STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Sampling methods and estimation

Research articles

Other articles:

Multivariate Statistical Analysis 2014, Łódź. Conference Papers

Volume 16, Number 1, Spring 2015

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 2015
Vol. 16, No. 1, pp. 1—3

FROM THE EDITOR

A set of eight articles that constitute this issue is organized, as usual, in three-part arrangement: *Sampling Methods and Estimation* (three articles), *Research Articles* (two articles), and *Other Articles* (three articles). They are produced by sixteen authors from different countries and continents. This gives a sense of the broadness of the journal in terms of its topical and geographical scope.

It starts with an article by **Elkasabi M. A.**, **Heeringa S. G.**, and **Lepkowski J. M.**, entitled ***Joint Calibration Estimator for Dual Frame Surveys*** in which a new Joint Calibration Estimator (JCE) is proposed. According to the authors, it has better performance when the auxiliary variables can fully explain the variability in the study variables or at least when the auxiliary variables are strong correlates of the estimation variables. It is asymptotically design unbiased conditional on the strong relationship between the estimation variable and the auxiliary variables employed in the calibration. In contrast to the standard dual frame estimators, the JCE does not require domain membership information and the effect of the randomly misclassified domains does not exceed the random measurement error effect. Therefore, the JCE tends to be robust for the misclassified domains if included in the auxiliary variables.

Swain A. K. P. C. and **Das M.** present ***Some Classes of Modified Ratio Type Estimators in Sample Surveys***, with additive and multiplicative adjustments made to the simple mean per unit estimator. Also, classical ratio estimators are suggested to obtain more efficient ratio type estimators compared to the classical one. Their biases and mean square errors are obtained and compared with first order approximations. Specifically, it is shown that even without assuming restrictive assumptions associated with the linear regression estimator, the proposed modified ratio-type estimators are asymptotically as efficient as the linear regression estimators.

In the next article, ***Improved Separate Ratio and Product Exponential Type Estimators in the Case of Post-Stratification,*** **Hilal A. Lone** and **Rajesh Tailor** address the problem of estimation of finite population mean in the case of post-stratification. Improved separate ratio and product exponential type estimators in the case of post-stratification are suggested. The biases and mean squared errors of the suggested estimators are obtained up to the first degree of approximation. Theoretical and empirical studies have been done to demonstrate better efficiencies of the suggested estimators than other considered estimators. In particular, the suggested estimators are recommended for use in practice for estimating the population mean when some conditions (discussed in the body of the text) are satisfied.

The *Research Articles* section opens **Nicholas T. Longford's** paper ***Policy-Oriented Inference and the Analyst-Client Cooperation. An Example from Small-Area Statistics*** which demonstrates that efficient estimation is not always conducive to good policy decisions because the established inferential procedures have no capacity to incorporate the priorities and preferences of the policy makers and the related consequences of incorrect decisions. A method that addresses these deficiencies is described along with an example of planning an intervention in a developing country's districts with high rate of illiteracy. In addition to exposing the deficiencies of the general concept of efficiency the example shows that the criterion for the quality of an estimator has to be formulated specifically for the problem at hand. In particular, it is shown that the established small-area estimators may perform poorly if the minimum mean squared error is an inappropriate criterion.

The problem of forecasting of prices of commodities is discussed by **Abhishek Singh** and **G. C. Mishra** in ***Application of Box-Jenkins Method and Artificial Neural Network Procedure for Time Series Forecasting of Prices.*** Focusing on prices of agricultural commodities which are considered especially difficult to forecast because they are not only governed by demand and supply but also by so many other factors which are beyond control (such as weather vagaries, storage capacity, transportation, etc.) the Authors employ to this aim time series models ARIMA (Autoregressive Integrated Moving Average) methodology given by Box and Jenkins. An example is provided for the case of forecasting prices of Groundnut oil in Mumbai. This approach has been compared with ANN (Artificial Neural Network) methodology. The results showed that ANN performed better than the ARIMA models in forecasting the prices. The reason for this may lay in the nature of the data which show chaotic behaviour and cannot be fully captured by the linear ARIMA model. Also, the neural network results conform to the theoretical proofs that a feed forward neural network with only one hidden layer can precisely and satisfactorily approximate any continuous function.

The last section contains three papers based on presentations given at the Multivariate Statistical Analysis conference held in Łódź, Autumn 2014. In the first paper **Tomasz Górecki**, **Mirosław Krzyśko** and **Waldemar Wołyński** discuss ***Classification Problems Based on Regression Models for Multi-Dimensional Functional Data.*** Data in the form of a continuous vector function on a given interval are referred to as multivariate functional data. These data are treated as realizations of multivariate random processes. The Authors use multivariate functional regression techniques for the classification of multivariate functional data along with illustration of application of the discussed approaches to two real data sets.

In **Radosław Pietrzyk's** and **Paweł Rokita's** paper ***Stochastic Goals in Financial Planning for a Two-Person Household*** addressed are two types of risk that are typically being taken into account in household financial planning. The first is life-long lasting risk and the other is financing related risk. Also variety of events associated with insurance are sometimes taken into account, e.g., health deterioration. However, there is no model addressing stochastic nature of household financial goals, and the problem of modelling goals themselves is not sufficiently explored yet. This article puts forward a proposition of working goal time and magnitude into a household financial plan accounting for their distributions in optimizing the plan. A two-person household model is used while the decision variables of the optimization task are consumption-investment proportion and intra-household investments, respectively.

In the last paper, ***Robust Regression in Monthly Business Survey***, **Grażyna Dehnel** discusses the distorting effect that outliers (extreme values) can have on classical statistical methods that are optimal under the assumption of normality or linearity, with special reference to population surveys in areas of business, agricultural, household and medicine. In particular, the presence of extreme observations may adversely affect estimation, especially when it is carried out at a low level of aggregation. To deal with this problem, several alternative techniques of estimation which seem to be less sensitive to outliers have been proposed in the statistical literature. Author attempts to apply and assess some robust regression methods (*LTS, M-estimation, S-estimation, MM-estimation*) in the business survey conducted within the framework of official statistics.

