. Construction of value function (being also a goal function for the optimization procedure) is presented in Section 3. It is a function of utilities calculated for consumption and bequest. What is also proposed is a simplification that allows avoiding taking all possible trajectories of bivariate survival process into account when searching for the maximum expected discounted utility. This simplification implies, at the same time, a straightforward definition of risk aversion measure (in respect of length-of-life related risk). Section 4 contains a step-by-step description of the procedure of matching multiple systematic-investment programs with a number of financial goals. Section 5 presents a numerical example. These are the results of the procedure described in Section 4 applied for a demonstration-case household. The last section concludes.

2. Basic concepts

When constructing a financial plan with a number of goals and multiple investment products available, two tasks are to be discussed. The first is selecting a combination from amongst available systematic-investment programs. The second is optimization of the term structure of household cash flows. While the question in the first task is which goals should be financed with which programs (assuming that one program may be used for financing more than one goal), in the second task the issue concerns the level of consumption, investments, as well as proportions in which household members participate in joint investments of the household, given all constraints, amongst which budget constraint is the most typical example. The procedure of carrying out the first task is described in Section 4. In the second task the goal function described in Section 3 is maximized. Both tasks are, of course, strictly connected. Combinations of investment programs chosen in preliminary selection as part of task 1 are evaluated by putting them into the model of household cash flows used in task 1 and calculating goal function value for each of them. The household cash flow model, together with the corresponding value function – being also the goal function of the optimization procedure in task 2 – is the basic construct to be discussed here. It is also the tool supporting financial plan construction (and, thus, among others, also carrying out task 1). In this section assumptions of the household cash flow model are provided and some basic notions shading light on how the model is constructed are discussed.

2.1. Assumptions

The model is based on a set of assumptions. They refer to the household itself, its incomes and expenses, construction of household goal function, and also to some elements of economic environment. They are as follows:

- Two-person household – if there are any persons other than the two *main household members*, they are represented in the model as elements of financial situation of the main two; a single person is treated just as a special case of a (reduced) pair.
- Both main members intend to remain in the household until their death.
- Goal function of the household is composed of two elements:
  - utility of consumption,
  - utility of bequest.
- Goal function s constructed using the concept of expected discounted utility.
- Survival processes of two main household members are independent.
- Joint utility function of the whole household is considered.
- Analytical form of utility function is the same for consumption and bequest.
- Household income in pre-retirement period is constant in real terms (inflation indexed).
- Fixed real rate of return on private pension plan.
- Pension income constant in real terms (inflation indexed).
- Fixed replacement rate (but may be different for women and men).
- Household members buy life annuity.
- Household consumption is fixed at the planned level unless running out of cumulated surplus (loosing liquidity).
- Optimization scope (not to be confused with domain of decision variables) is limited to the area determined by risk aversion of household members – *range of concern*.
- Risk aversion is limited to the length of life of household members; no other types of risk are considered.
- The surplus over consumption just cumulates – it is not invested, neither is it squandered.
- No will to work after retirement is taken into account. Thus, human capital in retirement is zero and the whole capital of the household that may be then

considered is reflected, on the asset side, just in form of cumulated investments into private pension plans and cumulated financial surplus.

- Bequest is not counted among financial goals of the household (but, if there is nonzero bequest motive, bequest-leaving potential is positively valued by the goal function).
- There is a constraint that all financial goals, together with retirement (that is, the main goal), must be realized – under expected values of death time no unutilized cumulated investment is left.
- Conditional survival probabilities used in discounted expected utility calculation may be obtained from any survival model, like Gompertz-Makeham (Gompertz 1825, Makeham 1860). This is, however, a secondary issue at this stage of research since the choice of mortality model does not influence in any way the very concept of the financial plan optimization procedure that is proposed here. It may, certainly, have impact on the final results of the optimization, which may require some detailed investigation at the later stage of the research (namely, when stability of the proposed model is be tested).

2.2. General characteristics of household cash flow model and consumption-bequest optimization

The largest and no doubt the most complicated building block of the whole financial plan optimization model is the model of household cash flows, also referred to as *household consumption model*. Its integral part is a value function reflecting preferences of the household. It is used as a goal function in household cash flow optimization. The value function is described in more details in Section 3, whereas main characteristics of the model are presented below and in Subsection 1.3.

- **Decision variables**

Decision variables with respect to which plan is optimized are: (a) proportion between consumption and investments and (b) division of total investments of the household between the two main household members. Proportions of the two persons in total investments are important. This is, among others, because life annuity of one person vanishes with death of this person and cannot be inherited. If the person who had bought higher life annuity died first, it would have much more severe consequences for the household finance than if it was the person who had bought lower annuity.

- **Incomes, consumption, investments**

The model is based on consumption utility, but the main diagnostics of financial situation throughout the whole life cycle of the household is the cumulated net cash flow (cumulated surplus). This is because the financial plan assumes preservation of some predefined standards of living. This means constant consumption in real terms (or constant growth of consumption in real terms).

Thus, consumption is not necessarily the whole difference between incomes and investments. The main types of cash flows in the model are: (a) basic incomes (without investment liquidation, etc.), (b) costs (basic consumption), then – dependent on g oals to be financed – also (c) cash flows resulting from pre-financing and post-financing of goals (investments, credit repayments, etc.). The difference between incomes and the sum of consumption and investments (and instalments) gives (d) the net cash flow. It cumulates over time. One of constraints imposed on financial plan is to secure household liquidity, thus not to let the cumulated net cash flow to fall below zero (the net cash flow of a given period may occasionally be negative if there is a potential to cover the shortfall in the future). Liquidation of investments, as well as transfer of credit capital to the household are additional incomes. Expenditures on realization of goals are additional elements of consumption, but they are treated separately from basic consumption. The separate treatment consists in calculating utility only of basic consumption. With a bequest motive, also the potential to leave bequest is taken into account in utility calculation.

- **Consumption-investment trade-off and risk aversion**

The decision about consumption determines the standard of live. The higher standard of life, other things unchanged, the lower value of the capital left unutilized. A need arises to find a trade-off between consumption in pre-retirement period and safety of consumption in retirement period. It is assumed that the sum of cumulated investments and cumulated net cash flow available after retirement must be sufficient to generate pay-offs that fill retirement gap. Retirement gap is understood as the difference between the last job income and retirement from compulsory public pension system. But the question for how long it should be sufficient is open to the decision of the household members. It depends on r isk aversion of the household. A simple way of grasping the notion of longevity risk aversion is asking these persons how many years after the expected time of death of the one who is expected to live longer a potential threat of permanent financial shortfall seems too abstract to be a cause for concern.

- **Role of bequest motive**

The higher cumulated surplus the better protection against longevity (and also premature death) risk. On the other hand, leaving any surplus or unutilized investment after the last household member dies may make sense only if the household wants to leave a l egacy to someone. Otherwise, it would be a suboptimal solution. The task of finding a trade-off between safety and economic efficiency of capital utilization will be different for the case with and without bequest motive. This difference would be particularly clear if the household showed no risk aversion at all. Then, for the case without the bequest motive the optimal plan would be such that its expected trajectory of cumulated surplus shrinks to zero at the date of the expected death of the last household member.

The household with no risk aversion, yet intending to leave some bequest, might, in turn, accept some unutilized capital at the end of their lifetime.

- **Changing parameters and constraints**

There are some quantities whose values may be changed in plan revision mode, which are not, however, decision variables. These are parameters and constraints that are subject to verification and adjustments by decision of the household. Main constraints that are adjusted in this way include financial goals. On the one hand goals must be met fully and on time. On the other hand, if this condition comes out to be infeasible (negative values in any point of the expected cumulated surplus trajectory), then goals are revised.

2.3. Input and output

Listed below are input and output arguments of the household consumption model. An important part of the model is the cash flow optimization procedure based on g oal function described in Section 3. For cash flow optimization, the starting values of decision variables are the main input. The output comprises: the optimum values of decision variables, the maximum of the goal function obtained as a result, and the expected trajectory of cumulated net cash flow for the optimal solution. The decision variables are: assumed consumption at the moment t_0 (C_{a_0}) and proportion of investments in private pension plans by Person 1 and Person 2 ($v_1, v_2 = 1 - v_1$). Apart from decision variables all initial values of variables and all parameters of the household cash flow model are certainly also the input of the optimization procedure.

- **Input:**
  - Age at t_0: $x_0^{(1)}, x_0^{(2)}$,
  - Retirement age: $zR1 = z(R1;\ x_0^{(1)}), zR2 = z(R2;\ x_0^{(2)}$,
    where $R1$ and $R2$ are retirement dates, $z(t, x)$ is age at the moment t of a person who was x years old at t_0,
  - Expected length of life at t_0: $E\left(D\middle|z\left(t_0;\ x_0^{(1)}\right)\right), E\left(D\middle|z\left(t_0;\ x_0^{(2)}\right)\right)$,
  - Income at t_0: $Ic_0^{(1)}, Ic_0^{(2)}$,$Ic_0^{(c)}$,
  - Income growth rate: g_1, g_2, g_c,
  - Replacement rate: χ,
  - Constant common consumption at t_0: FC,
  - Individual consumption at t_0: $VC_0^{(1)}, VC_0^{(2)}$,
  - Minimum acceptable consumption in any period: C_{min},
  - Consumption growth rate: h_{FC}, h_1 , h_2),
  - Proportion of investments in private pension plans by Person 1 and Person 2: $v_1, v_2 = 1 - v_1$,
  - Return on investment: r_{Iv},

- Return on “uninvested” surplus: r_{Spl},
- If with other goals (other than retirement):
  - Other goals ($\boldsymbol{G}$),

  where:

$$\boldsymbol{G} = [G_1, \dots, G_n] = \begin{bmatrix} T_1 & \cdots & T_n \\ M_1 & \cdots & M_n \end{bmatrix},$$

  $G_j = \begin{bmatrix} T_j \\ M_j \end{bmatrix}$ – j-th goal (denoted also as: $G_j = (T_j, M_j)$),

  T_j, M_j - planned time and magnitude of goal j,

  - Available investment programs for financing other goals than retirement ($L = [L_1 \quad \cdots \quad L_m]$).
  - Information about assignment of goals to financial programs (for explanation why it is input and not the output of the cash flow optimization task, refer to Section 4);

- **Output:**

  1) Direct:
  - Trajectories of consumption process,
  - Trajectories of surplus process,

  2) Indirect:
  - Income process,
  - Consumption process,
  - Cumulated investment process;

- **Relationships between some chosen input positions and basic household cash flows:**

  Consumption may be divided into three components:
  - Common consumption (fixed and not attributed to any particular person),
  - Consumption of Person 1,
  - Consumption of Person 2.

  Division of consumption between household members is vital for determining their contributions to private pension investment programs. The programs are separated and they do not depend on each other, however, if a person dies before retirement age, the amassed capital is transferred to the other one.

  Total consumption and savings of the household are given as (eq. 1, 2):

  - Assumed consumption:

$$C_{a_t} \equiv VC_t^{(1)} + VC_t^{(2)} + FC \tag{1}$$

  where: C_{a_t} – assumed consumption, $VC_t^{(i)}$ – variable costs assigned to i-th person, FC – fixed costs of the household.

- Savings (difference between incomes and consumption):

$$S_t = Ic_t - C_{a_t} = Ic_t^{(1)} + Ic_t^{(2)} + Ic_t^{(c)} - VC_t^{(1)} - VC_t^{(2)} - FC \tag{2}$$

where: Ic_t – joint income at the moment t, $Ic_t^{(1)}$ – income of the first person, $Ic_t^{(2)}$ – income of the second person, $Ic_t^{(c)}$ – income of the household that is not assigned to any person (e.g.: an income from renting out a real estate being a part of conjugal community).

Under the assumptions of the model, consumption needs are fixed or deterministically dependent on the life-cycle phase. Income, whether from labour or retirement, is either consumed or, in part that exceeds consumption needs, constitutes unconsumed and uninvested surplus. It is certainly also possible that the income of a given period does not cover consumption needs.

- Surplus – uninvested part of savings of a given period (eq. 3):

$$\begin{aligned} NCF_t &= S_t - Iv_t = Ic_t - C_{a_t} - Iv_t = \\ &= Ic_t^{(1)} + Ic_t^{(2)} + Ic_t^{(c)} - VC_t^{(1)} - VC_t^{(2)} - FC - Iv_t^{(1)} - Iv_t^{(2)} \\ &\qquad - Iv_t^{(c)} \end{aligned} \tag{3}$$

$$\begin{aligned} &(Ic_t = Ic_t^{(1)} + Ic_t^{(2)} + Ic_t^{(c)}; \\ &Iv_t = Iv_t^{(1)} + Iv_t^{(2)} + Iv_t^{(c)}; \\ &\text{if } t > R_i, \text{ then } Ic_t^{(i)} = Icb_t^{(i)} + Icc_t^{(i)}) \end{aligned}$$

where: Iv_t – investments of the household in period t, $Iv_t^{(1)}$ – investments of the first person in period t, $Iv_t^{(2)}$ – investments of the second person in period t, $Iv_t^{(c)}$ – investments of the household that are not assigned to any person in period t; moreover: $Icb_t^{(i)}$ – i-th person retirement income from a public pension system (all pillars included), $Icc_t^{(i)}$ – i-th person retirement income from private pension plan(s), R_i – retirement date of person i.

- Cumulated surplus – cumulated net cash flow (eq. 4):

$$CNCF_t = \sum_{\tau=0}^{t-1} NCF_\tau \tag{4}$$

- Maximum feasible consumption (eq. 5):

$$C_{ft}^* = Ic_t - Iv_t \tag{5}$$

(no cumulated surplus would be generated then because surplus of a given period would be consumed).

- Consumption that can be actually realized at a given moment t, assuming that until the moment only the assumed consumption was realized (eq. 6):

$$C_{f_t}^{*} = Ic_t + CNCF_t - Iv_t \tag{6}$$

- Consumption as understood in this model (i.e. assumed consumption, but up to the level that may be actually realized (eq. 7):

$$C_t = \min\{C_{a_t}, Ic_t + CNCF_t - Iv_t\} = \min\{C_{a_t}, C_{f_t}^{*}\} \tag{7}$$

or equivalently (eq. 8):

$$C_t = C_{a_t} + \min\{0, CNCF_t\} \tag{8}$$

In the formulas 6 and 7 there are no direct references to any further detailed decomposition of costs, incomes and investments. But it is important, after all, to be able to recognize individual contribution of each person to the total net cash flow of the household. This allows modelling the impact of stochastic elements of the model (namely, of the dates of person 1 or 2 deaths – $D1$ and $D2$).

3. Financial goals

Besides consumption sustaining, ensuring realization of the goals is the reason for which the financial plan is constructed. The goals differ in size, timing and other characteristic. In this section retirement-type goals and other goals are discussed.

3.1. Main financial goals of the household

The basic version of the model assumes only two main financial goals: retirement and bequest. These two goals have their unique feature – they cannot be post-financed. Therefore, the only way to realize them is to build up sufficient capital over the years. Retirement capital, as well as bequest capital, are usually very high in comparison to monthly income of the household. Thus, the earlier saving and investing are started the better.

The classical approach to consumption optimization assumes that: a) retirement income of the household should be at least as high as total consumption of the household, and b) individual retirement income of the household member should not be lower than his or her individual financial needs in retirement. This approach would be safe indeed, but rather inefficient due to overlapping coverage of household fixed costs, resulting in a considerable unutilized surplus. Neglecting this surplus would, in turn, lead to overestimating retirement capital needs and, consistently, paying unnecessarily high contributions to private pension plans in pre-retirement period. It is possible to propose such investment mix that would be less expensive than traditional approach, but would result in more risky retirement income. After all, it should be chosen so that it is suited to preference structure of household members. Generally speaking, the solutions differ in how much of household fixed and variable costs is covered by retirement income and in what proportions household members participate in them (comp. Feldman, Pietrzyk, Rokita, 2014a).

Emphasis should be also put on building the capital for bequest. It is modelled here as cumulated surplus (comp. eq. 4) that remains at the time of death of the last household member.

Taking bequest motive into consideration is necessary in this approach since the consumption may be the same in plans that differ much in respect of the term structure of cumulated surplus, and thus – financial situations of the household. Only shortfalls in cumulated surplus, driving consumption below its assumed value, would be visible for the utility of consumption. If, however, the last living household member dies, the uninvested and unconsumed surplus becomes visible in the form of bequest. Because this may happen with some probability at any moment, the value function takes account of cumulated surplus along the whole planning period.

These two above-mentioned financial goals are put in the centre of the model not only for their magnitude, but also because they usually are the last and often underestimated financial goals of the household. It is also obvious that the ability of achieving these goals depends significantly on household ability to save money. In most cases households spend their savings on some durable goods or other unplanned expenses, exchanging long term utility of sustaining high consumption level for short term utility of unplanned additional consumption (including realization of additional goals). However, households do not make a fully conscious choice in this respect. First of all, the decision makers often lack skills to estimate their retirement capital needs. Secondly, they often neglect the bequest motive and treat bequest as their estate. Thirdly, they are unaware of how the additional (unnecessary) expenditures affect their future financial situation.

3.2. Subordinate financial goals

There are two most commonly used approaches to determining financial goals: age based and life-event based (Nissenbaum, Raasch and Ratner, 2004). The first approach assumes financial goals are strictly dependent on the age of a decision maker. For instance, a 25-year old single male has different needs in terms of retirement planning than 40-year old male. The second approach focuses rather on the needs arising from particular events, determining some important elements of a decision maker's life situation. A single person usually has no bequest motive, while parents of two teenage children have a vital need to leave some legacy. Both approaches are justified to some extent and, in fact, result in similar outcome.

Apart from retirement and bequest, households have a wide variety of other financial goals. The most common include:

- Getting married,

- Buying a house,
- Raising a family,
- Funding education for children,
- Purchasing durable goods of high value.

Unlike retirement or bequest, all these financial goals may be post-financed. For the purposes of personal financial planning each financial goal has to be described by at least two characteristics: time of occurrence and value (assuming that goals are deterministic themselves). However, most decision makers do not have sufficient knowledge to determine these parameters. Firstly, due to lack of long-term planning (majority of individuals just do not plan). Secondly, because of insufficient information. Moreover, some of the parameters that need to be taken into consideration change in longer run. And they are in fact stochastic. For example, the question may arise of how real estate prices will change in the next 5 years.

It is important to point out a very significant difference between expenditures on the goal "child" (or "children") and funding education for children, on the one side, and realization of other financial goals, on the other. In most cases realization of a financial goal is an event occurring at particular point in time (e.g., purchase of a house, car, etc.). When it comes to raising children, the "realization" of that financial goal lasts in time. Therefore, it is assumed here that expenditures associated with children are treated as an increase in consumption. In order to stay in conformity with equation 8, these expenses increase the basic consumption by particular percentage (given as a model parameter) as long as the child remains in the household.

4. Goal function of the household

The goal function (value function) is intended to take into account both utility of consumption and utility of bequest (unconsumed and uninvested cumulated financial surplus). Utility function used for consumption and bequest is identical; just different arguments are put in. Apart from probabilities and discounting factors, these component utilities are multiplied by factors depending, among others, on attitude towards risk and bequest motive. The key concept of the goal function definition lies in these multipliers. One period value function is the weighted sum of component utilities. The goal function for the planning period has a form of expected discounted utility.

4.1. Utility

Household utility is split between the utility of consumption $U(C)$ and utility of bequest $U(B)$.

Moreover, utility of consumption is divided in two parts, with respect to time: the period before and after the expected death (see point 3.2).

Since the surplus, by definition, is what has not been consumed, it is not taken into account by utility of consumption. The surplus cumulated in previous periods may be partially consumed if current incomes are not sufficient to cover current expenses, but the utility of consumption does not recognize the sources from which the consumption is financed. This is why utility of consumption alone would be insufficient in this model.

Given other conditions and constraints (like financial goals and their financing) unchanged, the higher consumption the lower surplus left to build up the cumulated net cash flow. Since these two aspects of the financial plan are strictly contradictive, there has to be some trade-off between them. The trade-off is expressed with the following weights (eq. 9):

$$\alpha = 1 - \beta \tag{9}$$

where:

α – consumption preference parameter, β – bequest preference parameter.

Furthermore, the intertemporal consumption choice demands to discount the utility at some rate r_C. It is obvious that the utility of bequest should be also discounted, but at some other rate r_B. The relation between these rates should be (eq. 10):

$$r_B < r_C \tag{10}$$

The discount rate for the bequest has to be smaller because the household can postpone the realization of the bequest motive or even give it up, w hile the consumption at minimal level has to be achieved.

4.2. Risk aversion and optimization area

As it has been mentioned in the introduction, not all scenarios of the survival process are taken into account. The modification of the way the survival of two persons is worked into the model is twofold. The main concept of the simplification consists in considering only some periods before and after the expected time of death. This delineates a r ange in which premature death or unexpected longevity is recognized to be a concern for the household members. The *range of concern* defined in this way will be set in accordance to life-length risk aversion. In this approach optimization for a single person would be performed for the values of potential death time from within the interval of (eq. 11):

$$D_i^* \in [E(Di) - \gamma^*; E(Di) + \delta^*] \tag{11}$$

It is worth emphasizing that the range of concern should not be confused with domain of optimization, because it is not a set of decision variable values.

The second simplification is in probabilities used. Survival probabilities are not conditional probabilities for a given day but unconditional probabilities (conditional under the condition of surviving until the moment t_0). The second simplification is – from the point of view of the main idea – just a side issue, and may be refrained from in further stages of the research. One just needs to remember that when attempting to make the model more dynamic the whole history of the survival process would need to be considered for each scenario and each period (not just the state of the household in the preceding period). This is due to complicated interdependences between quantities used for cash flow calculation and cumulative nature of the net cash flow process.

For the household, the range of concern is a rectangle of (eq. 12):

$$Range_{Hh} = [E(D1) - \gamma^*; E(D1) + \delta^*] \times [E(D2) - \gamma^*; E(D2) + \delta^*] \qquad (12)$$

where:

γ^* – premature death risk aversion parameter (number of years that the household takes into consideration),

δ^* – longevity risk aversion parameter (also interpreted as the number of years),

There is one parameter γ^* and one δ^*, characteristic of the household, not individual person.