The issue is complemented by a brief note about the authors.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 5

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

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

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

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

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 7–36

JOINT CALIBRATION ESTIMATOR FOR DUAL FRAME SURVEYS

Mahmoud A. Elkasabi[1], **Steven G. Heeringa**[2], **James M. Lepkowski**[3]

ABSTRACT

Many dual frame estimators have been proposed in the statistics literature. Some of these estimators are theoretically optimal but hard to apply in practice, whereas others are applicable but have larger variances than the first group. In this paper, a Joint Calibration Estimator (JCE) is proposed that is simple to apply in practice and meets many desirable properties for dual frame estimators. The JCE is asymptotically design unbiased conditional on the strong relationship between the estimation variable and the auxiliary variables employed in the calibration. The JCE achieves better performance when the auxiliary variables can fully explain the variability in the study variables or at least when the auxiliary variables are strong correlates of the estimation variables. As opposed to the standard dual frame estimators, the JCE does not require domain membership information. Even if included in the JCE auxiliary variables, the effect of the randomly misclassified domains does not exceed the random measurement error effect. Therefore, the JCE tends to be robust for the misclassified domains if included in the auxiliary variables. Meanwhile, the misclassified domains can significantly affect the unbiasedness of the standard dual frame estimators as proved theoretically and empirically in this paper.

Key words: dual-frame estimation, calibration weighting, auxiliary variables, domain misclassification.

1. Introduction

With rapid changes in the cost of survey data collection, changes in population coverage patterns, and sample unit accessibility, dual frame sample surveys are becoming more common in survey practice. For example, dual frame telephone surveys that combine RDD landline telephone samples and cell phone samples emerged to reduce noncoverage due to “cell-only” households in Random-Digit-Dialing (RDD) landline telephone surveys (Brick et al., 2007;

[1] ICF International, Maryland, USA. E-mail: mahmoud.elkasabi@icfi.com.
[2] Institute for Social Research, University of Michigan, USA. E-mail: sheering@umich.edu.
[3] Institute for Social Research, University of Michigan, USA. E-mail: jimlep@umich.edu.

Link, Battaglia, Frankel, Osborn, & Mokdad, 2007). At the same time, Address Based Sampling (ABS) has been explored as a complement or an alternative to RDD telephone surveys (Link, Battaglia, Frankel, Osborn, & Mokdad, 2006, 2008; Link & Lai, 2011).

Estimation is not straightforward in dual frame surveys due to the overlap between the two frames. Simply adding the two estimated totals of the samples results in a biased estimate of the overall total. Standard dual frame estimators adjust for the overlap but present many methodological and practical problems in implementation (Lohr, 2011). In addition, standard dual frame estimation requires the correct identification of the design domain for each sample element. An error in the determination of design domain membership can affect the efficiency of the estimates (Lohr, 2011; Mecatti, 2007).

In this paper, the Joint Calibration Estimator (JCE) is introduced as a new dual frame estimator that relies on the general calibration approach introduced by Deville and Särndal (1992). Calibration generates unbiased estimates themselves for the auxiliary calibration variables under dual frame designs. The effectiveness of calibration for estimates for other variables not included in the calibration set is not completely understood in the dual frame context.

In this paper, we provide an overview of dual frame estimation and introduce a model-assisted design-based JCE under the 'ideal situation', with no errors present in the determination of sample domain and only sampling error for the estimate itself, and in the presence of domain misclassification, where dual frame domains are not correctly identified. The dual frame estimation and calibration approaches are discussed in Sections 2 and 3. The JCE is introduced in Sections 4 and 5, while in Section 6, the bias and variance estimate for the JCE are presented. The misclassification bias for the standard dual frame estimators is derived in Section 7. A simulation study of the performance of the JCE in comparison with standard dual frame estimators is described in Section 8, and the results are discussed in Section 9.

2. Dual frame estimation

Lohr (2011) identified the following five desirable properties for dual frame estimators: (1) unbiased for the corresponding finite population quantity; (2) internally consistent (that is, the multivariate relationships in the data should be preserved, such as the sum of the estimated totals for subgroups should equal the estimated overall); (3) efficient, with low Mean Square Error (MSE); (4) calculable with standard survey software (e.g., one set of weights is needed for all study variables; replicate weights are available for formula-based or replication-based variance estimation); and (5) robust to non-sampling errors.

In addition to Lohr's properties, we add the following three properties. (6) Data requirements for the estimator should be reasonable. For example, information about design domain membership or variance and covariance components is required for some estimators, but these may be poorly measured or unreliable components and add to the burden and complexity of computing the estimator. (7) An estimator should be robust to non-sampling errors in the estimator's auxiliary and domain membership variables or required variances and covariances. Although some estimators might theoretically be efficient, reporting errors in domain membership or biased estimates of required variance and covariance components could result in biased or non-optimal estimators. (8) An estimator should be readily applicable to dual and multiple (more than two) frame surveys.

2.1. Notation

Let $U=\{1,..,k,..,N\}$ denote a target population of *N* elements, and let $A=\{1,..,k,..,N_A\}$ and $B=\{1,..,k,..,N_B\}$ denote two overlapping frames. The two frames are not assumed to be exclusive, that is: $A \bigcap B = ab \neq \phi$ and $A \bigcup B = U$. The dual frame design sample *s* is composed of two samples $s_A\left(s_A \subseteq A\right)$ and $s_B\left(s_B \subseteq B\right)$ selected from the two overlapping frames *A* and *B* using a sample design with inclusion probabilities $\pi_k^A = p\left(k \in s_A\right)$ and $\pi_k^B = p\left(k \in s_B\right)$. Base weights to compensate for unequal selection probabilities are d_k, where $d_k = d_k^A = 1/\pi_k^A$ for $k \in s_A$ and $d_k = d_k^B = 1/\pi_k^B$ for $k \in s_B$. Let N_A and N_B denote the frame sizes and n_A and n_B denote the sample sizes for frames *A* and *B*, respectively. Let $a = A \cap B^c$ and $b = A^c \cap B$, where *c* denotes the complement of a set, and $s_a = a \cap s_A$, $s_b = b \cap s_B$, $s_{ab}^A = ab \cap s_A$ and $s_{ab}^B = ab \cap s_B$. Standard dual frame estimators of a population total take the form $\hat{Y} = \hat{Y}_a + \hat{Y}_{ab} + \hat{Y}_b$ estimating the true population total $Y = Y_a + Y_{ab} + Y_b$, where $Y = \sum_{k \in U} y_k$,$Y_a = \sum_{k \in a} y_k$, $Y_b = \sum_{k \in b} y_k$ and $Y_{ab} = \sum_{k \in ab} y_k$.

2.2. The standard dual frame estimators

The Horvitz-Thompson estimators of totals (Horvitz & Thompson, 1952) for domains *a* and *b* for characteristic *Y* are $\hat{Y}_a = \sum_{k \in s_a} d_k y_k$ and $\hat{Y}_b = \sum_{k \in s_b} d_k y_k$, and the estimators for the domain overlap are $\hat{Y}_{ab}^A = \sum_{k \in s_{ab}^A} d_k y_k$ and $\hat{Y}_{ab}^B = \sum_{k \in s_{ab}^B} d_k y_k$. For each sample, the estimators of population totals are unbiased for the corresponding domain totals Y_a, Y_{ab} and Y_b: $E\left[\hat{Y}_a + \hat{Y}_{ab}^A\right] = Y_a + Y_{ab}$ and $E\left[\hat{Y}_b + \hat{Y}_{ab}^B\right] = Y_b + Y_{ab}$, where $E(.)$ denotes design-based expectation. Therefore, adding the two sample estimated totals results in a biased population estimate $E\left[\hat{Y}_a+\hat{Y}_{ab}^A+\hat{Y}_b+\hat{Y}_{ab}^B\right]\approx Y_a+2Y_{ab}+Y_b \neq Y$.

An unbiased dual frame estimator for Y can be obtained by the weighted average of the estimators $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$,

$$\hat{Y}=\hat{Y}_a+\theta\hat{Y}_{ab}^A+(1-\theta)\hat{Y}_{ab}^B+\hat{Y}_b \tag{1}$$

where $\theta\in[0,1]$ is a composite factor combining $\hat{Y}_{ab}^A$ and $\hat{Y}_{ab}^B$. Estimators of domain sizes $\hat{N}_a^A$, $\hat{N}_{ab}^A$, $\hat{N}_{ab}^B$ and $\hat{N}_b^B$ are defined by setting $y_k=1$ for all $k\in s$ in $\hat{Y}_a^A$, $\hat{Y}_{ab}^A$, $\hat{Y}_{ab}^B$ and $\hat{Y}_b^B$, and the dual frame estimator in (1) can be used to find the population total estimate $\hat{N}$. Consequently, an unbiased dual frame estimator for the population mean $\bar{Y}$ can be written as $\bar{Y}=\hat{Y}/\hat{N}$. The weighted version of the estimated total in (1) can be written as

$$\hat{Y}=\sum_{k\in s_A} m_k d_k y_k+\sum_{k\in s_B} m_k d_k y_k \tag{2}$$

where the adjustment factor m_k can be written as

$$m_k=\begin{cases} 1 & k\in s_a, \\ \theta & k\in s_{ab}^A, \\ 1-\theta & k\in s_{ab}^B, \\ 1 & k\in s_b. \end{cases} \tag{3}$$

The approach used to determine the composite factor θ distinguishes standard dual frame estimators. Hartley (1962, 1974) proposed choosing the composite factor θ_{HT} to minimize the variance of $\hat{Y}$. Choosing any fixed value for the composite factor (e.g. θ=0.5) yields the unbiased Fixed Weight Estimator (FWE), which includes the optimum Hartley Estimator (HE) as a special case.

Fuller and Burmeister (1972) extended Hartley's estimator by using a maximum likelihood estimator $\hat{N}_{ab}$ of the overlap domain population size N_{ab}. Later, Skinner and Rao (1996) extended the Fuller-Burmeister (FB) estimator to achieve design-based consistency under complex designs using a Pseudo-Maximum Likelihood Estimator (PML). Rao and Wu (2010) proposed the Pseudo-Empirical Likelihood (PEL) estimator, which depends on adjustment factors based on probability measures p_a , p_{ab}^A, p_b and p_{ab}^B for a randomly selected case being in poststrata s_a, s_{ab}^A, s_b and s_{ab}^B.

Several single frame estimators have been proposed as alternatives. Bankier (1986) and Kalton and Anderson (1986) proposed the Single Frame Estimator (SFE) which treats the dual frame design as a stratified design consisting of three strata, one for each design domain, and calculates joint inclusion probabilities. Meccati (2007) introduced a simple dual frame estimator, the Multiplicity Estimator (ME), which depends on the number of the frames that case k belongs to, M_k , in order to combine domains.

With respect to their sampling variance, consistency, and practical utility, these estimators can be grouped into three types. First are the optimal estimators, HE, FB, and PEL. These are internally inconsistent since they generate weights that are dependent on the study variables. This restricts the practical application of the optimal estimators using standard survey software. At the same time, these optimal estimators require estimates of variance and covariance components for finding the composite factor θ . Biased estimates of the required components result in non-optimal estimates. Forms of these estimators for multiple frame surveys are complicated due to the need to estimate covariance terms in the composite factors (Lohr & Rao, 2000, 2006; Mecatti, 2007; Skinner, 1991).

The second type is the "practical" estimators, FWE, SFE and ME. Easier to compute in practice, these estimators achieve notably poorer efficiency relative to the optimal estimators. They are internally consistent since they generate only one set of weights for all study variables and standard survey software can be used to find the survey estimates. Deriving these estimators for multiple frame surveys is a straightforward task.

The third type includes just the PML, which has greater practical applicability than the optimal estimators and is more efficient than the practical estimators. PML has smaller MSE than FB and HE because it does not require estimation of variance components of the composite factors in FB and HE (Lohr & Rao, 2000; Skinner & Rao, 1996).

With respect to the eight desirable properties for dual frame estimators, all the standard dual frame estimators are unbiased, or approximately so. Not all of them are internally consistent, efficient, or calculatable with standard survey software. With regard to property (5) concerning non-sampling errors, dual frame estimators have a disadvantage compared to single frame surveys because of different levels of non-sampling errors associated with the frames (Brick, Flores-Cervantes, Lee, & Norman, 2011). These kinds of associations add to the complexity of the assessment and adjustment for these errors, adversely affecting property 6.

Nearly all of these dual frame estimators require accurate information about domain membership. But domain membership might be affected by reporting errors and leading to a biased estimate (property (7)). Finally, extending standard dual frame estimators to multiple frames is not readily achieved (property (8)).

3. The calibration approach

In the single frame survey design, where the sample $s\left(s \subseteq U\right)$ is selected from the population U using a sample design with inclusion probability of $\pi_k = p\left(k \in s\right)$, the base weights are equal to $d_k = 1/\pi_k$. Let $\mathbf{x}_k = \left(x_{k1},..,x_{kj},..,x_{kJ}\right)'$ denote an auxiliary variable vector of dimension $j = \left(1,...,J\right)$, where both y_k and $\mathbf{x}_k$ are observed for the sample elements $k \in s$. The Horvitz-Thompson estimator for the total $Y = \sum_{k\in U} y_k$ is $\hat{Y}_{HT} = \sum_{k\in s} d_k y_k$.