On the basis of risk aversion parameters (γ^* and δ^*), risk aversions measures, $\delta(t)$ and $\gamma(t)$, are constructed. These are then used as multipliers by which utility of consumption for periods before and after the expected time of the end of the household is multiplied. They are defined so that the premature-death risk aversion multiplier is the higher the earlier moment before the expected time of the end of the household decreases to 1 at the expected time of the end of the household, to decay afterwards, whereas the longevity risk multiplier reaches unity at the expected time of death and then increases with time. A proposed formal definition that holds these properties is given by eq. 13 and 14:

$$\gamma(t) = \begin{cases} \left(\frac{1}{1+\gamma^*}\right)^{\left(\frac{t-E(D)}{E(D)}\right)} & t \leq E(D) \\ 0 & t > E(D) \end{cases} \qquad (13)$$

$$\delta(t) = \begin{cases} (1+\delta^*)^{\left(\frac{t-E(D)}{\delta^*}\right)} & t > E(D) \\ 0 & t \leq E(D) \end{cases} \qquad (14)$$

where:

$E(D)$ is unconditional expected time of the end of the household, defined by eq. 15:

$$E(D) = max(E(D1), E(D2)) \qquad (15)$$

and $E(Di) = E(Di|Di > t_0)$ is unconditional expected time of death of Person i.

One of the merits of defining lifetime risk aversion in the form of γ^* and δ^* is that these parameters do not require estimation nor detailed inquiry. Their interpretation seems to be sufficiently intuitive for household members just to declare their values.

The optimization procedure differs significantly from the most commonly used ones. In classical approaches consumption is optimized across the whole life cycle of a decision maker. But that might result in excess saving and amassing too much retirement capital. The household would have to decrease its consumption in early years in order to fulfil optimization constraints at every point in time and for each combination of individual survival scenarios, even for those very unlikely (e.g., a man dies at the age of 25 and his wife lives up to 95 – possible, but it would be very likely that the young widow would find another lifetime partner and raise a new household for which the old financial plan would be utterly irrelevant). The model presented here focuses on the range of concern that corresponds to probabilities recognized arbitrarily by both decision makers as significant. Secondly, optimization over the range of all possible combination of dates when household members may die is very computationally-intensive. The number of possibilities increases proportionally to the square of the number of years taken into account.

4.3. Goal function

The goal function used here is based on the concept of expected discounted utility of consumption. It differs, however, from the one used in classical life cycle models, like that of Yaari (1965). It has two components: the first one is responsible for utility of consumption and the other reflects utility of unconsumed surplus (bequest). Both are joint utilities of the whole household. This is a necessary condition if one common life-long financial plan is to be constructed.

The goal function presented in eq. 16 is an expansion of that proposed by Feldman, Pietrzyk and Rokita (2014a). It is suited to the model of two-person household with rectangular range of concern.

$$V = \sum_{D_2^*=E(D2)-\gamma^*}^{E(D2)+\delta^*} \sum_{D_1^*=E(D1)-\gamma^*}^{E(D1)+\delta^*} p \left[\begin{array}{c} \alpha \left(\sum_{t=0}^{max\{D_1^*,D_2^*\}} \frac{1}{(1+r_c)^t} u\big(C(t;D_1^*,D_2^*)\big)\big(\gamma(t)+\delta(t)\big) \right) + \\ \beta \frac{1}{(1+r_B)^{max\{D_1^*;D_1^*\}}} u\big(B(max\{D_1^*,D_2^*\};D_1^*,D_2^*)\big) \end{array} \right] \longrightarrow max \tag{16}$$

where:

$u(.)$ – utility function (the same in all segments of the formula),
$C(t;D_1^*,D_2^*)$ – consumption at the moment t,
$B(t;D_1^*,D_2^*)$ – bequest (cumulated investments and surplus of both household members at the moment t,
$\gamma(t)$ – premature death risk aversion measure (depends on γ^*),
$\delta(t)$ – longevity risk aversion measure (depends on δ^*),
p – probability that at least one person is alive,

r_C – discount rate of consumption,
r_B – discount rate of bequest.

5. Technique of financing household's goals

The main purpose for which households prepare their financial plans is to achieve all their financial goals. Due to a wide variety of ways in which the goals may be financed and the fact that their examination and comparison is a complex task, households prefer ready-made investment products. One of the most common forms is systematic investment plans offered by mutual funds.

This solution is noticeably elastic and may be used for financing every financial goal. Although systematic investments require some discipline and may seem psychologically hard, they have many advantages over post-financing. Only part of the financial goals may be financed by debt (housing, cars, etc.). Moreover, taking loans is often more expensive (though easier in many respects). Households also may face limitations according to their credit standing.

The assumption that all goals must be realized is still sustained. Households usually take into consideration only the goals that are planned for the near future. Such goals as retirement or bequest are neglected (more or less on purpose). Such attitude affects significantly household's ability to realize all its financial goals. It may even make some of them unattainable. It is recommended that household members work all important goals they have into their financial plan.

That being so, the household faces the problem of how to finance n different goals when m possible investment programs are available. The problem is multidimensional not only because of the number of financial goals, but also for wide variety of parameters to be taken into consideration (i.e., rates of return, indexation, fees and charges, allocation rate, etc.).

It is proposed here to facilitate this process with an algorithm that seems to be indeed easy to understand and use. Such technique (algorithm) has to give a result that fulfils the following criteria:

- Expected term structure of cumulated net cash flow (obtained after application of the financing strategy selected by the algorithm) should be in compliance with life-length risk aversion and bequest motive.
- Financial liquidity of the household must be sustained.
- Net cash flow volatility is minimized.
- Costs of the investment plan combination are minimized.

We also assume that one program may be used to provide cash to cover more than one goal and the order may be other than sequential.

For illustrative reasons, let us assume that the household has three financial goals and there are three different investment programs available on the market. Each of the programs reflects the same level of risk. The household wants to find an optimal set of investment programs along with information which program is destined to cover which financial goal.

The investment mix selection is a multi-step process, whose three main steps are:

1) Minimization of contribution to investment programs.
2) Selection of effective investment mix.
3) Cash flow term structure fitting.

Before an analysis of concrete systematic-investment products is started, *general schemes* of financing are identified. A general scheme may be defined as a $(2 \times n)$ matrix $\boldsymbol{S}$, in the first row which includes indices denoting goals, whereas the second row is constructed using the following algorithm:

a) put any symbol, say "A", in the first field of the second row of matrix $\boldsymbol{S}$: $S_{2,1} = A$,

   a general way of financing goal 1 is then: $S_{:,1} = \begin{bmatrix} 1 \\ A \end{bmatrix}$;

b) for $j := 2\ to\ n$ repeat the following:

   c.1) if goal j is to be financed with the same program as some of the goals considered so far (i.e. goals: $[1, \dots, j-1]$), let it be a goal υ, $(1 \le \upsilon \le j-1)$, whose general way of financing is: $S_\upsilon = \begin{bmatrix} \upsilon \\ \Upsilon \end{bmatrix}$, then substitute

   $S_{2,j} = \Upsilon$,

   a general way of financing goals $[1, \dots, j]$ is then: $S_{:,[1:j]} = \begin{bmatrix} 1 \cdots \upsilon \cdots j \\ A \cdots \Upsilon \cdots \Upsilon \end{bmatrix}$.

   c.2) otherwise (goal j is intended to be financed with another program), assign a symbol that has not been used yet, say Ξ;, to goal j; thus, substitute:

   $S_{2,j} = \Xi$,

   a general way of financing goals $[1, \dots, j]$, is then: $S_{:,[1:j]} = \begin{bmatrix} 1 \cdots \upsilon \cdots j \\ A \cdots \Upsilon \cdots \Xi \end{bmatrix}$;

For 3 goals and 3 programs 4 general schemes of financing are possible:

$$\boldsymbol{S1} = \begin{bmatrix} 1\,2\,3 \\ AAA \end{bmatrix}, \ \boldsymbol{S2} = \begin{bmatrix} 1\,2\,3 \\ AAB \end{bmatrix}, \ \boldsymbol{S3} = \begin{bmatrix} 1\,2\,3 \\ ABB \end{bmatrix}, \ \boldsymbol{S4} = \begin{bmatrix} 1\,2\,3 \\ ABC \end{bmatrix}.$$

It should be stressed that sequences ABB and ACC are identical. The same refers, for example, to sequences AAB and AAC. General schemes inform whether goals are to be financed by the same or different programs, not determining, yet, which concrete programs might there be.

Then, the investment mix selection algorithm is executed.

In the first step (*Minimization of contribution*) we find the minimum contribution for each possible combination of goals and available investment programs. Due to incomparability of cash flow term structures of different investment programs, contribution is minimized for each possible general program scheme separately. Investment mix for a given general scheme that minimizes contribution is further on referred to as *efficient investment mix*.

Table 1. Types of investment schemes. Scheme depends on the number of different investment programs used to finance goals and the structure of which program is used to finance which goal

INVESTMENT MIX	GOAL 1	GOAL 2	GOAL 3	SCHEME
Investment mix 1	**Program 1**	**Program 1**	**Program 1**	AAA
Investment mix 2	Program 2	Program 2	Program 2	
Investment mix 3	Program 3	Program 3	Program 3	
Investment mix 4	Program 1	Program 1	Program 2	AAB
Investment mix 5	Program 1	Program 1	Program 3	
Investment mix 6	**Program 2**	**Program 2**	**Program 1**	
Investment mix 7	Program 1	Program 2	Program 2	ABB
Investment mix 8	**Program 1**	**Program 3**	**Program 3**	
Investment mix 9	Program 2	Program 1	Program 1	
Investment mix 10	**Program 1**	**Program 2**	**Program 3**	ABC
Investment mix 11	Program 2	Program 3	Program 1	
Investment mix 12	Program 3	Program 1	Program 2	

In the second step (*Selection of efficient investment mix*) efficient solutions are picked. A solution is said to be efficient if it requires the minimum contribution to finance household goals from amongst investment program combinations belonging to the same general scheme (compare Table 1).

In the third step (*Cash flow term structure fitting*) efficient solutions selected at the second stage are put into to the model of household cash flow. The optimal solution is such that the corresponding term structure of the cumulated net cash flow of the household best fits the preferred one. Two alternative approaches to evaluate the fit are proposed:

A. least squares method,
B. maximization of goal function by putting each of the efficient investment mixes into the model of household cash flow discussed in Sections 1-3, finding the value of the goal function (compare Section 3) for each of them, and picking up the one that maximizes the goal function maximum.

The optimal solution has to meet the following conditions: (1) consumption has to be higher than the minimum (set by the household), (2) at any point in time there has to be some nonnegative cumulated surplus, (3) goal function is to be maximized.

6. Numerical example

The following example shows in work the results of the algorithm described above. Let us assume that the future cumulated surplus term structure of the household (for the next 30 years) is given as below (Figure 1 a):

Cumulated surplus term structure before the financial goals

Cash flow term structure of financial goals

Cumulated surplus term structure after realization of financial goals

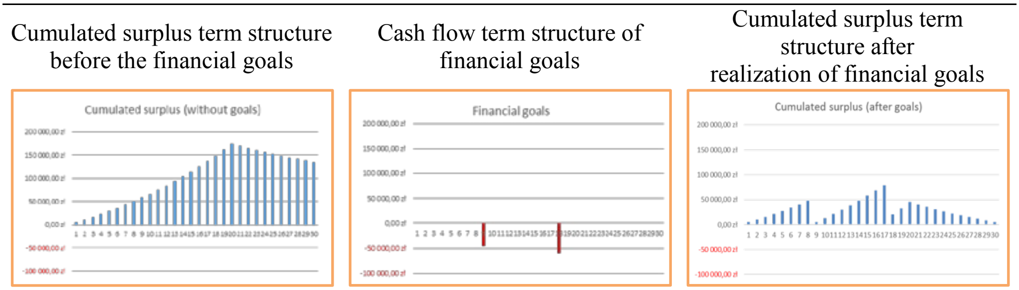

Figure 1. a) Figure 1. b) Figure 1. c)

Figure 1. Cumulated surplus term structure of the household before and after realization of financial goals

The household has two financial goals that are planned to be achieved in year 9 and 18. The size of these goals is known (or can be easily estimated) (compare Figure 1 b).

If the household decides to realize its financial goals from cumulated surplus, then the final term structure of cash flow will look like in the Figure 1 c.

The household may, however, invest some part of its surplus into an investment program. Let us assume that there are two programs available on the market. That gives two possible schemes to be analyzed (Figure 2 a, Figure 2 b). The first (scheme AB) uses two programs separately to finance two goals, and the second (scheme AA) uses one program to finance both goals. Negative cash flows in Figure 2 a and Figure 2 b are contributions to investment programs. Positive ones are pay-offs from the programs.

Realization of two financial goals with two different investment programs (scheme AB)

Realization of two financial goals with one investment programs (scheme AA)

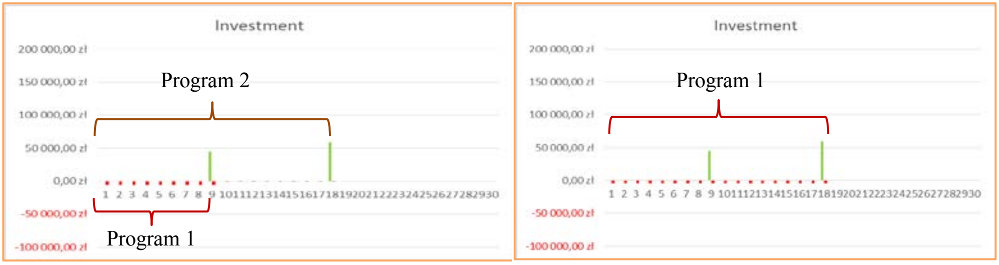

Figure 2. a) Figure 2. b)

Cumulated surplus term structure comparison

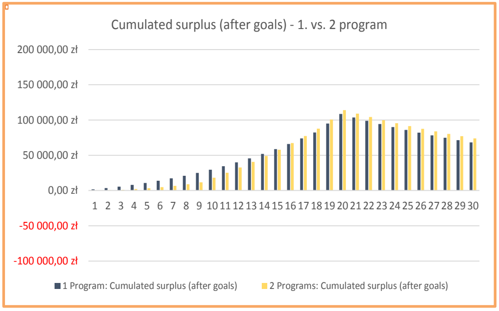

Figure 2. c)

Figure 2. Impact of different investment schemes on cumulated surplus term structure

Dependent on the scheme one uses, different cash flow term structures are obtained. The comparison of these structures is presented in Figure 2 c)

Then, both structures are compared with the optimal trajectory and the final result is given.

The optimal trajectory might be estimated for strategic asset allocation that reflects the risk level of investment programs, but not concrete programs themselves. Then the term structure of cash flow is calculated and compared to the cash flow term structure resulting from investment in particular efficient investment plan.

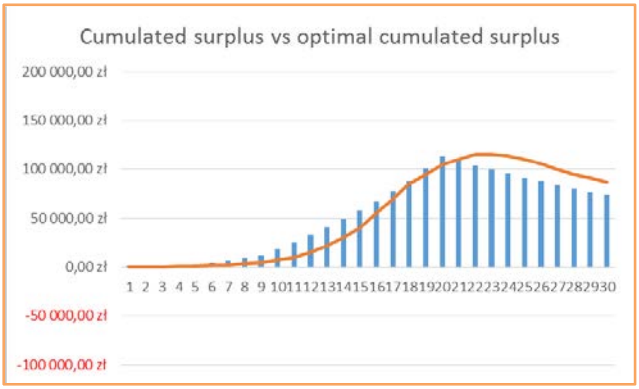

Figure 3. Cumulated surplus comparison with optimal trajectory

Another approach would be just calculating the value of household goal function for both program mixes and selecting the one with higher value. All conditions have been listed at the end of Section 4.

7. Technical issues

The formula of goal function (eq. 16) presented in section 3 does not specify in details the analytical form of the utility function. In the numerical example discussed in this article a sq uare root utility was used. This was, however, modified in such a way that it took on v alue zero for scenarios in which cumulated surplus fell below zero at any point in time. This, certainly, does not need to drive the goal function to zero because the goal is a sum of probability-weighted discounted expected values of utility for all scenarios within the range of concern. The argument for such solution is that a scenario cannot be "partially" satisfactory if it guarantees high level of consumption in some period and then leads to permanent shortfall (i.e. practical bankruptcy of the household). Within the bunch of scenarios there may by one or more such zero-utility scenarios. Their influence on the goal function depends on their probabilities.

Such construction of the goal function causes some technical inconvenience. The goal function becomes indifferentiable on vast parts of its domain. Moreover, there are not only unsmooth jumps in its value, but also local extremes. There is, however, a simple way to overcome this problem without reaching for very advanced optimization techniques. The goal function shows problematic properties mostly along one dimension, namely – the decision variable describing division of total investment between household members. Along the second one, that is consumption-investment proportion, it behaves in a much more conventional way. It is continuous, differentiable and unimodal up to the maximum, though indiferrentiable and showing local extremes after reaching the global maximum.

Along the first dimension (division of investment contributions) the function is sliced into a finite number of cross-sections. The range between 100% and 0% of the total household investment allocated to Person 1 may be divided into any number of scenarios. Then, for each of the slices a maximum along the second dimension is searched for.

It may be observed that the cross-section of the goal function along the second dimension (consumption-investment) always shows the following property: it is differentiable and increasing until it reaches global maximum for the given slice, then a downwards jump is encountered and then there may be a local maximum (always lower than the first maximum - walking from the left - for this particular slice), followed with a rapid drop. At this stage of analysis, continuous optimization may be used under the condition that the optimization algorithm starts from the lowest values of consumption and searches for the maximum of the goal in the direction of growing consumption.

Then, the maximum of maxima for each slice is taken as the global maximum for the whole goal function.

8. Summary

The model presented here involves some original approaches and solutions, and sheds some new light on household consumption optimization. It focuses on the household, not on a single decision maker only. The optimization area is strictly dependent on the risk aversion of household members and is narrowed to the most probable scenarios. This results in higher optimal consumption for the household than it would be derived from models taking the whole lifespan of decision makers into account. The risk aversion measures are very intuitive and their interpretation, calibration and use by decision makers is straightforward. Different discounting rates for utility are used. The same goal function as used in optimization model allows comparing different cash flow term structures. Thus, it may be used to facilitate choosing from amongst incomparable investment products.

Further research will focus on expanding its application to stochastic behaviour of financial goals. In particular, such goals as children should be treated in this way because of stochastic nature of a child's birth (conditional on planned time).

Furthermore, other types of risk than just risk related to length of life will be analyzed. In particular, risk connected with investments, mainly market risk (e.g.: interest rate risk, stock price risk, etc.) will be taken into consideration.

Together with taking account of investment risk, also risk of human capital will be a natural object of investigation. Adopting the approach by Bodie, Merton and Samuelson (1992), in which risk of human capital, increasing with age, is offset by decreasing riskiness of investments, may be useful in the next stages of research.

Also stability of the model will be analyzed. This is not only sensitivity of optimization results to changes of parameters that needs to be analyzed. What is also worth investigating is how the choice of the underlying survival process model will influence the final results.

Another area of research may be examining structure of hierarchy of financial goals and suggesting optimization procedures.

Acknowledgements

The research project was financed by The National Science Centre (NCN) grant, on the basis of the decision no. DEC-2012/05/B/HS4/04081.

REFERENCES

AINSLIE, G., (1991). Derivation of „Rational" Economic Behavior from Hyperbolic Discount Curves. The American Economic Review, 81(2), pp. 334–340.

AINSLIE, G., (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. Psychological Bulletin 82(4), pp. 463–496.

ANDO, A., MODIGLIANI, F., (1957). Tests of the Life Cycle Hypothesis of Saving: Comments and Suggestions. Oxford Institute of Statistics Bulletin, Vol. XIX (May), pp. 99–124.

BODIE, Z., DETEMPLE, J., OTRUBA, S., WALTER, S., (2004). Optimal Consumption-Portfolio Choices and Retirement Planning. Journal of Economic Dynamics and Control, 28, pp. 1115–1148.

BODIE, Z., MERTON, R. C., SAMUELSON, W. F., (1992). Labor Supply Flexibility and Portfolio Choice in a Life Cycle Model. Journal of Economic Dynamics and Control, 16(3-4), pp. 427–249.

BODIE, Z., TREUSSARD, J., WILLEN, P., (2007). The Theory of Life-Cycle Saving and Investing. [online] Federal Reserve Bank of Boston: Research Paper, no. 07-3. Available at: <http://ssrn.com/abstract=1002388> or <http://dx.doi.org/10.2139/ssrn.1002388> [Accessed 20 March 2013].

CARBONE, E., INFANTE, G., (2012). The Effect of a Short Planning Horizon on Intertemporal Consumption Choices. [online] LABSI: Research Paper. Available at: <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2187911> [Accessed: 28 March 2013].

DUS, I., MAURER, R., MITCHELL, O. S., (2004). Betting on Death and Capital Markets in Retirement: A Shortfall Risk Analysis of Life Annuities versus Phased Withdrawal Plans. [online] Goethe-Universität, Frankfurt am Main and The Wharton School, University of Pennsylvania: Working Paper. Available at: <http://www.actuaries.org/AFIR/Colloquia/Boston/Dus_Maurer_Mitchell.pdf > [Accessed: 27 March 2013].

FELDMAN, L., PIETRZYK, R., ROKITA, P., (2014a). A practical method of determining longevity and premature-death risk aversion in households and some proposals of its application, in: Spiliopoulou, M., Schmidt-Thieme, L., Janning, R. (Eds.), Data Analysis, Machine Learning and Knowledge Discovery, Studies in Classification, Data Analysis, and Knowledge Organization. Springer International Publishing, pp. 255–264.

FELDMAN, L., PIETRZYK, R., ROKITA, P., (2014b). General strategies to meet household pension goal versus longevity risk, in: Lisowski J., Łyskawa K. (Eds.), Insurance in view of longevity/old age risk. Poznań University of Economics Publishing House, pp. 37–49.

GEYER, A., HANKE, M., WEISSENSTEINER, A., (2009). Life-cycle asset allocation and consumption using stochastic linear programming. The Journal of Computational Finance, 12(4), pp. 29–50.

GOMPERTZ, B., (1825). On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. Philosophical Transactions of the Royal Society of London, 115, 513–585.

GONG, G., WEBB, A., (2008). Mortality Heterogeneity and the Distributional Consequences of Mandatory Annuitization. The Journal of Risk and Insurance, 75(4), pp. 1055–1079.

HUANG, H., MILEVSKY, M. A., (2011). Longevity Risk Aversion and Tax-Efficient Withdrawals. [online] SSRN. Available at: <http://ssrn.com/abstract=1961698> [Accessed: 22 March 2012].

HUANG, H., MILEVSKY, M. A., SALISBURY, T. S., (2012). Optimal Retirement Consumption with a Stochastic Force of Mortality. Insurance: Mathematics and Economics, 51(2), pp. 282–291.

KIRBY, K. N., HERRNSTEIN, R. J., (1995). Preference reversals due to myopic discounting of delayed reward. Psychological Science, 6(2), pp.83–89.

MAKEHAM, W. M., (1860). On the Law of Mortality and the Construction of Annuity Tables. Journal of the Institute of Actuaries and Assurance Magazine, 8, pp. 301–310.

MERTON, R. C., (1969). Lifetime portfolio selection under uncertainty: The continuous time case. The Review of Economics and Statistics, 51(3), pp. 247–257.