In a complete response situation, with known auxiliary totals for the $j = \left(1,..,J\right)$ auxiliary variables, $\mathbf{X} = \left(X_1,..,X_j,..,X_J\right)' = \left(\sum_{k\in U} x_{k1},..,\sum_{k\in U} x_{kj},..,\sum_{k\in U} x_{kJ}\right)'$, Deville and Särndal (1992) defined calibration as a method to find weights w_k which minimize a distance measure $G\left(w_k, d_k\right)$ between the calibrated weights w_k and the base weights d_k. This minimization of the distance function is subject to the constraint that the calibration-weighted total of the auxiliary variable values $\sum_{k\in s} w_k x_{kj}$ equals the known population total for the auxiliary X_j for $j = 1,\ldots,J$, or $\sum_{k\in s} w_k \mathbf{x}_k = \mathbf{X}$. This calibration approach results in final calibrated weights $w_k = d_k F\left(q_k \mathbf{x}'_k \lambda\right)$ where $F\left(q_k \mathbf{x}'_k \lambda\right)$ is the inverse of $\partial G\left(w_k, d_k\right)/\partial w_k$, λ denotes a vector of Lagrange multipliers in the minimization, and q_k is a positive value which scales the calibrated weights.

Many distance measures have been proposed for calibration, but empirically there are small differences in the calibrated estimates derived from alternative distance measures (Singh & Mohl, 1996; Stukel, Hidiroglou, & Särndal, 1996). We use the linear case with the chi-square distance function $\left(w_k - d_k\right)^2/2d_k$ and $q_k = 1$. The calibration obtains $w_k, k \in s$ by minimizing the distance function $\sum_{k\in s}\left(w_k - d_k^*\right)^2/2d_k^*$ subject to the calibration equation $\sum_{k\in s} w_k \mathbf{x}_k = \mathbf{X}$, where d_k^* are arbitrary initial weights (a base weight or an adjusted version).
The minimization generates the Lagrange multiplier vector $\lambda' = \left(\sum_{k\in U} \mathbf{x}_k - \sum_{k\in s} d_k^* \mathbf{x}_k\right)'\left(\sum_{k\in s} d_k^* \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$ and calibration factor is $g_k = \left(1 + \lambda' \mathbf{x}_k\right)$.
The final calibrated weights are $w_k = d_k^*\left(1 + \lambda' \mathbf{x}_k\right) = d_k^*\left[1 + \left(\sum_{k\in U} \mathbf{x}_k - \sum_{k\in s} d_k^* \mathbf{x}_k\right)\left(\sum_{k\in s} d_k^* \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \mathbf{x}_k\right]$ and the calibrated estimated total is $\hat{Y}_w = \sum_{k\in s} w_k y_k$.

As it will be shown in the next section, the main idea behind calibration, finding a set of weights which guarantee that estimated auxiliary totals conform to known population totals, can be used to combine two samples.

4. Joint Calibration Estimator

Under the dual frame design, let $E\left(\sum_{k\in s_A} d_k\mathbf{x}_k\right)=\mathbf{X}_A$, $E\left(\sum_{k\in s_B} d_k\mathbf{x}_k\right)=\mathbf{X}_B$ and $E\left(\sum_{k\in s_A} d_k\mathbf{x}_k+\sum_{k\in s_B} d_k\mathbf{x}_k\right)\neq\mathbf{X}$, where $\mathbf{X}_A=\left(\sum_{k\in A}x_{k1},..,\sum_{k\in A}x_{kj},..,\sum_{k\in A}x_{kJ}\right)'$ and $\mathbf{X}_B=\left(\sum_{k\in B}x_{k1},..,\sum_{k\in B}x_{kj},..,\sum_{k\in B}x_{kJ}\right)'$. Calibration conditioning on $\sum_{k\in s_A} w_k\mathbf{x}_k+\sum_{k\in s_B} w_k\mathbf{x}_k=\mathbf{X}$ should achieve $E\left(\sum_{k\in s_A} w_k\mathbf{x}_k+\sum_{k\in s_B} w_k\mathbf{x}_k\right)=\mathbf{X}$. Consequently, a set of auxiliary variables that are strong predictors for the study variable y should yield $E(\sum_{k\in s_A} w_k y_k+\sum_{k\in s_B} w_k y_k)\simeq Y$ (see Proposition 1 and Corollary 1).

Under complete response (i.e., no nonresponse), calibrated estimates can be parameterized for the dual frame design by deriving the calibration factors as explicit components for each sample of the dual frame sample. Calibration finds final weights w_k such that

$$\sum_{k\in s} w_k\mathbf{x}_k=\sum_{k\in s_A} w_k\mathbf{x}_k+\sum_{k\in s_B} w_k\mathbf{x}_k=\mathbf{X} \tag{4}$$

by minimizing the distance function $\sum_{k\in s_A}\left(w_k-d_k\right)^2/2d_k+\sum_{k\in s_B}\left(w_k-d_k\right)^2/2d_k$. The joint calibration weights are $w_k=d_k\left(1+\lambda'\mathbf{x}_k\right), k\in s$ where $\lambda'=\left(\sum_{k\in U}\mathbf{x}_k-\sum_{k\in s}d_k\mathbf{x}_k\right)'\left(\sum_{k\in s}d_k\mathbf{x}_k\mathbf{x}'_k\right)^{-1}$ with joint calibration factor $g_k=\left(1+\lambda'\mathbf{x}_k\right)$.

Therefore, the JCE for population total can be written as

$$\hat{Y}_{JCE}=\sum_{k\in s} w_k y_k=\sum_{k\in s_A} w_k y_k+\sum_{k\in s_B} w_k y_k \tag{5}$$

where $w_k=d_k\left(1+\lambda'\mathbf{x}_k\right)$ and

$$\lambda'=\left(\sum_{k\in U}\mathbf{x}_k-\sum_{k\in s_A}d_k\mathbf{x}_k-\sum_{k\in s_B}d_k\mathbf{x}_k\right)'\left(\sum_{k\in s_A}d_k\mathbf{x}_k\mathbf{x}'_k+\sum_{k\in s_B}d_k\mathbf{x}_k\mathbf{x}'_k\right)^{-1}.$$

The calibration constraints determine the form of the JCE. Some forms may be identical to the standard dual frame estimators. For example, the dual frame estimator for the total can be written as in equation (1), and the weighted version expressed as in equations (2) and (3), where an alternative expression for equation (2) is $\hat{Y}=\sum_{k\in s_a} d_k y_k+\sum_{k\in s_{ab}^A} m_k d_k y_k+\sum_{k\in s_{ab}^B} m_k d_k y_k+\sum_{k\in s_b} d_k y_k$, where m_k is defined in (3). When the auxiliary variable $\mathbf{x}_k=1$ for $k\in U$, under the JCE, the main constraint in (4) can be written as

$$\sum_s w_k = N \tag{6}$$

and the constraint $w_k=d_k\ \forall\ k\in s_a\cup s_b$ can be added to the calibration minimization problem. This constraint is identical to

$$\sum_{k\in s_a} w_k=\sum_{k\in s_a} d_k^* = N_a \tag{7}$$

and

$$\sum_{k\in s_b} w_k=\sum_{k\in s_b} d_k^* = N_b\,. \tag{8}$$

In (7) and (8), $d_k^*=\left(N_a\big/\sum_{k\in s_a} d_k\right)d_k$ and $d_k^*=\left(N_b\big/\sum_{k\in s_b} d_k\right)d_k$, respectively. Joint calibration with the three constraints (6), (7) and (8) is identical to post-stratifying the sample by the design domain totals N_a, N_{ab} and N_b, which yields the unbiased dual frame estimator (2), where the modification factors for the overlap domain have the same value $m_k=N_{ab}\Big/\left(\sum_{k\in s_{ab}^A} d_k+\sum_{k\in s_{ab}^B} d_k\right)\ \forall k\in s_{ab}^A\cup s_{ab}^B$. In this case, the joint calibration factor is

$$g_k=\begin{cases} N_a\big/\sum_{k\in s_a} d_k & k\in s_a, \\ N_{ab}\Big/\left(\sum_{k\in s_{ab}^A} d_k+\sum_{k\in s_{ab}^B} d_k\right) & k\in s_{ab}^A\cup s_{ab}^B, \\ N_b\big/\sum_{k\in s_b} d_k & k\in s_b. \end{cases} \tag{9}$$

The joint calibration factor in (9) yields the post-stratified version of the Fixed Weight Estimator (FWE), $\hat{Y}_{FWE}^{post}=\frac{N_a}{\hat{N}_a}\hat{Y}_a+\frac{N_{ab}}{\hat{N}_{ab}}\left(\theta\hat{Y}_{ab}^A+(1-\theta)\hat{Y}_{ab}^B\right)+\frac{N_b}{\hat{N}_b}\hat{Y}_b$ where $\theta=0.5$ and $\hat{N}_{ab}=\left(\theta\hat{N}_{ab}^A+(1-\theta)\hat{N}_{ab}^B\right)$.

The JCE can readily be adapted to multiple frames as well. Under multiple frame designs, with P domains, the JCE for population total of y can be written as $\hat{Y}_{JCE}=\sum_{p\in P}\sum_{k\in s_p} w_k y_k$ where $w_k=d_k\left(1+\lambda'\mathbf{x}_k\right)$ and λ' can be written as $\lambda'=\left(\sum_{k\in U}\mathbf{x}_k-\sum_{p\in P}\sum_{k\in s_p} d_k\mathbf{x}_k\right)'\left(\sum_{p\in P}\sum_{k\in s_p} d_k\mathbf{x}_k\mathbf{x}_k'\right)^{-1}$.

5. Examples of Joint Calibration Estimators

The auxiliary variable vector characterizes the final JCE for dual frame estimation. For example, under the univariate auxiliary variable $\mathbf{x}_k = 1$ for $k \in U$, we have the *common mean model*

$$\begin{cases} E_{\xi}(y_k) = \beta, \\ V_{\xi}(y_k) = \sigma^2, \end{cases} \tag{10}$$

where E_{ξ} and V_{ξ} denote the expectation and variance with respect to the calibration model ξ. For the overall population total $\mathbf{X} = N$, the joint calibration factor is $g_k = N\left(\sum_{k\in s_A} d_k + \sum_{k\in s_B} d_k\right)^{-1}$. By calibrating concatenated or "stacked" datasets for each frame's sample, $\sum_{k\in s_A} w_k\mathbf{x}_k + \sum_{k\in s_B} w_k\mathbf{x}_k = N$. This JCE estimate is appropriate when it is thought that the true common mean β is the same for all $k \in U$. However, when the β varies between design domains, another JCE uses the calibration factor in (8).

For $\mathbf{x}_k = x_k$ for $k \in U$, we can also consider the *ratio model*

$$\begin{cases} E_{\xi}(y_k) = \beta x_k, \\ V_{\xi}(y_k) = \sigma^2 x_k, \end{cases} \tag{11}$$

where $\mathbf{X} = X$. The joint calibration factor is $g_k = X\left(\sum_{k\in s_A} d_k x_k + \sum_{k\in s_B} d_k x_k\right)^{-1}$. Calibrating the stacked dataset, $\sum_{k\in s_A} w_k\mathbf{x}_k + \sum_{k\in s_B} w_k\mathbf{x}_k = X$. This JCE estimate is appropriate when it is thought that βx_k is the same, for all $k \in U$. Another JCE estimate is appropriate when it is thought that βx_k varies between design domains. This estimate uses the calibration factor

$$g_k = \begin{cases} X_a \Big/ \sum_{k\in s_a} d_k x_k & k \in s_a, \\ X_{ab} \Big/ \left(\sum_{k\in s_{ab}^A} d_k x_k + \sum_{k\in s_{ab}^B} d_k x_k\right) & k \in s_{ab}^A \cup s_{ab}^B, \\ X_b \Big/ \sum_{k\in s_b} d_k x_k & k \in s_b. \end{cases} \tag{12}$$

Obviously, this estimate requires knowledge of the separate totals (X_a, X_{ab}, X_b).

Under the multivariate auxiliary variable $\mathbf{x}_k = (1, x_k)$ for $k \in U$, consider *the simple regression model with intercept*

$$\begin{cases} E_\xi(y_k) = \alpha + \beta x_k, \\ V_\xi(y_k) = \sigma^2. \end{cases} \tag{13}$$

The calibrated estimate $\hat{Y}_{JCE}$, is $\hat{Y}_{JCE} = \hat{Y}_{HT}^A + \hat{Y}_{HT}^B + \left(\sum_{k\in U} x_k - \left(\sum_{k\in s_A} d_k x_k + \sum_{k\in s_B} d_k x_k\right)\right)\hat{B}_s^{A,B}$ where $\hat{B}_s^{A,B} = \left(\sum_{k\in s_A} d_k \mathbf{x}_k y_k + \sum_{k\in s_B} d_k \mathbf{x}_k y_k\right)\left(\sum_{k\in s_A} d_k \mathbf{x}_k \mathbf{x}_k' + \sum_{k\in s_B} d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}$. With more than one auxiliary variable, the multivariate estimator can be written as

$$\hat{Y}_{JCE} = \hat{Y}_{HT}^A + \hat{Y}_{HT}^B + \left(\sum_{k\in U} \mathbf{x}_k - \left(\sum_{k\in s_A} d_k \mathbf{x}_k + \sum_{k\in s_B} d_k \mathbf{x}_k\right)\right)\hat{B}_s^{A,B} \tag{14}$$

where $\mathbf{x}_k = \left(x_{k1}, .., x_{kj}, .., x_{kJ}\right)'$ is the auxiliary variable vector with $j = (1, ..., J)$. Since $\left(\sum_{k\in s_A} d_k \mathbf{x}_k + \sum_{k\in s_B} d_k \mathbf{x}_k\right)$ is always greater than $\sum_{k\in U} \mathbf{x}_k$, the term $\left(\sum_{k\in U} \mathbf{x}_k - \left(\sum_{k\in s_A} d_k \mathbf{x}_k + \sum_{k\in s_B} d_k \mathbf{x}_k\right)\right)\hat{B}_s^{A,B}$ in (14) can be viewed as a negative-sign correction factor for the biased summation of $\hat{Y}_{HT}^A$ and $\hat{Y}_{HT}^B$. All the JCE forms discussed above can be derived from the general JCE form in (14).