MERTON, R. C., (1971). Optimum consumption and portfolio rules in a continuous-time model. Journal of Economic Theory, 3(4), pp. 373–413.

MILEVSKY, M. A., HUANG, H., (2011). Spending Retirement on P lanet Vulcan: The Impact of Longevity Risk Aversion on Optimal Withdrawal Rates. Financial Analysts Journal, 67(2), pp. 45–58.

MODIGLIANI, F., BRUMBERG, R. H., (1954). Utility analysis and the consumption function: an interpretation of cross-section data. In: Kenneth K. Kurihara, ed. 1954. Post-Keynesian Economics. New Brunswick, NJ: Rutgers University Press. pp. 388–436.

NISSENBAUM, M., RAASCH, B. J., RATNER, C., (2004). Ernst & Young's Personal Financial Planning Guide. John Wiley & Sons, Inc. 237–522.

RICHARD, S. F., (1975). Optimal consumption, portfolio and life insurance rules for an uncertain lived individual in a continuous time model. Journal of Financial Economics, 2, pp. 187–203.

SCHOLZ, K., SESHADRI, A., (2012). Health and Wealth In a Lifecycle Model. [online] University of Wisconsin-Madison and NBER and University of Wisconsin-Madison: Working Paper. Available at http://www.ssc.wisc.edu/~scholz/Research/Health_and_Wealth_v16.pdf [Accessed: 23 Nov. 2013].

SUNDARESAN, S., ZAPATERO, F., (1997). Valuation, optimal asset allocation and retirement incentives of pension plans. Review of Financial Studies, 10, pp. 631–660.

YAARI, M. E., (1965). Uncertain Lifetime, Life Insurance and Theory of the Consumer. The Review of Economic Studies, 32(2), pp.137–150.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 269–282

INCOME INEQUALITY AND INCOME STRATIFICATION IN POLAND

Alina Jędrzejczak[1]

ABSTRACT

Income inequality refers to the degree of income differences among various individuals or segments of a population. When the population has been partitioned into subgroups, according to some criterion, one common application of inequality measures is evaluation of the relationship between inequality in the whole population and inequality in its constituent subgroups in order to work out the within and the between subgroups contributions to the overall inequality. In the paper selected decomposition methods of the well-known Gini concentration ratio were discussed and applied to the analysis of income distribution in Poland. The aim of the analysis was to verify to what extent the inequality in different subpopulations contributes to the overall income inequality in Poland and to what extent their members form distinct segments or strata. To provide the decomposition of the Gini index the population of households was partitioned into several socio-economic groups on the basis of the exclusive or primary source of maintenance. Moreover, the households were divided by economic regions using the Eurostat classification units NUTS 1 as well as by family type defined by the number of children.

Key words: income distribution, income inequality.

1. Introduction

In the analysis of income inequality it may be relevant to assign inequality contributions to various income components (such as labor income or property income) or to various population subgroups associated with socio-economic characteristics of individuals (age, sex, occupation, composition of their household, ethnic groups, etc.). Such an approach can be useful for social policy makers to better understand the influence of various socio-economic determinants on income levels and income inequality. In order to separate the within-groups inequality from the between-groups inequality a decomposable inequality measure has to be used. If the adopted inequality measure is additively

[1] Institute of Statistics and Demography, University of Lodz, Poland. E-mail: jedrzej@uni.lodz.pl.

decomposable, the overall inequality is equal to the sum of the within and between-groups inequality.

The Gini index is a well known and widely used synthetic inequality measure usually expressed in terms of the area under the Lorenz curve. In numerous works on income distribution it is considered the best single measure of income inequality (see e.g.: Morgan, 1962; Gastwirth, 1970), what is mainly due to its statistical properties. In contrast to many other inequality coefficients, measuring only the deviations from the mean and thus interlinking the concept of location with the concept of variability, the Gini index takes into account the income differences between each and every pair of individuals. It has also a clear economic interpretation and thus has been applied in various empirical studies and policy research. On the other hand, being sensitive to both the distribution of income and the distribution of ranks, the Gini index cannot be easily decomposed into two: between-groups and within-groups components. This property can be found a disadvantage of this index which was even claimed decomposable only when the subpopulations do not overlap (see: Shorrocks, 1984). Regardless of these difficulties, for the last 50 years a great effort has been made to specify the conditions under which the decomposition of the Gini coefficient is feasible and many interesting decompositions have been derived. Some of them provide us with the more complex but at the same time more informative tools for income inequality analysis than do many straightforward decompositions of additively decomposable inequality measures.

The first attempts to decompose the Gini index followed the classical Theil's approach and considered only two terms: the within-groups component and the between-groups component, the latter being generally based on the assumption that each individual receives the mean income of his own group. The pioneer Gini index decomposition by subgroups is due to Soltow (1960) who analyzed the effects of changes in education, age and occupation on income distribution. The first Gini index decomposition encompassing comparisons between pairs of subgroups is due to Bhattacharya and Mahalanobis (1967); actually the decomposition proposed by the authors refers first to the Gini mean difference Δ. The decomposition is based on a priori definition of the between-groups component, being the Gini mean difference evaluated among the subgroup means, and leaves the within term to be obtained as a residual.

Both the decompositions mentioned above were rather inadequate as they ignored the existence of overlapping as well as different variances and asymmetries of income distributions in subpopulations. In fact, when the groups ranges overlap the third component called "crossover term" or "interaction" arises, being rather difficult to interpret. The interaction term can be viewed as a measure of income stratification or the degree to which the incomes of different social groups cluster.

An interesting three-term decomposition and interpretation of the Gini coefficient was proposed by Pyatt (1976) in a game theory framework. Following the Pyatt idea, the Gini index can be perceived as an average gain to be expected if an individual had a choice between his own income or any other income selected at random from the population of income receivers. Pyatt split the Gini index into the sum of three non-negative terms: the first depends on the differences in mean incomes between subgroups and remains the only positive term in the special case when there is no variation within subgroups, the other two terms both depend on variation within subgroups. In particular, the second one depends on the Gini indices evaluated within each subgroup and the third term vanishes in the case when subgroups income ranges do not overlap, otherwise it is positive and measures the degree of overlapping. An analogous approach, based on matrix algebra, can be found in Silber (1989); the author decomposes the Gini index into the sum of the within, between and interaction terms giving a clear and intuitive interpretation to the latter in terms of individuals ranking. That "third component" was also discussed by Mehran (1975), Mookherjee and Shorrocks (1982), Yitzhaki and Lerman (1991), Deutsch and Silber (1999), to name only a few, what resulted in numerous interesting decomposition formulas. Some of them are computationally cumbersome and it is not always clear what meaningful interpretation each of the components has. Mehran defined "the third term" as interaction "*interpreted as a measure of income domination of one subgroup over the other apart from the differences between their mean incomes*". Yitzhaki and Lerman (1991), intended from a sociological point of view, proposed a decomposition of the Gini index into the sum of a within term, a between term, and a third term that accounts for subgroups stratification understood as "*a group's isolation from members of other groups*". The within- and between-group terms considered by the authors were based on the covariance formula so they are differently defined with respect to the ones considered above.

The most widespread approach to the decomposition of the Gini index that gives an important contribution to the understanding of the overlapping term was proposed by Dagum (1997). It introduces the concept of economic distance between distributions and relative economic affluence (REA) as an important element in the Gini index decomposition by subpopulation groups.

The objective of the paper is to discuss the most interesting decomposition procedures proposed by Dagum (1997) and Yitzhaki and Lerman (1991) and then apply them to the analysis of income inequality in Poland. The aim of the analysis was to verify to what extent the inequality in different subpopulations contributes to the overall income inequality in Poland and whether their members form distinct segments or strata.

2. The Gini index decomposition by subpopulations

The Gini index of inequality is usually defined by means of a geometric formula since it can be expressed as twice the area between the Lorenz curve and the straight line called the line of equal shares. The Gini index can also be seen as a relative dispersion measure when expressed by means of the mean difference

Δ - a dispersion measure which is defined as the average absolute difference between all possible pairs of observations. This concept can be called a statistical approach and was introduced by Gini (1912). It was subsequently used by many authors to derive various Gini index decompositions but the most widespread decomposition by subpopulations was undoubtedly proposed by Dagum (1997).

The starting point for this decomposition was the Gini index formula based on the Gini mean difference extended to the case of a population divided into *k* subpopulations (groups):

$$G = \frac{\Delta}{2\overline{Y}} = \sum_{r=1}^{n}\sum_{i=1}^{n} | Y_i - Y_r | \Big/ 2n^2\overline{Y} = \sum_{j=1}^{k}\sum_{h=1}^{k}\sum_{i=1}^{n_j}\sum_{r=1}^{n_h} | y_{ji} - y_{hr} | \Big/ 2n^2\overline{y} \tag{1}$$

The Gini index expressed in terms of the Gini mean difference can also be generalized for a two-populations case, measuring the between-populations (or intra-groups) inequality. Thus, the extended Gini index between groups *j* and *h* can be written as follows:

$$G_{jh} = \frac{\Delta_{jh}}{\overline{Y}_j + \overline{Y}_h} = \frac{1}{\overline{Y}_j + \overline{Y}_h}\sum_{i=1}^{n_j}\sum_{r=1}^{n_h} | y_{ji} - y_{hr} | \Big/ n_j n_h \tag{2}$$

where: Δ_{jh} - mean difference modified for two income distributions.

Dagum (1997) proved that the Gini ratio *G* for a population of economic units partitioned into *k* subpopulations n_j (j = 1,…, k) can be expressed as the weighted sum of the extended Gini ratios weighted by the products of the *j*-th group population share p_j and the *h*-th group income share s_h:

$$G = \sum_j \sum_h G_{jh} p_j s_h \tag{3}$$

Using the symmetry properties of G_{jh} and Δ_{jh} and the equation (3), the Gini index can be decomposed into two elements: the within G_w and gross-between G_{gb} inequality (Dagum, 1997):

$$G = \sum_{j=1}^{k} G_{jj} p_j s_j + \sum_{j=2}^{k}\sum_{h=1}^{j-1} G_{jh}(p_j s_h + p_h s_j) = G = G_w + G_{gb} \tag{4}$$

where: $G_{jj} = \frac{\Delta}{2\overline{y}_j} = \frac{1}{2\overline{y}_j}\sum_{r=1}^{n_h}\sum_{i=1}^{n_j} | y_{ji} - y_{jr} | \Big/ n_j^2$ is the Gini index for the subpopulation j $\overline{y}_j$ - mean income in group *j*, n$_j$ - frequency in group *j*.

As it can be easily noticed the Gini index provides an unusual “between-group” component. It measures the income inequality between each and every pair of subpopulations, whereas entropy and most of between-groups inequality measures yield only the income inequalities between the subpopulation means. The first component of the decomposition given by the formula (4) (G_w) describes the *contribution of the Gini inequality within subpopulations* to the total inequality of a population described by the Gini ratio *G*. The second component (G_{gb}) measures the *gross contribution of the extended Gini inequality between subpopulations* to the total Gini *G*. This component depends on the differences between subpopulations coming from both: *differences in mean* income levels and *differences in shape* (the populations differ in variance and asymmetry which implies that they have different inequality measures).

The income differences between the elements coming from various subgroups can be of the same or of opposite sign as the deviation in their corresponding means.

The interpretation of G_{gb} given above suggests the further decomposition of the Gini index by subgroups. The contribution of gross between-group inequality can be divided into two separate parts: the first one consistent with the differences between the means and the remaining part called transvariation:

$$G_{gb} = \sum_{j=2}^{k}\sum_{h=1}^{j-1} G_{jh}(p_j s_h + p_h s_j)D_{jh} + \sum_{j=2}^{k}\sum_{h=1}^{j-1} G_{jh}(p_j s_h + p_h s_j)(1 - D_{jh}) = G_b + G_t \tag{5}$$

G_b – the contribution of net between-groups inequality to the Gini index,
G_t – the contribution of ”transvariation”,
D_{jh} – “economic distance” ratio (Dagum, 1980).

The concept of transvariation (*transvariazione*) was originally introduced by Gini (1916) and it plays a crucial role in the Gini index decomposition by population subgroups. Transvariation between two populations exists when at least one income difference between individuals belonging to different groups has the sign opposite to the sign of the difference between their means. Obviously, the idea of transvariation is similar to the concept of distribution overlapping. The probability of transvariation can be simply defined (Gini, 1916) as the ratio of the actual number of transvarying pairs to its maximum. It takes values in the interval [0,1] and the more the two groups overlap the greater value it takes. Intensity of transvariation accounts not only for the frequency but also for the amount of income differences. The term Djh (eq. 5) called economic distance ratio or REA (relative economic affluence) is related to the normalized intensity of transvariation which is simply 1-Djh , and can be regarded as the measure of relative economic affluence of the j-th subpopulation with respect to the h-th subpopulation. It can be defined as the weighted sum of the income differences yji –yhr for all the members belonging to the population j-th with incomes greater than the income of all the members belonging to the population *h*-th, given that $\overline{Y}_j > \overline{Y}_h$ (for details see: Dagum, 1980).

As pointed out in Monti (2007), it is easy to verify that G_w, G_b and G_t of the Dagum decomposition (eq. 4 and eq. 5) equal, respectively, the within, the between and the interaction term of Mookherjee and Shorrocks decomposition and are also equivalent to Mehran's decomposition. It can be noted that the Dagum between-groups inequality (4) can be obtained without the rigorous assumption about equally distributed income groups. Moreover, it is worth mentioning that only the Dagum decomposition shows clearly how the overlapping term is connected both with between-groups and within-group inequality.

The inequality decomposition proposed by Yitzhaki and Lerman (1991) is based on the covariance formula, presented by the same authors (Lerman, Yitzhaki, 1985), where the Gini index is expressed in terms of twice the covariance between income and its rank divided by the overall mean income. Their decomposition encompasses an index of stratification that highlights the distinction between social stratification and inequality. It captures the extent to which population subgroups occupy distinct strata within an overall distribution. For the *i*-th subpopulation the index of stratification has the following form:

$$Q_i = \frac{\mathrm{cov}_i[F_i(y) - F_{n-i}(y), y]}{\mathrm{cov}_i[F_i(y), y]} \qquad (6)$$

where: $cov_i[F_i(y) - F_{n-i}(y), y]$ – covariance over group *i* between *y* and the difference between the ranking of a member of group *i* in his own group and the re-ranking he would have in the rest of the population,

$cov_i[F_i(y), y]$ - covariance over group *i* between *y* and its own ranking in group *i*.

The index of stratification given by (6) measures how members of a group differ from members of other groups. In this context stratification can be understood as "a group's isolation from members of other groups" (Yitzhaki, Lerman 1991). The index (6) has the following properties, making it sensitive to stratification of particular groups over an overall population:

- it measures the level of stratification for each group separately, taking into consideration the relation of its ranking in comparison with the rest of the population;
- Q_i declines when the number of the members of other groups being in the range of *i* increases;
- Q_i takes values from the interval <–1,1>. If Q_i = 1, a group *i* forms a perfect stratum - no members of other groups fall within its range of income. If Q_i = 0, a group *i* does not form a stratum at all - the ranking of all individuals within this group is identical to their ranking within the overall population (the groups completely overlap). Q = –1 in an extreme case when a group *i* is not well defined as being composed of two perfect strata placed at the tails of the distribution;

– given a number of the members of other groups who fall in the range of a group *i, Q_i* will be lower the closer the members of these groups are to the mean of *i*.

Income stratification is highly related to income inequality and can be a starting point to inequality decomposition by subpopulation groups. In general, high within-group inequality is likely to reduce a group stratification because it often increases overlapping of a group with other groups. On the other hand, high between-group inequality is likely to increase stratification by making the subpopulations more isolated from each other. Complicated connections between within-group inequality, between-group inequality and stratification can be revealed in detail by an unified framework given by a decomposition formula of Lerman and Yitzhaki (1991):

$$G = \sum_i s_i G_i + \sum_i s_i(p_i - 1)G_i Q_i + \sum_i \frac{2\operatorname{cov}[\bar{y}_i, \overline{F_i}(y)]}{\bar{y}} \tag{7}$$

where: $\overline{F_i}(y)$ – group *i*'s average rank.

The first component represents within-group inequality, the second component reflects the impact of stratification, described as intra-group inequality in overall ranks, while the third component accounts for the between-group inequality. Changes in income distribution may affect only one component of (7) or may have influence on all of them. High stratification implies low variability of ranks so the increases in group stratification exert negative impact on inequality. The between-group inequality is expressed as the between-group Gini index calculated on the basis of covariance between each mean income of a group and the average rank. As the authors point out, it is similar, but not identical to the between-group terms presented in Pyatt (1976), Mookherjee and Shorrocks (1982) and Silber (1989). The substantial difference is in the way the group ranks are established: in Lerman and Yitzhaki (1991) the ranking is obtained by averaging each ranking of observation within each subpopulation, while for the remaining authors it is simply the ranking of mean incomes. It is worth mentioning that when there is no overlapping between groups, all the methods yield the same results.

3. Application

The methods discussed above were applied to the analysis of income inequality in Poland by socio-economic groups, regions and family types. The basis for the calculations was micro data coming from the Household Budget Survey (HBS) conducted by Central Statistical Office in 2009. The data obtained from the HBS allow for the detailed analysis of the living conditions in Poland, being the basic source of information on the revenues and expenditure of the population. In 2009 the randomly selected sample covered 37,302 households, i.e. approximately 0.3% of the total number of households. The adopted sampling scheme was geographically stratified and two-stage one with different selection probability at the first stage. In the estimation of inequality measures and their decomposition the survey weights based on inverse inclusion probabilities were taken into consideration. In order to maintain the relation between the structure of the surveyed population and the socio-demographic structure of the total population, the data obtained from the HBS were weighted with the structure of households by the number of persons and class of locality coming from Population and Housing Census 2002.

The inequality analysis was conducted after separately dividing the overall sample: by region NUTS 1 constructed according to the Eurostat classification, by family type classified according to the number of children, and by socio-economic group established on the basis of the exclusive or primary source of maintenance. The variable of interest was household available income that can be considered the basic characteristic of its economic condition. It is defined as a sum of households' current incomes from various sources reduced by prepayments on personal income tax made on behalf of a tax payer by tax-remitter (this is the case of income derived from hired work and social security benefits and other social benefits); by tax on income from property; taxes paid by self-employed persons (including professionals and individual farmers), and by social security and health insurance premiums. To avoid interpretation problems, rare negative incomes were removed from the original sample.

Table 1 describes in detail the results of income inequality decomposition by socio- economic groups while tables 2 and 3 present the corresponding calculations outcome for the population divided by region and family type, respectively. To allow comparing the conditions of households of different sizes and different demographic structures, the square root scale, popular in recent OECD publications, was applied in the paper (table 3a). All the tables present statistical characteristics of household available income by population groups as well as the final results of inequality decomposition with respect to these groups. In particular, the within-groups, between-groups and "overlapping" components are reported for both Dagum (D) and Yitzhaki-Lerman (Y-L) approach (eq. (4), (5) ,(7)). As it has been mentioned above, these decompositions represent completely different concepts and thus provide us with inequality contributions that can be the basis of income inequality analysis from different perspectives. However, the main interest of this paper is groups overlapping and stratification. The overlapping component in the Dagum decomposition (called transvariation) is based on the relative economic affluence of one subpopulation with respect to another while the "third term" of Y-L method is based on ranking rather than income differences, and can only be regarded as a measure of groups separation. Similarly, the between-group component of the Dagum approach is based on income differences for each and every pair of households in contrast to the Y-L approach where only group means are considered. It results in higher sensitivity of the Dagum decomposition to changes in grouping factors, while the Y-L decomposition is by construction dominated by the within-group component (see. Tables 1-3).

Table 1. Decomposition of income inequality in Poland by socio-economic group

Measure	Socio-economic group					Total
	Employees	Farmers	Self-employed	Pensioners	Unearned sources	
Mean income $\bar{y}_i$ [1000 PLN]	3.781	4.556	4.738	2.108	1.695	3.186
Population proportion p_t	0.491	0.038	0.069	0.361	0.041	1
Income proportion s_i	0.583	0.054	0.103	0.239	0.021	1
Gini index G_i	0.293	0.483	0.319	0.306	0.370	0.352
Stratification index Q_i	0.313	–0.038	0.269	0.189	0.083	
Within-groups term (Y–L)	0.171	0.026	0.033	0.073	0.008	**0.311**
Between-groups term (Y–L)						**0.085**
Stratification term (Y–L)						**–0.044**
Within-groups term (D)	0.084	0.001	0.002	0.026	0.000	**0.114**
Between-groups term (D)						**0.154**
Transvariation (overlapping term) (D)						**0.085**

Source: Author's calculations.

Table 2. Decomposition of income inequality in Poland by region

Measure	Region of Poland						Total
	Central	Southern	Eastern	North-western	South-western	Northern	
Mean income $\bar{y}_i$ [1000 PLN]	3.554	3.093	2.861	3.227	3.159	3.122	3.186
Population proportion p_t	0.218	0.208	0.168	0.154	0.107	0.145	1
Income proportion s_i	0.243	0.202	0.151	0.156	0.106	0.142	1
Gini index G_i	0.381	0.318	0.355	0.342	0.352	0.348	0.352
Stratification index Q_i	–0.025	0.054	–0.023	0.031	–0.001	0.005	
Within-groups term (Y–L)	0.093	0.064	0.054	0.053	0.037	0.049	**0.351**
Between-groups term (Y–L)							**0.006**
Stratification term (Y–L)							**–0.003**
Within-groups term (D)	0.020	0.013	0.009	0.008	0.004	0.007	**0.062**
Between-groups term (D)							**0.042**
Transvariation (overlapping term) (D)							**0.249**

Source: Author's calculations

Table 3. Decomposition of income inequality in Poland by family type

Measure	Family type (number of children)						Total
	0	1	2	3	4	5…	
Mean income $\bar{y}_i$ [1000 PLN]	2.751	3.920	4.013	3.685	3.471	3.667	3.186
Population proportion p_t	0.643	0.183	0.126	0.035	0.009	0.004	1
Income proportion s_i	0.559	0.226	0.160	0.041	0.009	0.005	1
Gini index G_i	0.361	0.313	0.325	0.329	0.294	0.314	0.352
Stratification index Q_i	–0.028	0.169	0.165	0.108	0.104	0.107	
Within-groups term (Y–L)	0.201	0.071	0.052	0.013	0.003	0.002	**0.342**
Between-groups term (Y–L)							**0.027**
Stratification term (Y–L)							**–0.017**
Within groups term (D)	0.129	0.013	0.006	0.000	0.000	0.000	**0.150**
Between-groups term (D)							**0.071**
Transvariation (overlapping term) (D)							**0.131**

Source: Author's calculations.