Another multivariate calibration estimator is a post-stratified estimator, corresponding to *a group mean model*, calibrating on known post-stratified cell counts. When the sizes of the population groups N_p and the classification vector used to code membership in one of P mutually exclusive and exhaustive groups are known, and $\mathbf{x}_k = \gamma_k = \left(\gamma_{1k}, ..., \gamma_{pk}, ..., \gamma_{Pk}\right)'$ is the auxiliary variable vector, where $\gamma_{pk} = 1$ for $k \in p$ and 0 otherwise, the calibrated estimator is the standard post-stratified estimator. The joint calibration factor is $N_p \Big/ \left(\sum_{k\in s_p^A} d_k + \sum_{k\in s_p^B} d_k\right)$, where s_p^A denotes the sample cell $U_p \cap s_A$ and s_p^B denotes the sample cell $U_p \cap s_B$. The calibrated estimator of the total can be written as $\hat{Y}_{JCE} = \sum_P \frac{N_p}{\left(\sum_{k\in s_p^A} d_k + \sum_{k\in s_p^B} d_k\right)}\left(\sum_{k\in s_p^A} d_k y_k + \sum_{k\in s_p^B} d_k y_k\right)$. In this *group mean model*, it is implicitly assumed that mean and variance are shared by all elements within the same group p as

$$\begin{cases} E(y_k) = \beta_p, \\ V(y_k) = \sigma_p^2. \end{cases} \tag{15}$$

Similarly, when the group totals X_p are known and $\mathbf{x}_k = x_k\gamma_k = \left(x_{1k}\gamma_{1k},\dots,x_{pk}\gamma_{pk},\dots,x_{Pk}\gamma_{Pk}\right)'$ is used as the auxiliary variables vector, this corresponds to *the group ratio model*, where mean and variance are shared by all elements within the same group *p* as

$$\begin{cases} E\left(y_k\right) = \beta_p x_k, \\ V\left(y_k\right) = \sigma_p^2 x_k. \end{cases} \tag{16}$$

Both *the group mean model* and *the group ratio model* can be classified under *the group model* of Särndal, Swensson & Wretman (1992).

6. The bias and the variance of the Joint Calibration Estimator

The JCE is a model-assisted design-based estimator for which the design-based bias properties are affected by the association between the study variable *y* and the auxiliary variable vector **x**.

6.1. Proposition 1

The bias of the JCE estimator $\hat{Y}_{JCE}$, in (5), is given approximately by

$$Bias\left(\hat{Y}_{JCE}\right) = \sum\nolimits_{k\in U_{ab}} e_k^{A,B} \tag{17}$$

where $e_k^{A,B} = \left(y_k - \mathbf{x}_k'\mathrm{B}_U^{A,B}\right)$ and

$\mathrm{B}_U^{A,B} = \left(\sum\nolimits_{k\in U_A}\mathbf{x}_k\mathbf{x}_k' + \sum\nolimits_{k\in U_B}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum\nolimits_{k\in U_A}\mathbf{x}_k y_k + \sum\nolimits_{k\in U_B}\mathbf{x}_k y_k\right)$ (see the appendix for proof).

Note that the dual frame estimation bias can be derived from expression (1) as

$$Bias\left(\hat{Y}_A + \hat{Y}_B\right) = \sum\nolimits_{k\in U_{ab}} y_k \ . \tag{18}$$

This means that the joint calibration approach uses $\mathbf{x}_k'\mathrm{B}_U^{A,B}$ to attenuate the bias for each $k \in U_{ab}$ to reduce the bias in (18). Therefore, the reduction in dual frame estimation bias due to the joint calibration is $\sum\nolimits_{k\in U_{ab}}\mathbf{x}_k'\mathrm{B}_U^{A,B}$, which is the difference between (17) and (18).

Proposition 1 highlights the need to identify powerful auxiliary variables that can predict study variable *y*. The more $\mathbf{x}_k'\mathrm{B}_U^{A,B}$ is able to predict y_k for each $k \in U_{ab}$, the greater the reduction in bias. The bias of $\hat{Y}_{JCE}$ in (17) is independent of the sampling design used to draw s_A and s_B as long as the set of auxiliary variables $\mathbf{x}_k$ is the same.

6.2. Corollary 1

When a linear relationship exists between the study variable y_k and the auxiliary vector $\mathbf{x}_k$, as in $y_k{=}\mathbf{x}'_k\mathrm{B}_U$, for every $k \in U$, the bias of the JCE estimator in (17) can be written as $Bias\left(\hat{Y}_{JCE}\right) = \sum_{k\in U_{ab}} \mathbf{x}'_k \left(\mathrm{B}_U - \mathrm{B}_U^{A,B}\right) = 0$.

This corollary is true because when a linear relationship between y_k and x_k exists, $\mathrm{B}_U^{A,B}{=}\mathrm{B}_U$, and the bias of $\hat{Y}_{JCE}$ is a function of the difference between two regression vectors $\mathrm{B}_U^{A,B}$ and B_U. This linear relationship will not hold in practice, but the bias in (17) will be reduced if the relationship between y_k and $\mathbf{x}_k$ is linear or nearly linear. The JCE bias is reduced to the extent that there are auxiliary variables $\mathbf{x}_k$ such that the residuals $e_k^{A,B} = \left(y_k - \mathbf{x}'_k\mathrm{B}_U^{A,B}\right)$ are small. Using such a set of auxiliary variables $\mathbf{x}_k$ guarantees reduced bias and variance of the JCE. Thus, the properties of the JCE are controlled by the association between *y* and **x**, where the best performance occurs when **x** more closely matches the population model or **x** includes strong correlates of *y*.