Table 3a. Decomposition of income inequality in Poland by family type (equivalised income)

Measure	Family type (number of children)						Total
	0	1	2	3	4	5…	
Mean income $\bar{y}_i$ [1000 PLN]	1.947	2.066	1.910	1.563	1.330	1.260	1.942
Population proportion p_t	0.643	0.183	0.126	0.035	0.009	0.004	1
Income proportion s_i	0.645	0.194	0.124	0.028	0.006	0.003	1
Gini index G_i	0.308	0.308	0.322	0.319	0.282	0.293	0.312
Stratification index Q_i	0.034	0.018	-0. 033	-0.058	0.064	0.107	
Within-groups term (Y–L)	0.198	0.060	0.040	0.009	0.002	0.001	**0.310**
Between-groups term (Y–L)							**0.027**
Stratification term (Y–L)							**–0.002**
Within-groups term (D)	0.128	0.011	0.005	0.000	0.000	0.000	**0.144**
Between-groups term (D)							**0.021**
Transvariation (overlapping term) (D)							**0.147**

Source: Author's calculations.

The overall income inequality in Poland in 2009, measured by means of the Gini index and estimated on the basis of the Polish HBS, was 0.352 (for equivalent income G=0.312). These values confirm a high level of income inequality in Poland as compared with other European countries - according to EU-SILC in 2009 the Gini index calculated for equivalent disposable net income was at the level of 0.314 and in 2011 at the level 0.311, what was still above the EU average. It is worth mentioning that one can observe substantial differences in the values of inequality measures while using different data sources. The discrepancies between the values of the Gini index obtained on the basis of HBS, EU-SILC and Social Diagnosis for the same category of income may come from different sample sizes, different sampling designs and what seems the most important from the method of dealing with non-response. For example, the methodology of EU-SILC includes a requirement for the imputation of the missing income, what can lead to the underestimation of inequality measures and their standard errors. Moreover, one can run into difficulties while trying to compare the results over time - EU-SILC and Social Diagnosis are relatively new surveys and their implementation has been disturbed by many methodological changes. On the contrary, the Household Budget Survey is relatively stable and has the largest sample size, but even such a sample can be insufficient to provide reliable estimates in some divisions (see: Jędrzejczak, Kubacki, 2013).

The impact of the number of children on the distribution of household available income is presented in table 3. Applying the Dagum decomposition, the overall Gini index is due to within-group (43%) and overlapping (37%) components, while the contribution of the between-group term was found to be rather small (20%). The families without children form an untypical group (Q_0<0), which in fact consists of two smaller ones differing in average income level: a group of individuals (mainly retirees) and a group of couples without children. The significant stratification emerges only for the households with 1 or 2 children (Q_1=0,169; Q_2=0,165), identifying them as relatively similar within the groups and different from the outside. This result, however, can be misleading for two reasons. Firstly, the stratification indices Q_i proposed by Yitzhaki and Lerman ignore group sizes and can be negligible even for relatively separated groups when they are sufficiently small. Secondly, to compare subpopulations constructed on the basis of the number of children the equivalised income should be considered rather than the nominal one. After the transformation of available incomes with respect to household composition, the stratification indices, except for the first group, were found to be close to 0 (table 3a). Nevertheless, very high economic distance ratios D_{jh} were observed between small but the poorest groups of households (with 4 and 5 or more children) and the wealthiest group of families possessing only one child. They both exceed 60% so the families possessing only one child are 60% more affluent than the families with 4 and more children. The economic distance ratios D_{jh} consider pair comparisons between groups so they better detect income differences between various subpopulations than do Q indices.

The stratification and between-group inequality is much higher when the breakdown by socio-economic group is considered (table 1). The decomposition presented in table 1 takes into account the splitting up into households of self-employed, households of employees (managers, office workers, blue-collar workers, school teachers, etc.), households of not employed (retirees and pensioners) and households of other not employed (mainly unemployed). The households of farmers constitute a separate group.

Using the Dagum decomposition, the total income inequality in Poland by socio-economic group is dominated by between-group term that accounts for 44% of the overall Gini index. This result coincides with serious stratification indices, which were observed for several socio-economic groups and play an important role in Y-L decomposition. The within-group component (32%) reflects the inner polarization of the groups what gives rise to remarkable differentials in average income between managers and blue-collar workers within the group of *employees*, between entrepreneurs and the others within the group of *self-employed* or between retirees and pensioners within the fourth group. The households of *self-employed* are the wealthiest group, the one with the highest average income, but the group representing the highest level of inequality are *farmers* (*G*=0.48). The households of *employees* constitute a group with the highest share (24%) in the overall Gini index what is mainly due to its size and income share. The contribution of the overlapping component measured by transvariation is rather small (24%), contrary to high stratification indices for socio-economic groups except *farmers* and *unearned sources*. The negative value of the stratification index *Q* (and high *G*) observed for *farmers* suggests that this group is nonhomogeneous, being composed of the households that are not of the same kind (small and very large farms).

The impact of regional differences on income inequality in Poland can be observed in table 2. Contrary to family types and socio-economic groups, regional differences contribute slightly to the overall value of the Gini index. The between-group component accounts for only 12% of the overall income inequality. The Gini ratios and means within regions do not differ significantly so the contributions of particular subpopulations to the overall inequality are determined mainly by their sizes. The substantial contribution of transvariation, equal to 71% of the overall Gini index, is an evidence of notable overlapping of income distributions for NUTS 1 regions in Poland (see also: Jędrzejczak 2010).

4. Concluding remarks

Decomposition of the Gini index can be useful for social policy-makers in assessing the contributions of between-groups and within-groups inequalities to the overall inequality of a population. It can also be helpful in stratification and market segmentation by including the concept of overlapping.

The most widespread approach to the group decomposition of the Gini index was given by Dagum and it is based on the concepts of economic distance between distributions and relative economic affluence. It takes into account different variances and asymmetries of income distributions in subpopulations and gives an important contribution to the understanding of the overlapping term.

The Gini index decomposition proposed by Yitzhaki and Lerman encompasses the index of stratification by linking social stratification with inequality. It can be applied to assess isolation of social groups expressed in terms of income.

Estimation results obtained on the basis of Polish HBS revealed high discrepancies between socio-economic groups of households defined on the basis of primary source of maintenance, whereas regional differences were found to be relatively small and to contribute slightly to overall income inequality in Poland. Extremely large income differences were observed between some household groups differentiated by the number of children. One should also be conscious that the estimation results can be biased mainly because of a high non-response rate being an immanent feature of household budgets surveys all over the world.

Acknowledgements

The Author would like to thank the anonymous reviewer for her/his valuable comments and constructive suggestions which have improved the quality of the paper.

REFERENCES

BHATTACHARYA, N., MAHALANOBIS B., (1967). Regional disparities in household consumption in India, Journal of the American Statistical Association 62 (317), 143–161.

DAGUM, C., (1980). Inequality Measures Between Income Distributions with Application. Econometrica 48(7), 1791–1803.

DAGUM, C., (1997). A New Approach to the Decomposition of the Gini Income Inequality Ratio, Empirical Economics 22(4), 515–531.

DEUTSCH, I., SILBER J., (1999). Inequality Decomposition by Population Subgroups and the Analysis of Interdistributional Inequality, in: J. Silber, Handbook of Income Inequality Measurement, 363–397.

GASTWIRTH, J. L., (1972). The Estimation of the Lorenz Curve and the Gini Index, Review of Economics and Statistics 54(3), 306–316.

GINI, C., (1912). Variabilita e Mutabilita, Bologna, Tipografia di Pado Cuppini.

GINI, C., (1916). Il Concetto di Transvariazione e le sue Primi Applicazioni, Giornale degli Economisti e la Rivista Statistica, [in:] Gini (1959), 21–44.

GINI, C., (1959). Memorie di Metodologia Statistica, vol. II. Libreria Goliardica, Roma.

JĘDRZEJCZAK, A., (2010). Decomposition Analysis of Income Inequality in Poland by Subpopulations and Factor Components, Argumenta Oeconomica 1(24), 109–123.

JĘDRZEJCZAK, A., KUBACKI, J., (2013). Estimation of Income Inequality and the Poverty in Poland by Region and Family Type, Statistics in Transition 14(3), 359-378.

MEHRAN, F., (1975). A Statistical Analysis on Income Inequality Based on Decomposition of the Gini Index. Proceedings of the 40th Session of ISI.

MOOKHERJEE, D., SHORROCKS, A., (1982). A Decomposition Analysis of the Trend in UK Income Inequality, The Economic Journal 92(368), 886–902.

MONTI, M., (2007). Note on the Dagum Decomposition of the Gini Inequality Index, Università Degli Studi di Milano Working Papers 2007–16.

MORGAN, J., (1962). The Anatomy of Income Distribution, Review of Economics and Statistics 44(3), 270–282.

PYATT, G., (1976). On the Interpretation and Disaggregation of Gini Coefficient, The Economic Journal 86(342), 243–255.

RADAELLI, P., (2010). On the Decomposition by Subgroups of the Gini and Zenga‘s Uniformity and Inequality Indexes, International Statistical Review, 78(1), 81–101.

SHORROCKS, A., (1984). Inequality Decomposition by Population Subgroups, Econometrica 52(6), 1369–1385.

SOLTOW, L., (1960). The Distribution of Income Related to Changes in the Distributions of Education, Age, and Occupation, The Review of Economics and Statistics 42(4), 450–453.

SILBER, J., (1989). Factor Components, Population Subgroups and the Computation of the Gini Index of Inequality, The Review of Economics and Statistics 71(2), 107–115.

VERNIZZI, A., (2009). Applying the Hadamard Product to Decompose Gini, Concentration , Redistribution and Re-ranking Indices, Statistics in Transition 10 (3), 505–524.

YITZHAKI, S., (1994). Economic Distance and Overlapping of Distributions, Journal of Econometrics 61(1), 147–159.

YITZHAKI, S., LERMAN, R., (1984). A Note on the Calculation and Interpretation of the Gini Index, Economic Letters 15(3-4), 363–369.

YITZHAKI, S., LERMAN, R., (1991). Income Stratification and Income Inequality, Review of Income and Wealth 37(3), 313–329.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 283–298

APPLICATION OF COHERENT DISTORTION RISK MEASURES

Grażyna Trzpiot[1]

ABSTRACT

This paper concentrates on solving the portfolio selection problem. It starts with an extension of the well-known optimization framework for Conditional Value-at-Risk (CVaR)-based portfolio selection problems [1, 2] to optimization over a more general class of risk measure - known as the class of Coherent Distortion Risk Measure (CDRM). The CDRM class of risk measures is the intersection of Coherent Risk Measure (CRM) and Distortion Risk Measure (DRM). It concludes with showing that many of the well-known risk measures are of special cases of the CDRM class what may facilitate to deal with the portfolio optimization problem

Key words: coherent risk measure, distortion risk measure, coherent distortion risk measure.

1. Introduction

The problem of optimal portfolio selection is very important issue to investors, hedgers, fund managers and individual investors. The research on optimal portfolio selection has been growing rapidly. Researchers and practitioners are constantly looking for better and more sophisticated risk measures and reward trade-off in constructing optimal portfolios. The classical Markowitz[2] model used variance as the benchmark for risk measurement and this is perceived to be undesirable since it penalizes equally both sides, regardless of downside risk or upside potential. Consequently, other measures of risk have been proposed in connection with portfolio optimization. These include semi-variance[3],

[1] University of Economics in Katowice. E-mail: trzpiot@ue.katowice.pl.

[2] H. M. Markowitz. Portfolio selection. The Journal of Finance, 7(1): pp. 77–91, 1952.

[3] H. M. Markowitz. Portfolio selection: efficient diversification of investments. New Haven, CT: Cowles Foundation, 94, 1959.

partial moments[1], safety first principle[2], skewness and kurtosis[3], value-at-risk (VaR) and conditional value-at-risk (CVaR)[4].
Let us take some basic notation: let $\rho(\mathbf{x})$ be a function which measures the riskiness of a portfolio **x**. Then, in general, the portfolio selection problem seeks the solution to

$$\min_{\mathbf{x}\in\mathbf{S}} \rho(\mathbf{x}) \tag{1}$$

where the minimization is taken over all feasible portfolios **S**. If ρ corresponds to the variance of the return of the portfolio **x**, then this problem (1) reduces to the standard Markowitz model. We are concerned with the portfolio selection problem involving a particular class of risk measure known as the coherent distortion risk measure (CDRM). CDRM is the intersection of two important families of risk measures: the coherent risk measure (CRM)[5] and the distortion risk measure (DRM)[6]. We study the intersection of both classes: CVaR is an example of CDRM while VaR is neither CRM nor DRM, and hence not CDRM[7].

2. CVaR-based portfolio optimization model

Let $\boldsymbol{l} = f(\mathbf{x},\mathbf{y})$ be the portfolio loss associated with the decision vector **x** , to be chosen from a set $\mathbf{S} \subseteq n$, and the random vector $\mathbf{y} \in m$. The vector **x** represents what we may generally call a portfolio, with **S** capturing the set of all feasible portfolios subject to certain portfolio constraints. For every **x** , the loss $f(\mathbf{x},\mathbf{y})$ is a random variable having a distribution induced by the distribution of $\mathbf{y} \in m$.

[1] V. S. Bawa and E.B. Lindenberg. Capital market equilibrium in a mean-lower partial moment framework. Journal of Financial Economics, 5(2):189–200, 1977, Trzpiot G. (2005). Partial Moments and Negative Moments in Ordering Asymmetric Distribution, in: Daniel Baier and Klaus-Dieter Wernecke (eds.): Innovations in Classification, Data Science and Information Systems, Proc. 27th Annual GFKL Conference, University of Cottbus, March 11-14 2003. Springer-Verlag, Heidelberg-Berlin, 181-188.

[2] A. D. Roy. Safety first and the holding of assets. Econometrica: Journal of the Econometric Society, pages 431–449, 1952.

[3] C. R. Harvey, J. Liechty, M. Liechty, and P. Müller. Portfolio selection with higher moments. Quantitative Finance, 10(5):469–485, 2010.

[4] R.T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. Journal of risk, 2, 21–42, 2000.

[5] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath. Coherent measures of risk. Mathematical finance, 9(3):203–228, 1999.

[6] S. S. Wang. A class of distortion operators for pricing financial and insurance risks. The Journal of Risk and Insurance, 67(1):15–36, 2000.

[7] Trzpiot G. Własności transformujących miar ryzyka [Properties of transforming risk measures], Economic Studies of the University of Economics in Katowice - Faculty Scientific Papers No. 91, 21–36, 2012.

The underlying probability distribution of **y** is assumed to be discrete with probability masses ***p*** , i.e., P[***l*** = *L*(**x**, y_i)] = p_i for *i* = 1, … ,*m*.

We can notice that in many cases it is assumed that *X* - the portfolio loss has a discrete uniform distribution. This is not a very limiting assumption if we restrict ourselves to discrete portfolio loss distributions, which is typically the case if we are obtaining distributional information via scenario generation or from historical data. In addition, given any arbitrary discrete distribution representable with rational numbers, we may always convert it to discrete uniform distribution for some large enough *m*.

For every portfolio **x** denote the cumulative distribution function (cdf) of the portfolio loss ***l*** = *f* (**x**,**y**) as:

$$\Psi(x,\zeta)=\sum_{i=1}^{m} p_i I\{l_i \le \zeta\}$$

Then α -VaR and α -CVaR are defined as follows[1].

Definition 2.1. *Suppose for each* ***x***∈***S****, the distribution of the portfolio loss* ***l*** *= f(**x**,**y**) is concentrated in m ‹ ∞points, and Ψ(**x**, ·) is a step function with jumps at these points. Now, fixing* **x** *and* $l_{(1)}$ *<....<* $l_{(m)}$ *denoting the corresponding ordered portfolio loss points and* $p_{(i)}$ *> 0, i =1, …, m, represent the probability of realizing loss* $l_{(i)}$*.*

If

minρ ( )
x∈**S**

x denotes the unique index satisfying α then α *-VaR* and α *-CVaR* of the portfolio loss are given, respectively, by ζ_α (**x**) = $l_{(i\alpha)}$ and

$$\phi_\alpha(x)=\frac{1}{1-\alpha}\left[\left(\sum_{i=1}^{i\alpha} p_{(i)}-\alpha\right)l_{i\alpha}+\sum_{i=i\alpha}^{m} p_{(i)}l_{(i)}\right] \tag{2}$$

As pointed out, if ρ is set to VaR, then the resulting portfolio problem (1) is numerically challenging due to its lack of convexity. In contrast, the CVaR-based portfolio optimization problem (1) is a convex program and hence it is computationally amenable. The CVaR-based portfolio model becomes even more popular and more practical when [1] shows that the convex program can in fact be formulated as a l iner program. The key to Rockafellar-Uryasev's linear optimization scheme of CVaR-based portfolio selection problem is expressing ϕ_α(**x**) and ζ_α(**x**) in terms of the following special function:

$$F_\alpha(x,\zeta)=\zeta+\frac{1}{1-\alpha}E\left[(l(x,y)-\zeta)^+\right]=\zeta+\frac{1}{1-\alpha}\sum_{i=1}^{m} p_i(l_i-\zeta)^+ \tag{3}$$

[1] R.T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.

If f (**x**,**y**) is convex with respect to **x** , then $\varphi_\alpha(\mathbf{x})$ is convex with respect to **x** . In this case, $F_\alpha(\mathbf{x}, \zeta)$ is also jointly convex in $(\mathbf{x}, \zeta)$. Armed with these findings, Rockafellar and Uryasev derived the following equivalence formulation[1]:

Theorem 2.1. *Minimizing $\phi_\alpha(\mathbf{x})$ with respect to $\mathbf{x} \in \mathbf{S}$ is equivalent to minimizing $F_\alpha(\mathbf{x}, \zeta)$ over all $(\mathbf{x}, \zeta) \in \mathbf{S}\times$, in the sense that*

$$\min_{x\in S}\phi_\alpha(x) = \min_{(x,\zeta)\in S\times R} F_\alpha(x,\zeta) \tag{4}$$

where moreover

$$\left(x^*,\zeta^*\right) \in \min_{(x,\zeta)\in S\times R} F_\alpha(x,\zeta) \Leftrightarrow x^* \in \arg\min_{x\in S}\phi_\alpha(x),\ \zeta^* \in \min_{(x,\zeta)\in S\times R} F_\alpha\left(x^*,\zeta\right) \tag{5}$$

The above theorem links the representation (3) explicitly to both VaR and CVaR simultaneously. The theorem asserts that for the purpose of determining an optimal portfolio with respect to CVaR, we can replace $\phi_\alpha(\mathbf{x})$ by $F_\alpha(\mathbf{x}, \zeta)$ in portfolio selection problems. More importantly, by exploiting (3) the general convex programming of CVaR portfolio optimization problem can be linearized into a linear objective function with additional linear auxiliary constraints. With such linear representation we can cast any portfolio selection problem with CVaR objective and linear constraint(s) as a linear program.

3. Coherent Risk Measure (CRM) and Distortion Risk Measure (DRM)

The uncertainty for future value of an investment position is usually described by a function $X : \Omega \to$ R, where Ω is a fixed set of scenarios with a probability space (Ω, F, P). Let $\mathscr{X}$ be a linear space of random variables on Ω, i.e., a set of functions $X : \Omega \to$ R. Note that X can be thought of as a loss from an uncertain position.

We can find out some properties of risk measures.

Property 1. Law-invariance

Law-invariance states that a risk measure ρ(X) does not depend on a risk itself but only on i ts underlying distribution, i.e. $\rho(X) = \rho(F_X)$, where F_X is the distribution function of X. This condition ensures that F_X contains all the information needed to measure the riskiness of X. Law-invariance can be phrased as:

$$F_X = F_Y \Rightarrow \rho(X) = \rho(Y)$$

[1] Ibid.

for every random portfolio returns X and Y with distribution functions FX and FY. In other words, ρ is law-invariant in the sense that ρ(X) = ρ(Y), whenever X and Y have the same distribution with respect to the initial probability measure P. This assumption is essential for a risk measure to be estimated from empirical data, which ensures its applicability in practice.

Property 2. Positive homogeneity

Positive homogeneity (also known as positive scalability) formulates as follows: for each positive λ and random portfolio return $X \in \mathcal{X}$:

$$\rho(\lambda X) = \lambda^k \rho\,(X).$$

Positive homogeneity signifies that a measure has the same dimension (scalability) as a variable X. When the parameter k = 0, a risk measure does not depend on the scalability.

From a financial perspective, positive homogeneity implies that a linear increase of the return by a positive factor leads to a linear increase in risk by the same factor.

Property 3. Sums of risks

Consider two different financial instruments with random payoffs $X, Y \in \mathcal{X}$: The payoff of a portfolio consisting of these two instruments will equal X + Y .

Property 3.1. Sub-additivity[1]

Sub-additivity states that the risk of the portfolio is not greater than the sum of the risks of the portfolio components. In other words, "a merger does not create extra risk".

$$\rho(X + Y\,) \leq \rho(X) + \rho(Y\,)$$

Compliance with this property tends to the diversification effect. Although Artzner treat sub-additivity as a necessary requirement for constructing a risk measure in order for it to be coherent, empirical evidence suggests that sub-additivity does not always hold in reality[2].

Property 3.2. Additivity

The additivity property is expressed in the following form:

$$\rho(X + Y) = \rho(X) + \rho(Y\,)$$

This property is valid for independent and comonotonic[3] random variables X and Y. The comonotonic random variables with no-hedge condition result in comonotonic additivity.

[1] Artzner et al. (1999).

[2] Critiques of sub-additivity can be found in Dhaene et al. (2003) and Heyde et al. (2006).

[3] Comonotonic or common monotonic random variables (Yaari (1987), Schmeidler (1986), Dhaene et al. (2002a, 2002b)) are such that if the increase of one follows the increase of the other variable: $P[X \leq x, Y \leq y] = \min\{P[X \leq x], P[Y \leq y]\}$ for all $x, y \in R$.
Intuitively, such variables have a maximal level of dependency. Comonotonic random variables are necessarily positively correlated. In financial and insurance markets, this property appears quite frequently.

Property 3.3. Super-additivity

Super-additivity states that the portfolio risk estimate could be greater than the sum of the individual risk estimates.

$$\rho(X + Y) \geq \rho(X) + \rho(Y)$$

The super-additivity property is valid for risks which are positive (negative) dependent.

Property 4. Convexity

(1) For all X, Y ∈ *X*, $0 \leq \lambda \leq 1$, the following inequality is true:

$$\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y).$$

Convexity ensures the diversification property and relaxes the requirement that a risk measure must be more sensitive to aggregation of large risks.