Assuming that the same model holds for all units in the population, $\frac{1}{N_{ab}}\sum_{k\in U_{ab}} e_k^{A,B}$ is asymptotically $N(0,V)$ where V is $O\left(N_{ab}^{-1}\right)$. The bias $Bias\left(\hat{\bar{Y}}_{JCE}\right) = \frac{1}{N}\sum_{k\in U_{ab}} e_k^{A,B}$ (where $\hat{\bar{Y}}_{JCE} = \hat{Y}_{JCE}/N$) converges in probability to 0 in large populations because the variance of the estimator $\hat{\bar{Y}}_{JCE} = \frac{N_{ab}}{N}\frac{1}{N_{ab}}\sum_{k\in U_{ab}} e_k^{A,B}$ is proportional to $P_{ab}^2 O\left(N_{ab}^{-1}\right) \approx \frac{P_{ab}}{N}$, and $\frac{N_{ab}}{N} \to P_{ab}$, and $\frac{P_{ab}}{N} \to 0$ as $N \to \infty$. Thus, the JCE estimator of the mean, $\hat{\bar{Y}}_{JCE}$, is a consistent estimator of population mean, $\bar{Y}$.

Under dual frame design, variance of $\hat{Y}_{JCE}$ can be written as $V\left(\hat{Y}_{JCE}\right) = \sum\sum_{k,l\in U_A} \Delta_{kl}^A \left(\frac{e_k^A}{\pi_k^A}\right)\left(\frac{e_l^A}{\pi_l^A}\right) + \sum\sum_{k,l\in U_B} \Delta_{kl}^B \left(\frac{e_k^B}{\pi_k^B}\right)\left(\frac{e_l^B}{\pi_l^B}\right) + \sum\sum_{k,l\in U_{ab}} \Delta_{kl}^{ab} \left(\frac{e_k^{ab}}{\pi_k^{ab}}\right)\left(\frac{e_l^{ab}}{\pi_l^{ab}}\right)$ where $s_{ab} = s_A \cap s_B$, for $D = (A,B,ab)$, $\Delta_{kl}^D = \left(\pi_{kl}^D - \pi_k^D \pi_l^D\right)$, $\pi_{kl}^D = p\left(k \,\&\, l \in s_D\right)$, $\pi_k^D = p\left(k \in s_D\right)$, $\pi_l^D = p\left(l \in s_D\right)$, $e_k^D = y_k - \mathbf{x}'_k \mathrm{B}_{U_D}$, and $\mathrm{B}_{U_D} = \sum_{k\in U_D} \mathbf{x}_k y_k \left(\sum_{U_D} \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$. Assuming small values of π_{kl}^{ab}, π_k^{ab} and π_l^{ab}, the estimated variance reduces to $\hat{v}\left(\hat{Y}_{JCE}\right) = \sum\sum_{k,l\in s_A} \frac{\Delta_{kl}^A}{\pi_{kl}}\left(w_k \hat{e}_k^A\right)\left(w_l \hat{e}_l^A\right) + \sum\sum_{k,l\in s_B} \frac{\Delta_{kl}^B}{\pi_{kl}}\left(w_k \hat{e}_k^B\right)\left(w_l \hat{e}_l^B\right)$ where $\hat{e}_k^D = y_k - \mathbf{x}'_k \hat{\mathrm{B}}_{ws_D}$, and $\hat{\mathrm{B}}_{ws_D} = \sum_{k\in s_D} w_k \mathbf{x}_k y_k \left(\sum_{k\in s_D} w_k \mathbf{x}_k \mathbf{x}'_k\right)^{-1}$.

7. Domain misclassification bias in dual frame estimation

Standard dual frame estimators depend on identifying the design domains during the data collection. Consequently, the performance of these estimators is sensitive to the errors in measuring the domain membership (Mecatti, 2007). Since it is uncommon to have access to the domain membership information before collecting the survey data, this information should be obtained during the data collection. For example, information about landline telephone service should be obtained in the area-landline dual frame surveys (Lepkowski & Groves, 1986) or about the landline and cell phone services should be obtained in the landline-cell dual frame telephone surveys (Brick et al., 2006; Kennedy, 2007). Collecting this information could be burdensome for some respondents and could lead to more unit non-response. It is even worse when dealing with rare populations such as persons with a rare disease or for elusive or hidden populations such as the homeless, illegal immigrants or drug consumers (Lepkowski, 1991; Mecatti, 2007; Sudman & Kalton, 1986).

Besides identifying the domain membership for every sampled unit, ideally, such information should be free from reporting or measurement errors, but this is not typically the case (Lohr & Rao, 2006). The correct classification of the sampled units into the domains in each frame is required to apply the standard dual frame estimators. In practice, achieving the correct classification for all cases is almost impossible because, as any other study variable, the domain membership variable could be affected by the measurement or the reporting error. Therefore, the sampled units could be misclassified into the wrong domain, leading to ***the domain misclassification***. For example, in RDD-cell phone dual frame surveys, households owning both landline and cell phone can be misclassified as landline only households or vice versa. Generally, it is difficult to identify misclassified units, and to estimate the misclassification rate. This means that the optimal dual frame estimators could have less than optimal performance (Lohr, 2011; Lohr & Rao, 2006).

The bias due to domain misclassification affects the standard dual frame estimators, however it does not affect the JCE; the latter does not necessarily require any domain membership information. In the presence of domain misclassification and where s_{mis} is the domain-misclassified sample s, the unconditional bias of the standard dual frame estimators in (1), $\hat{Y}_{mis}$, can be evaluated jointly with respect to the sampling design $p(s)$ and the conditional misclassification distribution $q(s_{mis} \mid s)$ as

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = E_p\left(E_q\left(\hat{Y}_{mis} \mid s\right)\right) - Y = E_{pq}\left(\hat{Y}_{mis}\right) - Y \ . \qquad (19)$$