(2) For any λ, $\mu \geq 0$, $\lambda+\mu = 1$, and distribution functions *F,G*, the following inequality holds

$$\rho(\lambda F + \mu G) \leq \lambda\rho(F) + \mu\rho(G).$$

(3) Generalized convexity. For any λ, $\mu \geq 0$, $\lambda+\mu = 1$ and distribution functions U, V, H, such that the following random variables exist X, Y , $\lambda X + \mu Y$, for which $F_X = U$, $F_Y = V$, $F_{\lambda X+\mu Y} = H$, the inequality is true

$$\rho(H) \leq \lambda\rho(U) + \mu\rho(V).$$

Property 5. Monotonicity

For every random portfolio returns X and Y such that $X \geq Y$,

$$\rho(X) \leq \rho(Y).$$

Monotonicity implies that if one financial instrument with the payoff X is not less than the payoff Y of the other instrument, then the risk of the first instrument is not greater than the risk of the second financial instrument. Another presentation of the monotonicity property with a risk-free instrument is as follows:

$$X \geq 0 \Rightarrow\Rightarrow \rho(X) \leq \rho(0)$$

for X ∈ *X*,.

Property 6. Translation invariance

Property 6.1. For the non-negative number $\alpha \geq 0$ and C ∈ R, the property has the following form:

$$\rho(X + C) = \rho(X) - \alpha C.$$

This property states that if the payoff increases by a known constant, the risk correspondingly decreases. In practice, $\alpha = 0$ or $\alpha = 1$ are often used.

Property 6.2. When $\alpha = 0$, it implies that the addition of a certain wealth does not increase risk. This property is also known as the Gaivoronsky-Pflug (G-P) translation invariance[1] .

Property 6.3. The case when $\alpha = 1$ implies that by adding a certain payoff, the risk decreases by the same amount.

$$\rho(X + C) = \rho(X) - C.$$

Property 6.4. When a constant wealth has a positive value, i.e., $C \geq 0$, one gets

$$\rho(X + C) \leq \rho(X).$$

This result is in agreement with the monotonicity property of $X + C \geq X$.

Property 6.5. In particular, translation invariance involves

$$\rho(X + \rho(X)) = \rho(X) - \rho(X) = 0,$$

obtaining a risk-neutral position by adding $\rho(X)$ to the initial position X.

Property 7. Consistency

Property 7.1. Consistency with respect to n-order stochastic dominance has the following general form:

$$X \geq_n Y,\ \rho(X) \geq \rho\ (Y\ ).$$

In practice, the maximal value of $n = 2$; $n = 0$ just stands for a monotonicity property.

Property 7.2. Monotonic dominance of n-order

$$X \geq_{M(n)} Y, \text{ if } E[u(X)] \geq E[u(Y\ )]$$

for any monotonic of order n functions, that is $u^{(n)}(t) \geq 0$.

It is known, that $X \geq_1 Y$ is equivalent to $X \leq_{M(1)} Y$. $X \leq_{M(2)} Y$ is also called the Bishop-de Leeuw ordering or Lorenz dominance.

Property 7.3. First-order stochastic dominance (FSD)

$$\text{For } X \geq_1 Y,\ F_X(x) \leq F_Y\ (x)$$

If an investor prefers X to Y, then FSD will indicate that the risk of X is less than the risk of Y. In terms of utility function u, the following holds

$$\text{If } X \geq_1 Y, \text{ then } E[u(X)] \geq E[u(Y\ )]$$

[1] Gaivoronski A. and Pflug G., 2001, Value at risk in portfolio optimization: properties and computational approach, Technical Report, University of Vienna.

for all increasing utility functions u. FSD characterizes the preferences of risk-loving investors. Ortobelli et al. (2006) classified risk measures consistent with respect to FSD as *safety-risk measures*[1].

Property 7.4. Rothschild-Stiglitz stochastic order dominance (RSD)

RSD has the form[2]:

$$\text{If } X \leq_{RS} Y, \text{ then } E[u(X)] \geq E[u(Y\ )]$$

for any concave, not necessarily decreasing, utility function u. RSD describes preferences of risk-averse investors. *Dispersion measures* are normally consistent with RSD.

Property 7.5. Second-order stochastic dominance (SSD)

SSD has the following form:

$$\text{For } X \geq_2 Y,\ E[u(X)] \geq E[u(Y\ )]$$

for all increasing, concave utility functions u. SSD characterizes non-satiable risk averse investors[3].

Property 7.6. Stochastic order - stop-loss Y dominates X ($Y \geq_{SL} X$) in stop-loss order, if for any number α the following inequality is true:

$$E[(Y-\alpha)^+] \geq E[(X-\alpha)^+].$$

Here, $\alpha^+ = \max\{0, \alpha\}$. Such order is essential in the insurance industry. If the insurer takes the responsibility for the claims greater than α (deductible), then the expected claim Y is not smaller than X.

Property 7.7. Convex order Y dominates X with respect to convex order ($Y \geq_{CX} X$) if the relation $Y \geq_{SL} X$ is true and when $\alpha = -\infty$ in stop-loss order, i.e. $E[X] = E[Y]$. Convex ordering is related to the notion of risk aversion[4].

Consistency with the stochastic dominance is a necessary property for a risk measure, because it enables one to characterize the set of all optimal portfolio choices when either wealth distributions or expected utility functions depend on a finite number of parameters [5].

Property 8. Non-negativity

Property 8.1. $\rho(X) \geq 0$, while $\rho(X) > 0$ for all non-constant risk.

Property 8.2. If $X \geq 0$, then $\rho(X) \leq 0$; if $X \leq 0$, then $\rho(X) \geq 0$.

[1] In the portfolio selection literature, two disjoint categories of risk measures are defined: dispersion measures and safety-first risk measures. (Ortobelli S., Rachev S., Shalit H. and Fabozzi F., 2006).

[2] Rothschild M. and Stiglitz J., 1970.

[3] Hadar J. and Russell W., 1969.

[4] See also Kaas et al. (1994, 2001).

[5] Ortobelli S., 2001.

Property 9. Continuity

Property 9.1. P robability convergence continuity: If $X_n \xrightarrow{P} X$, then $\rho(X_n)$ converges and has the limit $\rho(X)$.

Property 9.2. Weak topology continuity: If $F_{Xn} \xrightarrow{w} F_X$, then $\rho(F_{Xn})$ converges and has the limit $\rho(F_X)$.

Property 9.3. Horizontal shift continuity: $\lim_{\delta\to 0} \rho(X + \delta) = \rho(X)$.

Property 9.4. Opportunity of arbitrary risk approximation with the finite carrier is expressed by the equality[1]:

$$\lim_{\sigma\to+\infty} \rho(\min\{X,\delta\}) = \lim_{\sigma\to-\infty} \rho(\min\{X,\delta\}) = \rho(X)$$

Property 9.5. Lower semi-continuity: For any $C \in R$, the set $\{X \in \mathit{X}: \rho(X) \leq C\}$ is $\sigma(L^{\infty},L^1)$ - closed.

Property 9.6. Fatough property[2]

For any bounded sequence (X_n) for which $X_n \xrightarrow{P} X$, the following holds:

$$\rho(X) \leq \lim_{n\to\infty} \inf \rho(X_n).$$

These properties are cardinally important. Nonfulfilment of the continuity property implies that even a small inaccuracy in a forecast can lead to the poor performance of a risk measure.

Property 10. Strictly expectation-boundedness

The risk of a portfolio is always greater than the negative of the expected portfolio return.

$$\rho(X) \geq -E[X], \text{ while } \rho\ (X) > -E[X] \text{ for all non-constant } X,$$

where E[X] is the mathematical expectation of X.

Property 11. Lower-range dominated

Deviation measures possess lower-range dominated property of the following form:

$$D(X) \leq E(X)$$

for a non-negative random variable. From property 10 and property 11 one can derive:

$D(X) = \rho(X - EX)$, $\rho(X) = D(X) - E(X)$

Property 12. Risk with risk-free return C

[1] Wang et al. (1997).

[2] In some contexts, it is equivalent to the upper semi-continuity condition with respect to $\sigma(L^{\infty},L^1)$.

Property 12.1. ρ (C) = −C, it follows from the invariance property 6.3. If C > 0, then the situation is stable, risk is negative. The opposite situation occurs with C < 0.

Property 12.2. ρ (C) = 0, risk does not deviate with the zero certain return.

Property 13. Symmetric property

(1) ρ(−X) = −ρ(X), which corresponds to property 8.1.

(2) ρ(−X) = ρ(X), this property makes sense for the measures with possible negative values (property 8.2 fulfilled).

Property 14. Allocation

A risk measure need not be defined on the whole set of values of a random variable. Formally, in a given set U, from the condition $F_X = F_Y$, when $x \notin U$, it follows that ρ(X) = ρ(Y). Apparently, this property holds only for law-invariant measures. Most often, some threshold value T is assigned, and the set U takes values U = (−∞, T] or U = [T,∞).

Property 15. Static and dynamic natures

It is useful to use a dynamic and multi-period framework to answer the following question: how should an institution proceed with new information in each period and how should it reconsider the risk of the new position? Riedel (2004) introduced the specific axioms such as p redictable translation invariance and dynamic consistency for a risk measure to capture the dynamic nature of financial markets.

CRM satisfies properties of *monotonicity, translation invariance, positive homogeneity,* and *subaddivity*[1].

DRM satisfies properties of conditional *state independence, monotonicity, comonotonic additivity* and *continuity*[2].

The notion of comonotonicity is central in risk measures[3]. Imposed axiom comonotonic additivity based on the argument that the comonotonic random variables do not hedge against each other, leads to additivity of risks.

It was also proved[4] that for all the *Bernoulli*(*n*, *p*) random variables, if $0 \le p \le 1$, then a D RM ρ satisfies ρ (1) = 1 if and only if ρ has a Choquet integral representation with respect to a distorted probability; i.e.

[1] P. Artzner, F. Delbaen, J.M. Eber, and D. Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

[2] Trzpiot G., O własnościach transformujących miar ryzyka, Studia Ekonomiczne UE w Katowicach nr 91, 21–36, 2012.

[3] J. Dhaene, S.S. Wang, V.R. Young, and M.J. Goovaerts.

[4] Wang, 2000.

$$\rho_g(X)=\int Xd(g\circ P)=\int_{-\infty}^{0}[g(P(X>x))-1)]dx+\int_{0}^{\infty}[g(P(X>x))dx \qquad (6)$$

where $g(\cdot)$ is known as the distortion function which is nondecreasing with $g(0)=0$,
and
$g(1)=1$ and $(g\circ P)(A):=g(P(A))$ is called the distorted probability. The Choquet integral representation of DRM can be used to explore its mathematical properties. Furthermore, calculations of DRMs can easily be done by taking the expected value of X under probability measure $\mathrm{P}^{*}:=g\circ \mathrm{P}$.

For discretely distributed portfolio losses random variable $\boldsymbol{l}$ □(l_1, …,l_m) with probability masses P[$\boldsymbol{l}$ = l_i] = p_i dla i = 1,…,m, with cdf done as $F_l(l)=\sum_{i=1}^{m} p_i\mathbf{1}_{\{l_i\le l\}}$
and the survival function $S_l(l)=1-F_l(l)$, (6) becomes

$$E_{P^*}(X)=\int_0^\infty S^*(x)dx-\int_0^\infty F^*(-x)dx=\int_0^\infty g[S(x)]dx-\int_0^\infty \tilde{g}[F(-x)]dx$$

$$=\int_0^\infty g[S(x)]dx-\int_0^\infty \{1-g[S(-x)]\}dx=\rho_g(X).$$

Here we list some commonly used distortion functions:

− CVaR distortion:

$$g_{CVaR}(x,\ \alpha)=\min\{x/(1-\alpha),\ 1\}\ \text{dla}\ \alpha\in[0,1) \qquad (7)$$

− Wang Transform (WT) distortion:

$$g_{WT}(x,\ \beta)=\Phi[\Phi^{-1}(x)-\Phi^{-1}(\beta)]\ \text{dla}\ \beta\in[0,1) \qquad (8)$$

- the dual-power g_{DP} distortion[1]:

$$g_{DP}(x,\nu)=1-(1-x)^{\frac{1}{\nu}}\ ,\ x\in[0,\ 1],\ \nu\le 1 \qquad (9)$$

– Proportional hazard (PH) distortion:

$$g_{PH}(x,\gamma)=x^{\gamma}\ ,\ x\in[0,\ 1],\ \gamma\le 1\ . \qquad (10)$$

[1] Wirch J, Hardy MR (2001), Trzpiot (2004a, 2006).

4. Coherent Distortion Risk Measure (CDRM)-based portfolio selection

Recall that CDRM is the intersection of CRM and DRM. There are two ways to define CDRM.

Definition 4.1. *We say ρ is a coherent distortion risk measure (CDRM) if:*

a) $\rho\, g$ is a distortion risk measure (DRM) with a concave distortion function g

or equivalently

b) ρ is a coherent risk measure (CRM) that is also comonotonic and law-invariant.

The following representation theorem for CDRM is the key result that enables us to use a convex optimization framework for any CDRM portfolio selection problem.

Theorem 4.1.[1] *For any random variable X and a given concave distortion function g, risk measure ρ_g is a CDRM if and only if there exists a function w:[0,1] → [0,1], satisfying* $\int_{\alpha=0}^{1} w(\alpha)d\alpha = 1$ *such that*

$$\rho_\alpha(X) = \int_{\alpha=0}^{1} w(\alpha)\phi_\alpha(X)d\alpha \tag{11}$$

where (X) $\alpha\,\phi$ is the α-*CVaR* of X.

This representation theorem says that any CDRM can be represented as a convex combination of $CVaR_\alpha\ (X)$, $\alpha \in [0, 1]$ and we can construct any CDRM based on a convex combination of $CVaR_\alpha\ (X)$. Such result was proved for continuous portfolio loss distributions[2]. Proved and strengthened was the representation theorem that any CDRM can be represented as a convex combination of finite number of $CVaR\alpha\ (X)$ under the assumption that the portfolio loss has a discrete uniform distribution[3].

For solving convex programming formulation CDRM portfolio selection problem, we need a theorem for CDRM to general discrete loss distributions. We notice the following definition[4]:

[1] S. Kusuoka, On law invariant coherent risk measures. Advances in mathematical economics, 3:83–95, 2001.

[2] Ibid.

[3] D. Bertsimas and D.B. Brown. Constructing uncertainty sets for robust linear optimization. Operations research, 57(6):1483–1495, 2009.

[4] Feng M. B., Tan K. S. (2012).

Definition 4.2. *For a given loss observation* $\boldsymbol{l} = (l_1, \ldots, l_m)$ *and its ordered losses* $l_{(1)} < l_{(2)} < \ldots < l_{(m)}$, $p_{(i)}$ *be the probability of realizing* $l(i)$, $i = 1, m$ *and* $S_l\left(l_{(i)}\right) = 1 - \sum_{j=1}^{i} p_{(i)}$. *Define a* $CVaR_a$ *matrix* $\mathbf{Q} \in \mathrm{R}^m \times \mathrm{R}^m$ *with columns m* $\mathbf{Q}_i \in \mathrm{R}^m$, $i = 1, m$ *as*

$$\mathbf{Q} = [\mathbf{Q}_{1,} \mathbf{Q}_2, \ldots, \mathbf{Q}_m] = \begin{bmatrix} p(1) & 0 & 0 & \ldots & 0 \\ p(2) & \frac{p(2)}{1-S_l\left(l_{(1)}\right)} & 0 & \ldots & 0 \\ p(3) & \frac{p(3)}{1-S_l\left(l_{(1)}\right)} & \frac{p(3)}{1-S_l\left(l_{(2)}\right)} & \ldots & 0 \\ .. & \ldots & \ldots & \ldots & \ldots \\ p(m) & \frac{p(m)}{1-S_l\left(l_{(1)}\right)} & \frac{p(m)}{1-S_l\left(l_{(2)}\right)} & \ldots & \frac{p(m)}{1-S_l\left(l_{(m-1)}\right)} \end{bmatrix}$$

Since portfolio losses are discretely distributed at m points, there are m jumps in the cumulative function of $\boldsymbol{l}$.

By defining

$$\alpha_i = \begin{cases} 0, & dla \quad i = 1 \\ \sum_{j=1}^{i-1} p_{(j)} & dla \quad i = 2, \ldots m \end{cases} \tag{12}$$

at these m jumps, the m CVaRs at these probability levels are given by

$$\phi_{\alpha_i}(l) = \frac{1}{1-\alpha_i} \sum_{j=i}^{m} p(j) l(j) = \sum_{j=i}^{m} \frac{p(j)}{1-S_l\left(l_{(m-1)}\right)} l(j) = \sum_{j=i}^{m} Q_{ij} l(j) \tag{13}$$

for $i = 1, m$ and Q_{ij} is the (i, j) -th entry of $\mathbf{Q}$. Note that column $\mathbf{Q}_i$ is essential to the calculation of $CVaR_{(i-1)/m}(\boldsymbol{l})$ and hence it explains the name of the matrix.

We consider the following special function for some $w(\alpha) \geq 0$ and $\int_{\alpha=0}^{1} w(\alpha) d\alpha = 1$,

$$M_g(x, \zeta) = \int_{\alpha=0}^{1} w(\alpha) F(x, \zeta_\alpha) d\alpha. \tag{14}$$

Theorem 4.1. of CDRM ensures the existence of $w(\alpha)$, $\alpha \in [0,1]$ and defines CDRM for a given set of weights. For each α there is a corresponding auxiliary variable ζ_α . Taking partial derivatives w.r.t. all ζ_α and setting them equal to zeros give the extremal properties of *Mg* $(\mathbf{x},\zeta)$. This provides more insights about the connection between a particular CDRM, $\rho_g(\mathbf{x})$, and its convex representation *Mg* $(\mathbf{x},\zeta)$[1] . Yet ζ may have infinitely many entries ζ_α.

Taking partial derivative w.r.t. all $\zeta\alpha$ for $\alpha \in [0,1]$ requires calculus of variations, which is outside the scope of this thesis. We alleviate such difficulty by applying properties of Choquet integrals because CDRM is a subclass of DRM.

We conclude by presenting this generalized CVaR-based portfolio model of [1] to the more general class of CDRM-based portfolio model:

Theorem 4.2 *Let* $\rho_g(\mathbf{x})$ *be a CDRM with a corresponding distortion function g. Minimizing* $\rho_g(\mathbf{x})$ *with respect to* $\mathbf{x}\in\mathbf{S}$ *is equivalent to minimizing Mg* $(\mathbf{x},\zeta)$ *over all* $(\mathbf{x}, \zeta) \in \mathbf{S}\times |\zeta|$, *in the sense that*

$$\min_{x\in S} \rho_g(x) = \min_{(x,\zeta)\in S\times R^{\varsigma}} M_g(x,\zeta) \tag{15}$$

where moreover

$$\left(x^*,\zeta^*\right) \in \underset{(x,\zeta)\in S\times R^{\varsigma}}{\arg\min}\ M_g(x,\zeta) \Leftrightarrow x^* \in \underset{x\in S}{\arg\min}\, \rho_g(x),\ \zeta^* \in \underset{\zeta\in R^{\varsigma}}{\arg\min}\, M_g\left(x^*,\zeta\right) \tag{16}$$

As some remarks we can notice that:

- It is of interest to note that the PH portfolio optimization is almost equivalent to optimization over two extreme CVaR-based portfolios: one with $\alpha = 0.99$ and the other with $\alpha = 0$. Recall that minimizing CVaR with high value of α implies that you are someone who is very risk averse and hence is interested in risk minimization. In contrast, minimizing CVaR with α close to 0 implies an investor is a risk seeker and is only interested in maximizing expected return.
- Consistent with the classical trade-off theory on risk and reward, a more risk averse investor seeks an optimal portfolio with lower risk (as measured by the respective CDRM) but at the expense of lower expected return. Hence, the expected return of the optimal portfolio decreases with α for CVaR, decreases with β for WT, increases with γ for PH, and increases with δ for LB.[2]

[1] Ibid.

[2] Trzpiot G. (2010). Pessimistic portfolio optimization, In: Modelling of preferences and risk '09, Scientific Papers, University of Economics in Katowice, 121–128.

5. Conclusion

In this paper we present extension of the well-known linear optimization framework for CVaR to a general class of risk measure known as the CDRM. We generalized the finite generation theorem for CDRM and showed that any CDRM can be defined as a convex combination of ordered portfolio losses and equivalently a co nvex combination of CVaRs. We make use of the latter to develop a CDRM-based portfolio optimization framework.

REFERENCES

ARTZNER, P., DELBAEN F., EBER J. M., HEATH, D., (1999). Coherent measures of risk. Mathematical finance, 9(3):203–228.

BAWA, V. S., LINDENBERG E. B., (1977). Capital market equilibrium in a mean-lower partial moment framework. Journal of Financial Economics, 5(2): 189–200.

BERTSIMAS, D. D. B., (2009). Brown, Constructing uncertainty sets for robust linear optimization. Operations research, 57(6):1483–1495.

DENNEBERG, D., (1994). Non-additive measure and integral. Kluwer, Dordrecht.

DHAENE, J., WANG, S. S., YOUNG, V. R., GOOVAERTS, M. J., (2000). Comonotonicity and maximal stop-loss premiums. Bulletin of the Swiss Association of Actuaries, 2:99–113.

FENG, M. B., TAN, K. S., (2012). Coherent Distortion Risk Measures in Portfolio Selection, System Engineering Procedia 4, 25–34.

HADAR, J., RUSSELL, W., (1969). Rules for ordering uncertain prospects, American Economic Review 59, 25–34.

HARVEY, C. R. J., LIECHTY, M., LIECHTY, P., (2010). Müller. Portfolio selection with higher moments. Quantitative Finance, 10(5):469–485.

KUSUOKA, S., (2001). On law invariant coherent risk measures. Advances in mathematical economics, 3:83–95.

MARKOWITZ, H. M., (1952). Portfolio selection. The Journal of Finance, 7(1):77–91.

MARKOWITZ, H. M., (1959). Portfolio selection: efficient diversification of investments. New Haven, CT: Cowles Foundation, 94.

ORTOBELLI, S., (2001). The classification of parametric choices under uncertainty: analysis of the portfolio choice problem, Theory and Decision 51, 297–327.

ORTOBELLI, S., RACHEV, S., SHALIT, H., FABOZZI, F., (2006). Risk probability function- analysis and probability metrics applied to portfolio theory, http://www.statistik.unikarlsruhe.de.

ROCKAFELLAR, R. T., URYASEV, S., (2002). Conditional value-at-risk for general loss distributions. Journal of Banking & Finance, 26(7):1443–1471.

ROCKAFELLAR, R. T., URYASEV, S., (2000). Optimization of conditional value-at-risk. Journal of risk, 2:21–42.

ROTHSCHILD, M., STIGLITZ, J., (1970). Increasing risk. I.a. definition, Journal of Economic Theory 2, 225–243.

ROY, A. D., (1952). Safety first and the holding of assets. Econometrica: Journal of the Econometric Society, pages 431–449.

TRZPIOT, G., (2005). Partial Moments and Negative Moments in Ordering Asymmetric Distribution, in: Daniel Baier and Klaus-Dieter Wernecke (eds.): Innovations in Classification, Data Science and Information Systems, Proc. 27th Annual GFKL Conference, University of Cottbus, March 11-14 2003. Springer-Verlag, Heidelberg-Berlin, 181–188.

TRZPIOT, G., (2010). Pesymistyczna optymalizacja portfelowa [Pessimistic portfolio optimization], In: Modelling of preferences and risk '09, Scientific Papers, University of Economics in Katowice, 121–128

TRZPIOT, G., (2012). Własności transformujących miar ryzyka [Properties of transforming risk measures], Economic Studies of the University of Economics in Katowice - Faculty Scientific Papers No. 91, Katowice, 21–36

WANG, S. S., (2000). A class of distortion operators for pricing financial and insurance risks. The Journal of Risk and Insurance, 67(1):15–36.

WIRCH, J., HARDY, M. R., (2001). Distortion risk measures: coherence and stochastic dominance. Working paper.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 299–308

SELECTED TESTS COMPARING THE ACCURACY OF INFLATION RATE FORECASTS CONSTRUCTED BY DIFFERENT METHODS

Agnieszka Przybylska-Mazur[1]

ABSTRACT

The forecasts of macroeconomic variables including the forecasts of inflation rate play an important role in estimating future situation in the economy. Knowledge of effective forecasts allows making optimal business, financial and investment decisions. The forecasts of macroeconomic variables and as a result also inflation rate forecasts can be determined by different methods often giving different results. Therefore, in this paper we apply selected tests to the evaluation of the accuracy of inflation rate forecasts determined by different methods.

Key words: forecast accuracy, parametric tests, Morgan-Granger-Newbold test, Meese-Rogoff test and Diebold-Mariano test.

1. Introduction

The forecasts of macroeconomic variables and therefore also inflation rate forecasts can be determined by different methods often giving different results (Dittmann, 2008). The purpose of the paper is to apply selected statistical tests to the evaluation of the accuracy of inflation rate forecasts constructed by different methods. Of particular and practical importance are tests which do not need to know the model on which the forecasts that allow comparing the accuracy of forecasts constructed by different methods were determined. This group of parametric tests include: Morgan-Granger-Newbold test, Meese-Rogoff test and Diebold-Mariano test. For this group of tests - the model-free tests - we assume that we have the actual values and the set or sets of forecasts of the prediction.

At the beginning we present the tests which assumes the squared-error loss and zero-mean, serially uncorrelated forecast errors in the context of the application of this tests to the evaluation of the accuracy of inflation rate forecasts determined by different methods. Next, we present tests that are asymptotically valid under more general conditions allowing loss functions other than the

[1] Ph.D., Department of Statistical and Mathematical Methods in Economy, University of Economics in Katowice.

quadratic and covering situations when forecast errors are non-Gaussian, non-zero-mean, serially correlated, and contemporaneously correlated. These tests are applied also to the evaluation of the accuracy of inflation rate forecasts.

2. Preliminary notions

We assume that the available information consists of the following:

- actual values of the inflation rate $\pi_t, t=1,2,...,T$,
- two forecasts: $\hat{\pi}_{1t}, t=1,2,...,T$ and $\hat{\pi}_{2t}, t=1,2,...,T$.

We define the forecast errors as

$$Q_{it} = \hat{\pi}_{it} - \hat{\pi}_t \text{ for } i=1,2,\ t=1,2,...,T \tag{1}$$

Moreover, we assume that the loss associated with the forecast *i* is a function of the actual and forecast values only through the forecast error, and is denoted by:

$$L(\pi_t, \hat{\pi}_{it}) = L(\hat{\pi}_{it} - \pi_t) = L(Q_{it}) \tag{2}$$

The error loss function *L* can take various forms. Typically, we take into consideration the squared-error loss of the form $L(Q_{it}) = Q_{it}^2$ or the absolute error loss of the form $L(Q_{it}) = \left|Q_{it}\right|$.

We also denote the loss difference between the two forecasts by

$$d_t = L(Q_{1t}) - L(Q_{2t}) \text{ for } t=1,2,...,T \tag{3}$$

Since the tests are presented below verified forecast accuracy, now we define the concept of equality accuracy of inflation rate forecasts. We say that the two inflation rate forecasts have equal accuracy if and only if the loss difference has zero expectation for all *t*.

3. Application of Morgan-Granger-Newbold test to compare the accuracy of inflation rate forecasts

We can apply the Morgan-Granger-Newbold test when the inflation forecasts errors are:

- zero mean,
- Gaussian,
- serially uncorrelated,
- contemporaneously uncorrelated.

Furthermore, we assume the squared-error loss. Moreover, this test is applicable only to one-step predictions.

We would like to test the null hypothesis

$H_0 : E(d_t) = 0$ for all $t = 1, 2, ..., T$

versus the alternative hypothesis

$H_1 : E(d_t) = c \neq 0$.

Therefore, the test statistics is (Diebold, Mariano, 1995, Clements (ed.), Hendry (ed.), 2004):

$$MGN = \frac{r}{\sqrt{\frac{1-r^2}{T-1}}} \tag{4}$$

where:

$$r = \frac{x^T \cdot z}{\sqrt{(x^T \cdot x) \cdot (z^T \cdot z)}},$$

x is the $T \times 1$ matrix with *t*-th element x_t,

z is the $T \times 1$ matrix with *t*-th element z_t,

$x_t = Q_{1t} + Q_{2t}$, $z_t = Q_{1t} - Q_{2t}$.

The *MGN* statistics has a t-distribution with $T - 1$ degrees of freedom.

4. Use of Meese-Rogoff test to compare the accuracy of inflation rate forecasts

The Meese-Rogoff test is the test of equal forecast accuracy when the forecast errors are serially and contemporaneously correlated, have zero mean and are Gaussian. In this test we assume also the squared-error loss.

We would like to test the null hypothesis: $H_0 : E(d_t) = 0$ for all $t = 1, 2, ..., T$ versus the alternative hypothesis $H_1 : E(d_t) = c \neq 0$.

Verifying the null hypothesis of equal accuracy of inflation rate forecasts we use also the series: $x_t = Q_{1t} + Q_{2t}$, $z_t = Q_{1t} - Q_{2t}$ for $t = 1, 2, ..., T$.

The statistics for Meese-Rogoff test is then (Rossi, 2005)

$$MR = \frac{\hat{\gamma}_{xz}(0)}{\sqrt{\frac{\hat{\Omega}}{T}}} \tag{5}$$

where:

$$\hat{\gamma}_{xz}(0) = \frac{x^T \cdot z}{T},$$

$$\hat{\Omega} = \sum_{k=-m(T)}^{m(T)} \left(1 - \frac{|k|}{T}\right) \cdot [\hat{\gamma}_{xx}(k) \cdot \hat{\gamma}_{zz}(k) + \hat{\gamma}_{xz}(k) \cdot \hat{\gamma}_{zx}(k)]$$

$\hat{\gamma}_{xz}(k)$, $\hat{\gamma}_{zx}(k)$ - cross-autocovariances,

$\hat{\gamma}_{xx}(k)$, $\hat{\gamma}_{zz}(k)$ - own-autocovariances,

$\hat{\gamma}_{xz}(k) = \mathrm{cov}(x_t, z_{t-k})$

$\hat{\gamma}_{zx}(k) = \mathrm{cov}(z_t, x_{t-k})$

$\hat{\gamma}_{xx}(k) = \mathrm{cov}(x_t, x_{t-k})$

$\hat{\gamma}_{zz}(k) = \mathrm{cov}(z_t, z_{t-k})$

$m(T)$ - the truncation lag that increases with sample size *T*.

Given the maintained assumptions, the following result holds under the hypothesis of equal forecast accuracy $\sqrt{T} \cdot \hat{\gamma}_{xz}(0) \to N(0, \Omega)$ in distribution.

5. Application of Diebold-Mariano test to compare the accuracy of inflation rate forecasts

Diebold and Mariano (1995) consider model-free tests of inflation rate forecast accuracy that are directly applicable to non-quadratic loss functions, multi-period inflation rate forecasts, and inflation rate forecast errors that are non-Gaussian, non-zero-mean, serially correlated, and contemporaneously correlated.

We use this test when sample sizes are large.

The Diebold-Mariano test verify the null hypothesis $H_0 : E(d_t) = 0$ for all $t = 1, 2, ..., T$ versus the alternative hypothesis $H_1 : E(d_t) \neq 0$.

Assuming covariance stationarity of the process d_t, we have the following Diebold-Mariano statistics when the sample size is large (Diebold, Mariano, 1995):

$$DM = \frac{\bar{d}}{\sqrt{\dfrac{2\pi \cdot \hat{f}_d(0)}{T}}} \tag{6}$$

where:

$$\bar{d} = \frac{1}{T} \sum_{t=1}^{T} [L(Q_{1t}) - L(Q_{2t})]$$

$\hat{f}_d(0)$ is a consistent estimate of $f_d(0)$,

$$\hat{f}_d(0) = \frac{1}{2\pi} \sum_{k=-m(T)}^{m(T)} w\left(\frac{k}{m(T)}\right) \cdot \hat{\gamma}_d(k)$$

$$\hat{\gamma}_d(k) = \frac{1}{T} \sum_{t=|k|+1}^{T} (d_t - \bar{d}) \cdot (d_{t-|k|} - \bar{d})$$

$m(T)$ - the bandwidth or lag truncation that increases with *T*,

$w(\cdot)$ - the weighting scheme or kernel.

One weighting scheme, called the truncated rectangular kernel and used in Diebold and Mariano (1995), is the indicator function that takes the value of unity when the argument has an absolute value less than one, thus $w(x) = 1(|x| < 1)$.

The statistics $\sqrt{T} \cdot (\bar{d} - c) \rightarrow N(0, 2\pi \cdot f_d(0))$ in distribution, where: $f_d(\cdot)$ is the spectral density of d_t for $t = 1, 2, ..., T$, $f_d(\lambda) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma_d(k) \cdot e^{-ik\lambda}$ for $-\pi \le \lambda \le \pi$, $\gamma_d(k)$ is the autocovariance of d_t at displacement *k*.

The null hypothesis is rejected in favour of the alternative hypothesis when *DM*, in absolute value, exceeds the critical value of a standard unit Gaussian distribution.

Harvey, Leybourne and Newbold (1997) propose a small-sample modification of Diebold-Mariano test.

When we assume that the inflation rate forecast accuracy is measured in terms of mean-squared prediction error, and optimal *h*-step ahead inflation rate predictions are likely to have forecast errors that are $MA(h-1)$ moving average process of order $h-1$, we have autocovariances $\gamma(k) = 0$ for $k \ge h$ and $\hat{\gamma}_d(k) = \frac{1}{T} \sum_{t=k+1}^{T} (d_t - \bar{d}) \cdot (d_{t-k} - \bar{d})$ for $0 \le k \le h$.

Then, the test statistics that is the modification of *DM* test statistics, is the following (Harvey, Leybourne, Newbold, 1997):

$$DM^* = \frac{DM}{\sqrt{\frac{T + 1 - 2h + \frac{h(h-1)}{T}}{T}}} \tag{7}$$

To make a d ecision of rejection or acceptance of the null hypothesis the empirical value with critical value from the t-distribution with $(T-1)$ degrees of freedom should be compared.

6. Empirical analysis

Below we present the results of testing inflation rate forecast accuracy for monthly inflation rate determined on the basis of the autoregressive model and the traditional VAR monetary policy model, and also for the quarterly inflation rate that come from the reports "Inflation projection of the NBP based on the NECMOD model".

6.1. Comparison of the accuracy of inflation rate forecasts for the forecasts obtained from the autoregressive model and from the traditional VAR monetary policy model

When testing the equality of inflation rate forecasts accuracy we take into account one-step forecast monthly inflation rates determined on the basis of the first-order autoregression model and on the basis of the traditional VAR monetary policy model, which contains three variables: inflation rate, industrial production growth rate and reference rate. We assume the significance level equals 0,01. The data concerning the monthly forecast of inflation rates, the real values of inflation rate and the forecast errors are presented in the table below.

Table 1. Forecasts and real values of monthly inflation rate and forecast errors

Time	Inflation forecasts determined on the basis of the first-order autoregression model	Inflation forecasts determined on the basis of the traditional VAR monetary policy model	Real values of inflation rate	Q_{1t}	Q_{2t}
April 2011	4.59	4.55	4.5	0.09	0.05
May 2011	4.71	4.33	5	-0.29	-0.67
June 2011	4.76	4.83	4.2	0.56	0.63
July 2011	4.78	4.9	4.1	0.68	0.8
August 2011	4.79	4.54	4.3	0.49	0.24

Source: Own calculations.

Since the inflation forecasts errors do not have zero mean ($\overline{Q_{1t}} = 0.306$, $\overline{Q_{2t}} = 0.21$), then to test the equality of forecasts accuracy we use the modification of Diebold-Mariano test for a small sample proposed by Harvey, Leybourne and Newbold.

The null hypothesis and alternative hypothesis are the following:

$H_0 : E(d_t) = 0$ for all $t = 1, 2, ..., 5$

$H_1 : E(d_t) \neq 0$

The test statistics is given by (7)

$$DM^* = \frac{DM}{\sqrt{\frac{T + 1 - 2h + \frac{h(h-1)}{T}}{T}}}$$

Assuming $m(T) = 3$ we obtain DM = -1.28 and DM^*_{emp}= -1.43. Because the critical value read from the table of t-distribution with 4 degrees of freedom is equal to $DM_\alpha = t_{0,01;4} = 4{,}604$, then for this significance level we have $\left|DM^*_{emp}\right| < DM_\alpha$, thus there is no evidence to reject the null hypothesis of equal forecast accuracy of monthly inflation rates determined on the basis of the first-order autoregression model and on the basis of the traditional VAR monetary policy model.

6.2. Comparison of the accuracy of inflation rate forecasts and "Inflation projection of the NBP based on the NECMOD model"

When comparing the accuracy of inflation rate forecasts we now take into account forecasts of quarterly inflation rates provided in the report "Inflation projection of the NBP based on the NECMOD model". We assume the significance level equals 0.01.

The obtained data concerning the quarterly forecasts of inflation rate, the real values of inflation rate, the forecast errors and the loss difference are presented in the tables below.

Table 2. Forecasts and real values of quarterly inflation rate

Year	Quarter	Inflation forecasts from given report	Inflation forecasts from the next report	Real values of inflation rate
2008	I	4.2	4.3	4.3
	II	4.6	4.7	4.7
	III	3.8	3.8	3.8
2009	I	3.4	3.3	3.3
	II	3.3	3.7	3.7
	III	3.6	3.6	3.5
	IV	3.0	3.3	3.3

Table 2. Forecasts and real values of quarterly inflation rate (cont.)

Year	Quarter	Inflation forecasts from given report	Inflation forecasts from the next report	Real values of inflation rate
2010	I	2.6	3.0	3.0
	II	2.4	2.3	2.3
	III	2.1	2.2	2.2
	IV	2.9	2.9	2.9
2011	I	3.5	3.8	3.8
	II	4.3	4.6	4.6
	III	4.1	4.1	4.1
	IV	4.6	4.6	4.6
2012	I	4.3	4.1	4.1
	II	3.9	4.0	4.0
	III	3.9	3.8	3.9
	IV	3.1	2.9	2.9
2013	I	1.7	1.3	1.3
	II	1.4	0.6	0.5

Source: Report "Inflation projection of the NBP based on the NECMOD model".

Table 3. The forecast errors and the loss difference

Year	Quarter	Q_{1t}	Q_{2t}	Loss difference d_t
2008	I	0.0	-0.1	-0.01
	II	0.0	-0.1	-0.01
	III	0.0	0.0	0.00
2009	I	0.0	0.1	-0.01
	II	0.0	-0.4	-0.16
	III	0.1	0.1	0.00
	IV	0.0	-0.3	-0.09
2010	I	0.0	-0.4	-0.16
	II	0.0	0.1	-0.01
	III	0.0	-0.1	-0.01
	IV	0.0	0.0	0.00
2011	I	0.0	-0.3	-0.09
	II	0.0	-0.3	-0.09
	III	0.0	0.0	0.00
	IV	0.0	0.0	0.00
2012	I	0.0	0.2	-0.04
	II	0.0	-0.1	-0.01
	III	-0.1	0.0	0.01
	IV	0.0	0.2	-0.04
2013	I	0.0	0.4	-0.16
	II	0.1	0.9	-0.80

Source: Own calculations.

In this case to compare the forecasts accuracy we use the modification of Diebold-Mariano test for a small sample.

The null hypothesis and the alternative hypothesis are as follow:

$H_0 : E(d_t) = 0$ for all $t = 1, 2, ..., 21$

$H_1 : E(d_t) \neq 0$

The test statistics is given by (7)

Assuming $m(T) = 5$ we obtain $DM = 2{,}22$ and $DM^*_{emp} = 2{,}39$. Because the critical value read from the table of t-distribution with 20 degrees of freedom is equal to $DM_\alpha = t_{0{,}01;20} = 2{,}845$, then for the significance level which equals 0.01 we have $\left|DM^*_{emp}\right| < DM_\alpha$. Subsequently, there is no evidence to reject the null hypothesis of equal forecast accuracy of quarterly forecasts of inflation rate determined on the basis of NECMOD model. Therefore, all determined forecasts have equal accuracy. The differences in values result from the change in the assumptions about the projections in the individual reports.

7. Conclusion

It follows from the analyses that the most frequently used test for the comparison of the accuracy of inflation rate forecasts (the forecasts constructed by different methods) is the modification of Diebold-Mariano test for a small sample proposed by Harvey, Leybourne and Newbold. It can be concluded that there is no evidence to reject the null hypothesis of equal forecast accuracy of monthly inflation rates determined on the basis of the first-order autoregression model and on the basis of the traditional VAR monetary policy model. We also conclude that the quarterly forecasts of inflation rate determined on the basis of NECMOD model and presented in two subsequent reports have equal accuracy. The differences in values result from the change in the assumptions about the projections in the individual reports.

REFERENCES

CLEMENTS, M. P. (Eds.), HENDRY D. F. (Eds.), (2004). A Companion to Economic Forecasting, Wiley-Blackwell.

DIEBOLD, F. X., MARIANO R. S., (1995). Comparing Predictive Accuracy, Journal of Business and Economic Statistics, 13, pp. 253–265.

DITTMANN, P., (2008). Forecasting in enterprise. Methods and their use, Oficyna Ekonomiczna Publishing House.

HARVEY, D., LEYBOURNE, S., NEWBOLD, P., (1997). Testing the equality of prediction mean squared errors. International Journal of Forecasting, 13, pp. 281–291.

ROSSI, B., (2005). Testing Long-horizon Predictive Ability with High Persistence, and the Meese-Rogoff Puzzle, International Economic Review, vol. 46, issue 1, pp. 61–92.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 309–316

TESTS FOR CONNECTION BETWEEN CLUSTERING OF POLISH COUNTIES AND PROVINCE STRUCTURE

Małgorzata Markowska[1], **Marek Sobolewski**[2], **Andrzej Sokołowski**[3], **Danuta Strahl**[4]

ABSTRACT

The general idea of statistical tests which allow testing the influence of geographical or administrative units of upper level on clustering results of lower level units is presented, basing on the authors' earlier works. The so-called "active border" notion is used in these methods. If two counties *(powiats)* have been classified into different clusters then the border between them is called active. This border can be also the border between provinces. The number and length of active borders are used in the proposed test statistics. Their distribution depends on the actual geographic division of a given country. In this paper we present results for Poland and division for provinces and counties. Tables for test critical values and the approximation functions are given.

Key words: cluster analysis, NUTS, comparing partitions.

1. Introduction

In Sokolowski *et al.* (2013a), (2013b) the general idea of statistical tests which allow testing the influence of geographical or administrative units of upper level on clustering results of lower level units was presented. The so-called "active border" notion is used in these methods. If two counties have been classified into different clusters then the border between them is called active. This border can be also the border between provinces. If the upper level has no influence on the lower level partition results, then only randomness should decide if the active border is also the upper level border. Distribution of test statistic depends on the actual geographic division of a given country. In this paper we present results for Poland and division for provinces and counties. There are 978 borders between counties and 210 of them are also elements of between-province

[1] Wroclaw University of Economics. E-mail: malgorzata.markowska@ue.wroc.pl.
[2] Rzeszow University of Technology. E-mail: mareksobol@poczta.onet.pl.
[3] Cracow University of Economics. E-mail: andrzej.sokolowski@uek.krakow.pl.
[4] Wroclaw University of Economics. E-mail: danuta.strahl@ue.wroc.pl.

border. The total length of borders between counties sums up to 88,869 km, including 16,062 km of borders between provinces.

2. Test statistics

Both of the proposed test statistics are being used for testing the same following hypotheses:

H_0: Province level has no influence on the partition results obtained for counties

H_1: Province level influences significantly the results of counties partition

It seems natural that if the upper level has no influence on the lower level partition results, then only randomness should decide if the active border is also the upper level border. If "too many" active borders between counties are also borders between provinces one should reject the null hypothesis. Thus, the proposed tests have right-sided critical region. The following two test statistics are considered:

$$L1 = \frac{Number\ of\ active\ borders\ which\ are\ borders\ between\ provinces}{Total\ number\ of\ observed\ active\ borders} \tag{1}$$

$$L2 = \frac{Length\ of\ active\ borders\ which\ are\ borders\ between\ provinces}{Total\ length\ of\ observed\ active\ borders} \tag{2}$$

3. Simulation study

For partition simulations we use the simplified version (with given number of groups) of random partition generator proposed by Sokolowski (1979):

- set *k* (number of groups),
- assign random number from uniform distribution to each object,
- order objects according to values of this random variable,
- now we have *(n-1)* potential borders between objects,
- assign random number from uniform distribution to each potential border,
- make "active" first *k* borders with the biggest values of these random numbers.

We have considered partitions of Poland's counties from 2 to 16 groups. With 1000 simulation runs we have found that the distribution of *L1* and *L2* statistics can be approximated by normal distribution. Empirical distribution for *k=5* as an example is presented on Fig 1.

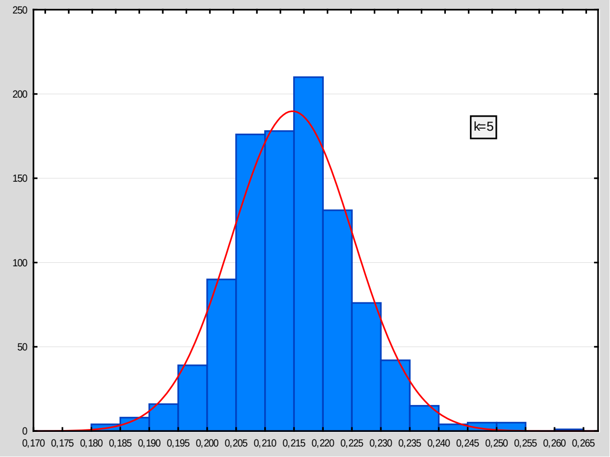

Figure 1. Empirical distribution of L1 statistic under null

Expected value of L1 equals 210/978=0.2147, while standard distribution depends on the number of clusters, but it can be very well approximated by (1), see Fig.2

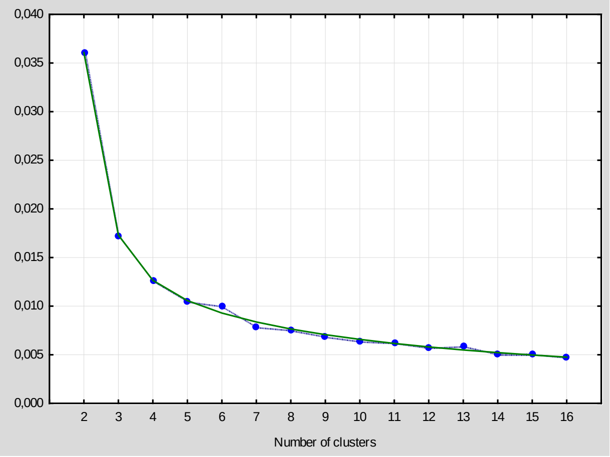

Figure 2. Standard deviation of L1 depending on the number of clusters

Critical values for 0.05 and 0.10 significance levels can also be approximated by fractional polynomials. Fig. 3 gives just one example of goodness-of-fit.

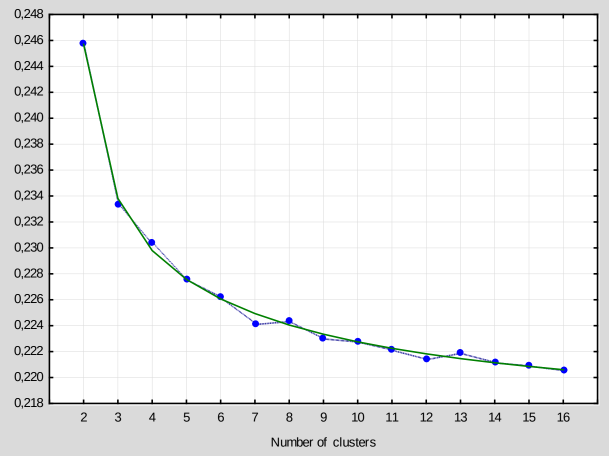

Figure 3. Critical value of L1 approximation for α=0.10

We have found that the distribution of L2 statistic can be also approximated by normal distribution. Expected value of L2 under null equals 16062/88869=0.1807 while standard deviation and critical values follows well fitted functions. Approximating functions are given in Table 1 and smoothed critical values in Table 2.

Table 1. Approximating functions

Parameter	Function	Adjusted coefficient of determination	Standard error of residuals
SD(L1)	$0.000802+0.074792k^{-1}-0.210665k^{-2}+0.404299k^{-3}$	0.999	0.0003
$Q_{0.90}$(L1)	$0.216349+0.076325k^{-1}-0.145002k^{-2}+0.219535k^{-3}$	0.996	0.0004
$Q_{0.95}$(L1)	$0.215771+0.119359k^{-1}-0.289479k^{-2}+0.446677k^{-3}$	0.995	0.0006
SD(L2)	$0.002043+0.052128k^{-1}-0.098926k^{-2}+0.230704k^{-3}$	0.999	0.0002
$Q_{0.90}$(L2)	$0.184717+0.037557k^{-1}+0.032073k^{-2}$	0.996	0.0004
$Q_{0.95}$(L2)	$0.184124+0.080401k^{-1}-0.121491k^{-2}+0.233468k^{-3}$	0.997	0.0005

Table 2. Critical values

Number of clusters	L1		L2	
	α=0.10	α=0.05	α=0.10	α=0.05
2	0.2457	0.2589	0.2115	0.2231
3	0.2338	0.2399	0.2008	0.2061
4	0.2298	0.2345	0.1961	0.2003
5	0.2276	0.2316	0.1935	0.1972
6	0.2261	0.2297	0.1919	0.1952
7	0.2249	0.2282	0.1907	0.1938
8	0.2241	0.2270	0.1899	0.1927
9	0.2233	0.2261	0.1893	0.1919
10	0.2228	0.2253	0.1888	0.1912
11	0.2223	0.2246	0.1884	0.1906
12	0.2218	0.2240	0.1881	0.1901
13	0.2215	0.2234	0.1878	0.1897
14	0.2211	0.2230	0.1876	0.1893
15	0.2209	0.2226	0.1874	0.1890
16	0.2206	0.2222	0.1872	0.1887

4. Example

We have taken four variables characterizing Polish counties: number of births per 1000 population, unemployment rate, average salary and number of new flats per 1000 population. Ward’s agglomerative clustering method suggests the division into seven clusters (see Fig. 4)

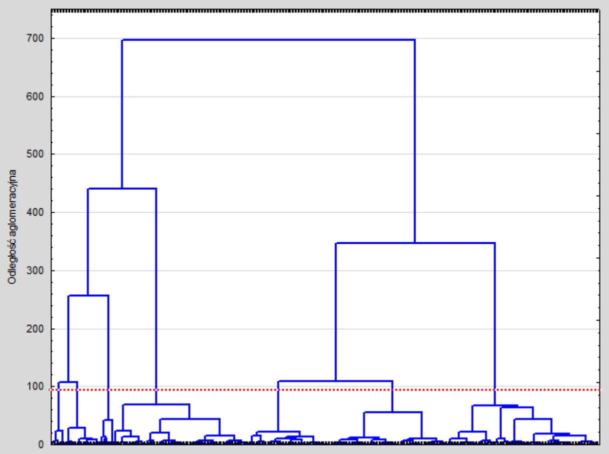

Figure 4. Ward's dendrogram of Polish counties

On Fig. 5 we can see the geographical distribution of clusters together with borders between provinces.

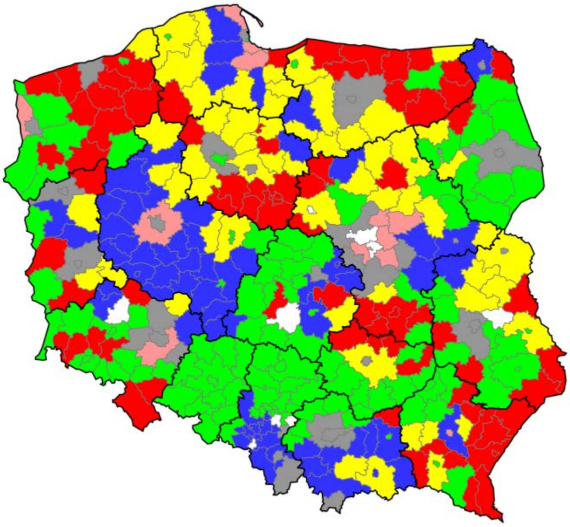

Figure 5. Clusters of Polish counties

Observed value of L1 equals 0.215 while critical value for α=0.10 is 0.2249 (p=0.450), and L2=0.185 with critical value 0.1907 and p=0.273. It is clear from both test statistics that there is no proof for statistically significant influence of province level on counties partition based on four considered variables.

5. Conclusions

It has been found that critical values for both proposed test statistics can be very well approximated by relatively simple functions while testing the influence of voivodship level of Polish provinces on county level. The example provided is only an illustrative effort. The proposed test can be widely used in testing the relations between administrative levels in Poland with respect to economic phenomena, politics, public administration and quality of life.

Acknowledgements

The project has been financed by the Polish National Centre for Science, decision DEC-2013/09/B/HS4/00509.

REFERENCES

SOKOŁOWSKI, A., (1979). Generowanie losowego podziału zbioru skończonego. Prace Naukowe Akademii Ekonomicznej we Wrocławiu [Generating random distribution of a finite set. Scientific Papers of Wrocław University of Economics], 1979, 160 (182), 413–415.

SOKOŁOWSKI, A., STRAHL, D., MARKOWSKA, M., SOBOLEWSKI, M., (2013a). The influence of upper level NUTS on lower level classification of EU regions. European Conference on Data Analysis, Luxembourg, July 10–12, 2013.

SOKOŁOWSKI, A., STRAHL, D., MARKOWSKA, M., SOBOLEWSKI, M., (2013b). The hierarchy test of geographic units based on bor der lengths, Conference of the International Federation of Classification Societies IFCS 2013, Tilburg, Netherlands, July 14–17, 2013. Abstract in Program and Book of Abstracts, Tilburg University, 46.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 317–326

ON CERTAIN A-OPTIMAL BIASED SPRING BALANCE WEIGHING DESIGNS

Bronisław Ceranka[1], **Małgorzata Graczyk**[2]

ABSTRACT

In the paper, the estimation of unknown measurements of *p* objects in the experiment, according to the model of the spring balance weighing design, is discussed. The weighing design is called biased if the first column of the design matrix has elements equal to one only. The A-optimal design is a design in which the trace of the inverse of information matrix is minimal. The main result is the broadening of the class of experimental designs so that we are able to determine the regular A-optimal design. We give the lowest bound of the covariance matrix of errors and the conditions under which this lowest bound is attained. Moreover, we give new construction methods of theregular A-optimal spring balance weighing design based on the incidence matrices of the balanced incomplete block designs. The example is also given.

Key words: A-optimal design, spring balance weighing design.

1. Introduction

Let us consider $\mathbf{\Phi}_{n\times p}(0,1)$, the class of all possible $n \times p$ matrices of the elements equal to zero or one and, moreover, the first column of this matrix consists only of ones. Any matrix $\mathbf{X} \in \mathbf{\Phi}_{n\times p}(0,1)$ is called the design matrix of the biased spring balance weighing design if the result of the experiment we are able to present in the form $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \mathbf{e}$, where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{w}^*$ is a $p \times 1$ vector of unknown parameters and $\mathbf{e}$ is an $n \times 1$ vector of random errors. Furthermore, assume that $\mathrm{E}(\mathbf{e}) = \mathbf{0}_n$ and $\mathrm{Var}(\mathbf{e}) = \sigma^2\mathbf{G}$, where $\mathbf{0}_n$ is vector of zeros, σ^2 is the constant variance of errors, $\mathbf{G}$ is the $n \times n$ symmetric positive definite diagonal matrix of known elements.

[1] Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland. E-mail: bronicer@up.poznan.pl.

[2] Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, Poland. E-mail: magra@up.poznan.pl.

The normal equations estimating $\mathbf{w}^*$ are of the form $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\hat{\mathbf{w}}^* = \mathbf{X}'\mathbf{G}^{-1}\mathbf{y}$, where $\hat{\mathbf{w}}^*$ is the vector of the weights estimated by the least squares method. Any weighing design is nonsingular if the matrix $\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}$ is nonsingular. It is obvious that $\mathbf{G}$ is the symmetric positive definite matrix, and any weighing design is nonsingular if and only if the matrix $\mathbf{X}'\mathbf{X}$ is nonsingular and then in that case all the parameters are estimable. The estimator of the vector representing unknown measurements of objects $\mathbf{w}^*$ is equal to $\hat{\mathbf{w}}^* = \left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{y}$ assuming that $\mathbf{X}$ is of full column rank. The covariance matrix of $\hat{\mathbf{w}}^*$ is given by $\mathrm{Var}\left(\hat{\mathbf{w}}^*\right) = \sigma^2\left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)^{-1}$.

In the special case of experimental designs, when bias is present, then $\mathbf{w}^* = \left[w_1 \quad \mathbf{w}\right]'$ is the $p \times 1$ vector of unknown measurements of objects, w_1 is the parameter corresponding to the bias (systematic error), $\mathbf{w} = \left[w_2 \quad w_3 \quad \cdots \quad w_p\right]'$ is the $(p-1)\times 1$ vector of unknown measurements of object excluding bias. In such experiment we assume that there is one object whose value is estimated by taking the column of ones in the design matrix $\mathbf{X}$ corresponding to the bias. Thus, we consider the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n\times p}(0,1)$ in the following form

$$\mathbf{X} = \left[\mathbf{1}_n \quad \mathbf{X}_1\right], \tag{1}$$

where $\mathbf{1}_n$ is $n \times 1$ vector of ones, $\mathbf{X}_1$ is $n \times (p-1)$ matrix of elements equal to zero or one.

It is worth emphasizing that for each pattern of $\mathbf{G}$ the conditions determining optimal design must be separately investigated. For the case $\mathbf{G} = \mathbf{I}_n$, Banerjee (1975), Raghavarao (1971) and Katulska (1989) present the problems related to the biased spring balance weighing designs. Some considerations connected with the diagonal covariance matrix of errors $\sigma^2\mathbf{G}$ are presented in Ceranka and Katulska (1990, 1992).

2. The main result

The statistical problem is to determine the most efficient design in some sense by a proper choice of the design matrix $\mathbf{X}$ among many at our disposals in $\mathbf{\Phi}_{n\times p}(0,1)$. Some optimal criteria have been considered in the literature, see Pukelsheim (1993). One of them is A-optimality which minimizes the average variance of the estimator of unknown measurements of the objects.

For the case $\mathbf{G}$ there is a positive definite diagonal matrix of known elements. The problems related to the regular A-optimal biased spring balance weighing design have been considered in the literature, see, for instance, Graczyk (2011). In this paper, the following definition is presented.

Definition 1. Any nonsingular $\mathbf{X} \in \mathbf{\Phi}_{n \times p}(0,1)$ of the form (1) with the diagonal covariance matrix of errors $\sigma^2\mathbf{G}$ is called the regular A-optimal biased spring balance weighing design for estimation of $\hat{\mathbf{w}}$ if $\mathrm{tr}(\mathrm{Var}(\hat{\mathbf{w}})) = \sigma^2 \frac{4(p-1)}{\mathrm{tr}(\mathbf{G}^{-1})}$.

In addition, in the same paper, the following corollaries are presented.

Corollary 1. Any nonsingular $\mathbf{X} \in \mathbf{\Phi}_{n \times p}(0,1)$ of the form (1) with the diagonal covariance matrix of errors $\sigma^2\mathbf{G}$ is called the regular A-optimal biased spring balance weighing design for estimation of $\hat{\mathbf{w}}$ if and only if $\mathbf{X}_1^{'}\mathbf{G}^{-1}\mathbf{X}_1 = \frac{\mathrm{tr}(\mathbf{G}^{-1})}{4}\left(\mathbf{I}_{p-1} + \mathbf{1}_{p-1}\mathbf{1}_{p-1}^{'}\right)$.

Corollary 2. In the regular A-optimal biased spring balance weighing design $\mathbf{X} \in \mathbf{\Phi}_{n \times p}(0,1)$ of the form (1) with the diagonal covariance matrix of errors $\sigma^2\mathbf{G}$, $\mathrm{Var}(\hat{w}_1) = \frac{p\sigma^2}{\mathrm{tr}(G^{-1})}$, where $\hat{w}_1$ is the estimator of the bias.

In the present paper, we construct the regular A-optimal biased spring balance weighing design with the covariance matrix of errors $\sigma^2\mathbf{G}$ for $\mathbf{G}$ of the form

$$\mathbf{G} = \begin{bmatrix} g^{-1}\mathbf{I}_c & \mathbf{0}_c\mathbf{0}_d^{'} & \mathbf{0}_c\mathbf{0}_{n-c-d}^{'} \\ \mathbf{0}_d\mathbf{0}_c^{'} & g_1^{-1}\mathbf{I}_d & \mathbf{0}_d\mathbf{0}_{n-c-d}^{'} \\ \mathbf{0}_{n-c-d}\mathbf{0}_c^{'} & \mathbf{0}_{n-c-d}\mathbf{0}_d^{'} & \mathbf{I}_{n-c-d} \end{bmatrix}, \quad g, g_1 > 0, c, d \geq 0. \qquad (2)$$

Suppose further that the design $\mathbf{X} \in \mathbf{\Phi}_{n \times p}(0,1)$ is partitioned in the same way as the matrix $\mathbf{G}$, i.e. we have

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_c & \mathbf{0}_c\mathbf{0}_{p-1}^{'} \\ \mathbf{1}_d & \mathbf{1}_d\mathbf{1}_{p-1}^{'} \\ \mathbf{1}_{n-c-d} & \mathbf{X}_2 \end{bmatrix}, \quad c, d \geq 0. \qquad (3)$$

In the special case $c = 0$ (or $d = 0$), the respective element of the matrix does not exist. That way we obtain the Theorem.

Theorem 1. Any nonsingular biased spring balance weighing design $\mathbf{X} \in \mathbf{\Phi}_{n\times p}(0,1)$ given by (3) with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2), is the regular A-optimal if and only if

$$\mathbf{X}_2'\mathbf{X}_2 = \frac{n+c(g-1)+d(g_1-1)}{4}\mathbf{I}_{p-1} + \frac{n+c(g-1)-d(3g_1+1)}{4}\mathbf{1}_{p-1}\mathbf{1}_{p-1}', \quad (4)$$

$g, g_1 > 0,\ c, d \ge 0$.

Proof. If $\mathbf{X}_1 = \left[\mathbf{0}_{p-1}\mathbf{0}_c' \quad \mathbf{1}_{p-1}'\mathbf{1}_d \quad \mathbf{X}_2'\right]'$ then $\mathbf{X}_1'\mathbf{G}^{-1}\mathbf{X}_1 = g_1 d\mathbf{1}_{p-1}\mathbf{1}_{p-1}' + \mathbf{X}_2'\mathbf{X}_2$. Consequently, in that case $\mathrm{tr}\left(\mathbf{G}^{-1}\right) = n + c(g-1) + d(g_1-1)$. From the above and from Corollary 1, $\mathbf{X}_1'\mathbf{G}^{-1}\mathbf{X}_1 = \frac{n+c(g-1)+d(g_1-1)}{4}\left(\mathbf{I}_{p-1} + \mathbf{1}_{p-1}\mathbf{1}_{p-1}'\right)$. This gives

$$\mathbf{X}_2'\mathbf{X}_2 = \frac{n+c(g-1)+d(g_1-1)}{4}\mathbf{I}_{p-1} + \frac{n+c(g-1)-d(3g_1+1)}{4}\mathbf{1}_{p-1}\mathbf{1}_{p-1}',$$

when $g, g_1 > 0,\ c, d \ge 0$. Hence the Theorem follows.

It is worth noting that the condition (4) implies that for the matrix $\mathbf{X}_2'\mathbf{X}_2$, diagonal elements satisfy the condition $n + c(g-1) - d(g_1+1) \equiv 0 \bmod (2)$ and off-diagonal elements satisfy the condition $n + c(g-1) + d(g_1-1) \equiv 0 \bmod (4)$. Afterwards, we have to determine the matrix $\mathbf{X}_2$ of elements equal to 1 or 0 which satisfied these conditions. Several methods of construction of the design matrix of the optimal spring balance weighing design are presented in the literature. Some of them are based on the incidence matrices of known block designs, another ones rely on using some algorithms.

In the following part of the paper we present the problem of constructing a regular A-optimal biased spring balance weighing design based on the incidence matrix of the balanced incomplete block design.

3. Regular A-optimal designs

Here, we present the application of the incidence matrix of balanced incomplete block design to the construction of the design matrix $\mathbf{X} \in \mathbf{\Phi}_{n\times p}(0,1)$ of the regular A-optimal spring balance weighing design. The balanced incomplete block design with the parameters v, b, r, k, λ is the design where we replace v objects in b blocks, each of size k. That is why each object occurs r times altogether and each pair of different objects occurs together in λ blocks. For more details see Raghavarao and Padgett (2005). Let us denote by $\mathbf{N}$ the

$v \times b$ incidence matrix of the binary incomplete block design. Then the matrix $\mathbf{X}$ is shown by the equations

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_c & \mathbf{0}_c\mathbf{0}_v^{'} \\ \mathbf{1}_d & \mathbf{1}_d\mathbf{1}_v^{'} \\ \mathbf{1}_b & \mathbf{N}^{'} \end{bmatrix}, \; c,d \geq 0 . \tag{5}$$

$\mathbf{X} \in \mathbf{\Phi}_{n \times p}(0,1)$ in the form (5) is the matrix of the biased spring balance weighing design. In this design we determine unknown measurements of $p = v + 1$ in $n = b + c + d$ measurement operations.

Theorem 2. The biased spring balance weighing design $\mathbf{X} \in \mathbf{\Phi}_{n \times p}(0,1)$ given by (5) with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2) is the regular A-optimal if and only if

$$b + cg - 3dg_1 \equiv 0 \,\mathrm{mod}(4) \tag{6}$$

and

$$r = 2\lambda + dg_1 . \tag{7}$$

Proof. The main idea of the proof is to show that any biased spring balance weighing design $\mathbf{X}$ given by (5) with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2), is the regular A-optimal if and only if $\mathbf{NN}^{'} = \frac{b + cg + dg_1}{4}\mathbf{I}_v + \frac{b + cg - 3dg_1}{4}\mathbf{1}_v\mathbf{1}_v^{'}$, $c,d \geq 0$, which follows from Theorem 1. Assume the formula $\mathbf{NN}^{'} = (r - \lambda)\mathbf{I}_v + \lambda\mathbf{1}_v\mathbf{1}_v^{'}$ holds, then we obtain the equalities (6) and (7) that is our claim.

Theorem 3. If there exists a balanced incomplete block design with the parameters v, b, r, k, λ and the design matrix $\mathbf{X}$ given by (5) is the matrix of the regular A-optimal biased spring balance weighing design with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2), then

$$\begin{aligned} v &= \frac{(4\lambda + 3dg_1 - cg)(\lambda + dg_1)}{dg_1\lambda + cg\lambda + d^2g_1^2} \\ b &= 4\lambda + 3dg_1 - cg \\ r &= 2\lambda + dg_1 \\ k &= \frac{(2\lambda + dg_1)(\lambda + dg_1)}{dg_1\lambda + cg\lambda + d^2g_1^2} \end{aligned} \; . \tag{8}$$

Proof. Let us first observe that from (6) and (7) it follows that $r = 2\lambda + dg_1$ and $b = 4\lambda + 3dg_1 - cg$. The proof is completed by showing that if the parameters v, b, r, k, λ of the balanced incomplete block design satisfy the conditions $vr = bk$ and $\lambda(v-1) = r(k-1)$, then v and k are given as in (8).

We have seen in Theorem 3 that if the parameters of the balanced incomplete block design satisfy the condition (8) then the biased spring balance weighing design $\mathbf{X}$ given by formula (5) with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2), is the regular A-optimal. The parameters v, b, r, k, λ must be positive integers as the parameters of the balanced incomplete block design. From the above reasoning and the condition (8) we obtain the theorem.

Theorem 4. For any positive definite integer λ and integers $c, d \geq 0$ the parameters v, b, r, k given by (8) are positive integers if and only if one of the following conditions holds:

(i) $d = 0$, $c > 0$, $2\lambda \equiv 0 \bmod(cg)$,

(ii) $d > 0$, $c = 0$, $2\lambda \equiv 0 \bmod(dg_1)$,

(iii) $d > 0$, $c = \dfrac{2(\lambda + dg_1)}{g}$, $2(\lambda + dg_1) \equiv 0 \bmod(g)$,

(iv) $dg_1 = cg$, $\lambda \equiv 0 \bmod(dg_1)$.

Proof. It is sufficient to show that from the condition (8) it follows that $v = 2k + a$, where $a = \dfrac{(dg_1 - cg)(\lambda + dg_1)}{dg_1\lambda + cg\lambda + d^2g_1^2}$ is such an integer that v is a positive integer. If $d = 0$, $c > 0$ then $a = -1$ and $k = \dfrac{2\lambda}{cg}$. It implies that $2\lambda \equiv 0 \bmod(cg)$, i.e. the condition (i) is fulfilled. By similar arguments, if $d > 0$, $c = 0$ then $a = 1$ and $k = 1 + \dfrac{2\lambda}{dg_1}$, and it implies that $2\lambda \equiv 0 \bmod(dg_1)$, i.e. the condition (ii) holds. If $d > 0$, $c = \dfrac{2(\lambda + dg_1)}{g}$, then $2(\lambda + dg_1) \equiv 0 \bmod(g)$, $a = -1$, $k = 1$, which means the condition (iii) is true. If $dg_1 = cg$ then $a = 0$ and $k = 1 + \dfrac{\lambda}{dg_1}$, it implies $\lambda \equiv 0 \bmod(dg_1)$. If one of the conditions (i)-(iv) is fulfilled then it is obvious that v, b, r, k are positive integers, which is the desired conclusion.

Based on the theoretical results presented in Theorems 3 and 4 we can formulate the following Corollaries.

Corollary 3. If there exists a balanced incomplete block design with the parameters v, b, r, k, λ and the matrix $\mathbf{X}$ given by (5) is the matrix of the regular A-optimal biased spring balance weighing design with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2), then

(i) $v = 2k - 1$, $b = 4\lambda - cg$, $r = 2\lambda$, $k = \frac{2\lambda}{cg}$ if $c > 0$ and $2\lambda \equiv 0 \operatorname{mod}(cg)$,

(ii) $v = 2k + 1$, $b = 4\lambda + 3dg_1$, $r = 2\lambda + dg_1$, $k = 1 + \frac{2\lambda}{dg_1}$ if $d > 0$ and $2\lambda \equiv 0 \operatorname{mod}(dg_1)$,

(iii) $v = 2k$, $b = 2(2\lambda + dg_1)$, $r = 2\lambda + dg_1$, $k = 1 + \frac{\lambda}{dg_1}$ if $d > 0$ and $\lambda \equiv 0 \operatorname{mod}(dg_1)$.

Corollary 4. If $d > 0$, $c = 2g^{-1}(\lambda + dg_1)$, $2(\lambda + dg_1) \equiv 0 \operatorname{mod}(g)$, then $v = k = 1$, $b = r = 2\lambda + dg_1$ and $\mathbf{X} = \begin{bmatrix} \mathbf{1}_c & \mathbf{0}_c \\ \mathbf{1}_d & \mathbf{1}_d \\ \mathbf{1}_b & \mathbf{1}_b \end{bmatrix}$ is the matrix of the regular A-optimal biased spring balance weighing design with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2), for two objects.

We have seen in Corollary 3 that if the parameters of the balanced incomplete block design are of the form (i)-(iii), then a biased spring balance weighing design $\mathbf{X}$, given by (5) with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2), is the regular A-optimal. Then we obtain the series of the parameters of the balanced incomplete block designs. Based on these parameters we form the incidence matrix $\mathbf{N}$ and then the design matrix $\mathbf{X}$.

Corollary 5. Let $d = 0$ and let $\mathbf{N}$ be the incidence matrix of the balanced incomplete block design with the parameters

(i) $v = 4t - 1$, $b = cg(4t - 1)$, $r = 2cgt$, $k = 2t$, $\lambda = cgt$, $t = 1,2,...$, for odd cg,

(ii) $v = 2t - 1$, $b = cg(2t - 1)$, $r = cgt$, $k = t$, $\lambda = \frac{cgt}{2}$, $t = 2,3,...$, for even cg,

then the matrix $\mathbf{X}$ given by (5) is the regular A-optimal biased spring balance weighing design with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2).

Corollary 6. Let $c = 0$ and let $\mathbf{N}$ be the incidence matrix of the balanced incomplete block design with the parameters

(i) $v = 4t + 3$, $b = dg_1(4t + 3)$, $r = dg_1(2t + 1)$, $k = 2t + 1$, $\lambda = dg_1 t$ for odd dg_1,

(ii) $v = 2t + 1$, $b = dg_1(2t + 1)$, $r = dg_1(t + 1)$, $k = t + 1$, $\lambda = \frac{dg_1 t}{2}$ for even dg_1,

$t = 1,2,...$, then the matrix $\mathbf{X}$ given by (5) is the regular A-optimal biased spring balance weighing design with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2).

Corollary 7. Let $dg_1 = cg$ and let $\mathbf{N}$ be the incidence matrix of the balanced incomplete block design with the parameters $v = 2(t + 1)$, $b = 2dg_1(2t + 1)$, $r = dg_1(2t + 1)$, $k = t + 1$, $\lambda = dg_1 t$, $t = 1,2,...$, then the matrix $\mathbf{X}$ given by (5) is the regular A-optimal biased spring balance weighing design with the covariance matrix of errors $\sigma^2\mathbf{G}$, where $\mathbf{G}$ is given by (2).

4. Example

Weighing designs can be applied in all experiments in which the experimental factors are on two levels. Let us suppose we study four marketing factors: the kind of advertisement (television or outdoor), assortment (basic or complementary), personal promotion (present or not), sale (directly by producer, mail-order). We will study the level of sale of chosen product on the basis on the nationwide range. From the statistical point of view, we are interested in determining the influences of these factors using sixteen different combinations. In the notation of weighing designs we determine unknown measurements of $p = 5$ objects in $n = 16$ measurements, so $\mathbf{X} \in \mathbf{\Phi}_{16\times 5}(0,1)$. In order to illustrate the theory given above we consider the case that we compare the influence of these factors in three different cities. Thus, the variance matrix of errors $\sigma^2\mathbf{G}$ is given by the matrix $\mathbf{G} = \begin{bmatrix} \frac{1}{2} & \mathbf{0}_3' & \mathbf{0}_{12}' \\ \mathbf{0}_{12} & \frac{3}{2}\mathbf{I}_3 & \mathbf{0}_3\mathbf{0}_{12}' \\ \mathbf{0}_{12} & \mathbf{0}_{12}\mathbf{0}_3' & \mathbf{I}_{12} \end{bmatrix}$ for $c = 1, d = 3, g = 2, g_1 = \frac{2}{3}$.

Moreover, $\mathrm{tr}(\mathbf{G}^{-1}) = 16$. Then we form the design matrix $\mathbf{X}$ of the form (5) for the case $c, d > 0$. That is why we consider $\mathbf{N} = \mathbf{1}_2 \otimes \mathbf{N}_1$, where $\mathbf{N}_1$ is the incidence matrix of the balanced incomplete block design with the parameters

$v=4$, $b=6$, $r=2$, $k=3$, $\lambda=1$ given by $\mathbf{N}_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$. Thus we obtain

$$\mathbf{X} = \begin{bmatrix} 1 & \mathbf{0}_4' \\ \mathbf{1}_3 & \mathbf{1}_3\mathbf{1}_4' \\ \mathbf{1}_{12} & \mathbf{N}' \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}'$$ and

$\mathbf{X}_1'\mathbf{G}^{-1}\mathbf{X}_1 = 4\left[\mathbf{I}_4 + \mathbf{1}_4\mathbf{1}_4'\right]$, i.e. the design $\mathbf{X}$ is the regular A-optimal (see Corollary 1). The first column of the design matrix $\mathbf{X}$ responds to the influence of a nationwide range, the second one to the kind of advertisement, the third one to the influence of the assortment. The next column exposes a p ersonal promotion, and finally the kind of sale. The form of the matrix can be interpreted in the following sense: the eighth row indicates that we take the nationwide range, a basic assortment and a personal promotion. Let us suppose $\mathbf{y}$ be the 16×1 vector of the results of the experiment. Thus $\hat{\mathbf{w}}^* = \left(\mathbf{X}'\mathbf{G}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{G}^{-1}\mathbf{y} =$

$$\frac{1}{24}\begin{bmatrix} 15 & -3 & -3 & -3 & 15 & 15 & 15 & 15 & 15 & 15 & 15 & 15 & 15 & 15 & 15 & 15 \\ -6 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ -6 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ -6 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \\ -6 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 & 3 \end{bmatrix}\mathbf{y}.$$

5. Discussion

In the paper, some problems related to A-optimality criterion are presented. The special class of experimental designs, i.e. biased spring balance weighing designs are considered here. It is not possible to determine a regular A-optimal biased spring balance weighing design in any class $\mathbf{X} \in \mathbf{\Phi}_{n\times p}(0,1)$. Therefore, in the literature new construction methods of A-optimal designs have been presented. In the most cases the construction of such design is based on the incidence matrices of some known block designs. It is worth emphasizing that in the regular A-optimal biased spring balance weighing design we are able to determine unknown measurements of the object with a minimal average variance. From the viewpoint of the experimenters such property is expectable.

It is clear that in the case presented in the example, the experimental design 2^4 may be used. It should be underlined that, for $\mathbf{G} = \mathbf{I}_n$, the sum of variances of estimators of the vector of unknown parameters in both designs is the same. When $\mathbf{G}$ is any positive definite diagonal matrix, the sum of variances of estimators of the vector of unknown parameters in the regular A-optimal spring balance weighing design is less than the sum of variances of estimators of the vector of unknown parameters in the design 2^4.

REFERENCES

BANERJEE, K. S., (1975). Weighing Designs for Chemistry, Medicine, Economics, Operations Research, Statistics. Marcel Dekker Inc., New York.

CERANKA, B., KATULSKA, K., (1990). Constructions of optimum biased spring balance weighing designs with diagonal covariance matrix of errors. Computational Statistics and Data Analysis 10, pp. 121–131.

CERANKA, B., KATULSKA K., (1992). Optimum biased spring balance weighing designs with non-homogeneity of the variances of errors. Journal of Statistical Planning and Inference 30, pp. 185–193.

GRACZYK, M., (2011). A-optimal biased spring balance weighing design. Kybernetika 47, pp. 893–901.

KATULSKA, K., (1989). Optimum biased spring balance weighing designs. Statistics and Probability Letters 8, pp. 267–271.

PUKELSHEIM, F., (1993). Optimal design of experiment. John Wiley and Sons. New York.

RAGHAVARAO, D., (1971). Constructions and combinatorial problems in design of experiment. John Wiley and Sons. New York.

RAGHAVARAO, D., PADGETT, L. V., (2005). Block Designs, Analysis, Combinatorics and Applications. Series of Applied Mathematics 17, Word Scientific Publishing Co. Pte. Ltd., Singapore.

STATISTICS IN TRANSITION new series, Spring 2014
Vol. 15, No. 2, pp. 327—330

REPORT

Conference on Coherence Policy and the Development of Cross-border Areas Along the European Union's External Border, 27-28 June 2014, Krasiczyn-Arlamow, Poland

The conference ***Coherence Policy and the Development of Cross-border Areas Along the European Union's External Border*** (***Polityka spójności a rozwój obszarów transgranicznych na zewnętrznej granicy Unii Europejskiej)*** was held in Krasiczyn and Arlamow from 27th to 28th June 2014. The Conference was organized by the Central Statistical Office of Poland and the Statistical Office in Rzeszow, under the honorary patronage of: Marshal of Podkarpackie Voivodship – **Wladyslaw Ortyl**, Rector of the University of Rzeszow – **Prof. Aleksander Bobko**, Director of the Regional Research Institute of the Ukrainian Academy of Sciences – **Prof. Vasyl Kravtsiv** and Rector of the School of Economics and Management of Public Administration in Bratislava – **Prof. Viera Cibáková**.

Members of the Conference Scientific Council

- Prof. Janusz Witkowski – Chairman, Central Statistical Office of Poland.
- Dr. Marek Cierpial-Wolan – Statistical Office in Rzeszow.
- Dr Jana Fiserova – University of Staffordshire, Huddersfield, Great Britain.
- Prof. Vladimir Gozora – Higher School of Economics and Management in Public Administration, Slovakia.
- Prof. Monika Hudakova – Higher School of Economics and Management in Public Administration, Slovakia.
- Dr. Krzysztof Jakobik – Statistical Office in Krakow.
- Prof. Vasyl Kravtsiv – Institute of Regional Research, Ukrainian Academy of Science.
- Prof. Franciszek Kubiczek – Council of Statistics.
- Dr. Krzysztof Markowski – Statistical Office in Lublin.
- Prof. Semen Matkovskyy – State Statistics Service of Ukraine.
- Prof. Janusz Merski – ALMAMER Higher School, Warsaw.
- Dr. Marek Mroczkowski – Central Statistical Office of Poland.
- Prof. Wlodzimierz Okrasa – Central Statistical Office of Poland.
- Prof. Oleksandr Osaulenko – State Statistics Service of Ukraine.
- Dr. Tomasz Potocki – University of Rzeszow.
- Prof. Iva Ritschelová – The Czech Statistical Office.
- Prof. Vasily Simchera – Russian Academy of Economic Science.

- Prof. Grzegorz Slusarz – University of Rzeszow.
- Dr. Bogdan Wierzbinski – University of Rzeszow.
- Prof. Stanislaw Zieba – ALMAMER Higher School, Warsaw.

The Conference was an interdisciplinary venture. Representatives of official statistics and the scientific institutions of Poland, Ukraine and Slovakia, as well as representatives of political and administrative authorities involved in the regional development, attended the conference (a total of about 70 people). Topics discussed covered the achievements and experience in research and monitoring of the processes of socio-economic development of cross-border areas, along the role of official statistics in the formulation and implementation of territorially targeted cohesion policy.

The Conference was opened by the President of the Central Statistical Office of Poland **Prof. Janusz Witkowski,** followed by a greeting address of **Prof. Jozefa Hrynkiewicz**, Member of the Parliament (Sejm) of the Republic of Poland.

The plenary session was devoted to ***Monitoring socio-economic processes in the border areas in the context of the objectives of cohesion policy***. It was led by **Prof. Grzegorz Slusarz**, the Dean of the Faculty of Economics of the University of Rzeszow. The opening lecture was delivered by **Prof. Janusz Witkowski** on ***Transborder statistics in monitoring development***, stressing, among others, information needs for border areas as a new type of economic space, pointing to challenges to the statistics of the border areas and the need for international cooperation of statisticians. Then **Prof. Oleksandr Osaulenko**, President of the State Statistics Service of Ukraine, presented the topic ***Improvement to statistical support for monitoring of cross-border cooperation***, pointing, in particular, to the regional statistics as a source of information for monitoring the effectiveness of cross-border cooperation, the challenges of Ukraine's cooperation with the EU in the field of regional statistics, or the need for improvement of statistical performance cross-border cooperation in order to create a coherent system of cross-border statistics. Next, **Mr. Juraj Horkay**, Vice President of the Statistical Office of the Slovak Republic, presented ***Draft cross-border cooperation***, funded by the Norwegian Financial Mechanism, which aims to create a statistical system for the border regions of Slovakia and Ukraine.

The plenary session was followed by two parallel panel sessions. First session was dedicated to ***Information needs of the coherence policy in the regional approach***. It was chaired by **Prof. Osaulenko**. Second panel session, led by the President of the Statistical Council, **Prof. Franciszek Kubiczek**, was devoted to ***Spatial differentiation of development level alongside the European Union's external border***. On the second day of the Conference the third panel session took place, which was led by **Prof. Witkowski.** The session was devoted to ***The role of statistics in monitoring the socio-economic development of cross-border areas***.

Altogether, 22 papers were delivered during the plenary session and three panel sessions. The first panel session was opened by **Dominika Rogalinska's** (CSO) presentation ***Official statistics in monitoring the territorial dimension of coherence policy***, noting, among other, new challenges facing the official statistics, the STRATEG portal, project "Support to the monitoring system of cohesion policy" and the role of the CSO in monitoring the cohesion policy in financial perspective in the 2014-2020. Then **Prof. Wlodzimierz Okrasa**, Advisor to the President of the CSO, presented a paper on ***Statistical aspects of coherence: Towards a spatial integration of local development indicators***. He pointed out, among others, information needs of coherence policy in the regional approach, the local dimension of cohesion policy and recognition of multifaceted character of local development. He also presented the conceptualisation of local development and well-being, research on spatial inequalities of local and individual welfare as well as the impact of the allocation of development funds to reduce local deprivation at the NUTS5 level. **Prof. Stanislaw Zieba** (Almamer University) gave a presentation on ***Macroeconomic situation in Poland in the context of coherence policy for the period 2004-2015***. He focused mainly on the results of evaluation studies of the impact of EU cohesion funds for the period 2004-2015, including the study of selected macroeconomic indicators as factors affecting the economic and social cohesion of the country. Then **Prof. Semen Matkovskyi**, Advisor to the President of the State Statistics Service of Ukraine, presented a paper ***Development of regional statistics in the context of the Ukrainian-Polish cross-border cooperation***. The first session was concluded by papers on special aspects of cross-border cooperation: ***Impact of financial and economic crisis on the economy of border regions*** by **Prof. Larisa Yaremko** and **Yuliia Poliakova, PhD** (Lviv Academy of Commerce), ***Monitoring of business processes in the Euroregion Tatry area using the data of official statistics of Poland and Slovakia*** by **Krzysztof Jakobik, PhD** (Statistical Office in Krakow) and ***The role of cross-border cooperation in regional development*** by **Ivan Ustich, PhD** (Foundation for the Development of Cross-Border Cooperation and Special Economic Zones, Uzhhorod).

During the second session **Prof. Romuald Polinski** (Almamer University) and **Marian Szolucha, PhD** (Vistula University) presented the paper ***Economic development in Poland and Slovakia and the financial crisis in the years 2008-2013***. Another paper presented by **Prof. Vladimir Gozora** (VSEMVS in Bratislava), entitled ***Economic and social diversity of Polish and Slovakian border areas in specific breakdowns***, presented the results of research of small and medium-sized enterprises in the regions of Nitra, Zilina and Presov, and ideas for the development of trans-border cooperation. During the second session the following papers were also presented: ***The role of social development indicators in ensuring coherence between border regions of Slovakia and Poland*** by **Prof. Monika Hudáková** (VSEMVS in Bratislava), ***Development of entrepreneurship in border areas of the Euroregion Bug in the years 2000-2013*** by **Krzysztof**

Markowski, PhD (Statistical Office in Lublin), ***Respect the border*** by **Marek Morze** (Statistical Office in Olsztyn) and ***Monitoring of threats to coherence in the Polish-Slovakian border area during the European financial crisis in the years 2008-2013*** by **Ewa Gucwa-Lesny, PhD** (Almamer University).

On the second day of the Conference (during the third session) the opening lecture on ***Statistics in stimulating the development of cross-border regions*** was delivered by **Prof. Grzegorz Slusarz** (University of Rzeszow). He drew attention to a need to recognize the complexity of the development processes in the context of the use of statistics (broadly defined) as a tool to stimulate the development of territorial units with special reference to border areas. Then **Prof. Vasyl Kravtsiv** (Regional Research Institute of the Ukrainian Academy of Sciences) presented a paper entitled ***Problems of information-statistical support for the development of the model of administrative-territorial reform of Lviv region as a border region***. **Prof. Franciszek Kubiczek** (Statistical Council) in his speech entitled ***The role of the Statistical Council in monitoring the socio-economic processes and phenomena observed in cross-border areas***, stressed the importance of the work carried out by official statistics for the study of border areas. He also presented two interesting initiatives on Human Developmet Index in local terms and Doing Business. Then **Ján Cuper** (Statistical Office in Prešov) gave a presentation on ***Tourism in Slovakia and in Prešov region*** and **Roman Fedak** (Statistical Office in Zielona Gora) shared practical insights in the presentation ***The role of regional statistics in the process of meeting the information needs of coherence policy. The experience of SO Zielona Gora***.

At the end of the third session, **Maria Jeznach** (CSO), **Jozef Sobota** (NBP) and **Marek Cierpial-Wolan, PhD** (Statistical Office in Rzeszow) together presented the subject of ***Multi-method surveys of travel for the needs of tourism statistics, national accounts and balance of payments – the use of multiple research methods***. M. Cierpiał-Wolan, PhD, pointed out, among other, the effects of integration and disintegration of economic processes, presented a coherent research system for cross-border areas, which due to the nature of the studied phenomena requires the use of different research methods. In turn, M. Jeznach stressed, among other, the importance of using these research results in the statistics of national accounts and balance of payments and the transnational and multidimensional nature of cross-border processes, which results in the need for international cooperation to carry out the research. J. Sobota drew attention to the essential role of statistics of cross-border areas in the work of the Statistics Department of the NBP. In summary of the last presentation Prof. J. Witkowski stressed the importance of research conducted jointly by the various institutions, in this case, CSO, NBP and MSiT, which provide a wide range of methodologically consistent data and to satisfy different needs.

At the end of the conference the President of the Central Statistical Office handed the honorary badges ***For services for Statistics of the RP*** to people who work for years with the Polish Statistics: **Mr. Juraj Horkay**, Vice President of the Statistical Office of the Slovak Republic, **Prof. Vasyl Kravtsiv**, Director of the Regional Research Institute of the Ukrainian Academy of Sciences, **Mrs. Svitlana Zymovina**, Director of the Main Statistical Office in Lviv Region.

The conference was concluded by **Prof. Janusz Witkowski** as very effective. He also emphasized the role and the importance of such meetings. The conference provided a platform for presentation of the current state of knowledge and innovative research approaches, and allowed to identify directions for further research.

Prepared by:

Marek Cierpial-Wolan
Elzbieta Wojnar
Statistical Office in Rzeszow

ERRATUM

The proper title of the paper by Dorota Pekasiewicz published in the previous issue of the *Statistics in Transition new series* (Volume 15, No 1) should read as follows: “Application of quantile methods to estimation of Cauchy distribution parameters”. Also, appearing in the text the term ‘quintile’ should be replaced by ‘quantile’.

We apologize to the Author and to the readers for this error.