7.1. Proposition 2

In the presence of the two-way misclassification (TWM), where $I_k^{ab,c}$ is a misclassification indicator for observation *k* from the overlapping domains s_{ab}^A and s_{ab}^B misclassified into non-overlapping domains s_a and s_b, respectively, and $I_k^{c,ab}$ is a misclassification indicator for observation *k* from s_a and s_b misclassified into s_{ab}^A and s_{ab}^B, respectively, a general expression for the unconditional bias that assumes each element *k* in the overlapping domain has a misclassification probability $E\left(I_k^{ab,c}\right)=\gamma_k^{ab,c}$ and each element *k* in the non-overlapping domains has a misclassification probability $E\left(I_k^{c,ab}\right)=\gamma_k^{c,ab}$, as derived in the appendix, can be written as

$$\begin{aligned}Bias_{pq}\left(\hat{Y}_{mis}\right) &= N_{ab}\left(\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right)+\bar{\gamma}^{ab,c}\bar{Y}_{ab}\right)-\\ &(1-\theta)N_a\left(\varsigma_a\left(\gamma_k^{c,ab}, y_k\right)+\bar{\gamma}_a^{c,ab}\bar{Y}_a\right)-\theta N_b\left(\varsigma_b\left(\gamma_k^{c,ab}, y_k\right)+\bar{\gamma}_b^{c,ab}\bar{Y}_b\right)\end{aligned} \tag{20}$$

where $\bar{Y}_{ab}=\sum_{k\in ab} y_k \big/ N_{ab}$, $\bar{\gamma}^{ab,c}=\sum_{k\in ab}\gamma_k^{ab,c}\big/ N_{ab}$, $\bar{Y}_a=\sum_{k\in a} y_k\big/ N_a$, $\bar{\gamma}_a^{c,ab}=\sum_{k\in a}\gamma_k^{c,ab}\big/ N_a$, $\bar{Y}_b=\sum_{k\in b} y_k\big/ N_b$ and $\bar{\gamma}_b^{c,ab}=\sum_{k\in b}\gamma_k^{c,ab}\big/ N_b$. $\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right)$ is the population covariance between the misclassification probabilities $\gamma_k^{ab,c}$ and the values of the target variable y_k within the overlapping domains *ab*. Also, $\varsigma_a\left(\gamma_k^{c,ab}, y_k\right)$ and $\varsigma_b\left(\gamma_k^{c,ab}, y_k\right)$ are the population covariance between the misclassification probabilities $\gamma_k^{c,ab}$ and the values of the target variable y_k within the non-overlapping domains *a* and *b*, respectively. These covariances can be written as follows

$$\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right)=\sum_{k\in ab}\left(\gamma_k^{ab,c}-\bar{\gamma}^{ab,c}\right)\left(y_k-\bar{Y}_{ab}\right)\big/ N_{ab}, \tag{21}$$

$$\varsigma_a\left(\gamma_k^{c,ab}, y_k\right)=\sum_{k\in a}\left(\gamma_k^{c,ab}-\bar{\gamma}_a^{c,ab}\right)\left(y_k-\bar{Y}_a\right)\big/ N_a, \tag{22}$$

$$\varsigma_b\left(\gamma_k^{c,ab}, y_k\right)=\sum_{k\in b}\left(\gamma_k^{c,ab}-\bar{\gamma}_b^{c,ab}\right)\left(y_k-\bar{Y}_b\right)\big/ N_b. \tag{23}$$

This means that the misclassification bias depends on two components:

a) The expected total of y_k for the misclassified cases within each domain, $N_{ab}\bar{\gamma}^{ab,c}\bar{Y}_{ab}$, $N_a\bar{\gamma}_a^{c,ab}\bar{Y}_a$ and $N_b\bar{\gamma}_b^{c,ab}\bar{Y}_b$.

b) The correlation between the misclassifications probabilities and the study variable *y* within the different design domains, supported by the within domains covariances, $\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right)$, $\varsigma_a\left(\gamma_k^{c,ab}, y_k\right)$ and $\varsigma_b\left(\gamma_k^{c,ab}, y_k\right)$.

In general, the misclassification bias can be controlled during the data collection process by following the best practices that decrease the measurement error in reporting the domain membership variable. At the same time, the misclassification bias can be adjusted based on the second component by implicitly predicting the misclassification probabilities. This can be performed by calibrating the data by an auxiliary variable that is correlated with the study variable y and the misclassification probabilities. This step can be performed either in the standard dual frame estimators or in the JCE. In the standard dual frame estimators, the calibration step comes after combining the data based on the misclassified domains. When misclassification probabilities are known, Lohr (2011) proposed an adjustment factor for the misclassification bias for the FWE estimator, which is consistent with our derivations of the misclassification bias.

In the JCE, the domain misclassification does not affect the estimates as long as no domain membership information was added to the auxiliary variable vector, $\mathbf{x}$. However, even if misclassified domain membership information was added to the auxiliary variable vector, adding more auxiliary variables which are correlated with the study variable y and the misclassification probabilities is enough to reduce the bias resulted from the misclassified domain. Moreover, the effect of using the misclassified domains as the sole auxiliary variable in the JCE is less significant than the effect of the domain misclassification in the standard dual frame estimators. This is due the fact that in the standard dual frame estimators, classifying the sampling units into the domain correctly is required before applying the composite factor θ . However, in the JCE, this misclassification error is accounted for as a measurement in the auxiliary variables.

8. Simulation studies

In this section, two simulation studies are presented. The first one is to examine the performance of the JCE estimator in comparison with the FWE estimator under different population models. These population models determine the relationship between the study variable and the calibration auxiliary variables. The second simulation study considers the domain misclassification errors and examines the performance of the JCE and FWE estimators in the presence of these errors.

8.1. The first study: design

A simulation study was used to evaluate the performance of the JCE relative to the FWE dual frame estimator. A finite population of size $N = 100{,}000$ with domains population sizes $N_a = 40{,}000$, $N_{ab} = 50{,}000$ and $N_b = 10{,}000$ was generated with frame sizes $N_A = 90{,}000$ (all cases in domains a and ab) and $N_B = 60{,}000$ (all cases in domains ab and b). H = 6 population strata had sizes $N_1 = 10{,}000$, $N_2 = 20{,}000$, $N_3 = 30{,}000$, $N_4 = 25{,}000$, $N_5 = 5{,}000$ and $N_6 = 10{,}000$.

The distribution of the population elements over the strata and the domains is presented in Table 1. As shown, strata 1 and 2 are unique to frame A, strata 3-5 are in both frame A and B and stratum 6 elements are present only on frame B.

Table 1. Distribution of the population elements over the six strata and the three domains.

Strata	Frames and domains			
	A			Total
		B		
	a	*ab*	*b*	
1	10,000			10,000
2	20,000			20,000
3	10,000	20,000		30,000
4		25,000		25,000
5		5,000		5,000
6			10,000	10,000
Total	40,000	50,000	10,000	100,000

Source: Own elaboration.

The data values for the variable of interest, *y*, were generated under two models. The first population model is a *common linear regression model* (CLR), $y_{jk} = x_{jk} + \varepsilon_{jk}$, for $k = 1,..,N$ and $j = 1,...,6$ strata, where $x_{jk} \sim N\left(\mu_x, \sigma_x\right)$ and $\varepsilon_{jk} \sim N\left(\mu_x, \sigma_x\right)$. Here, the mean of *y* is the same for all population strata and design domains. The second population model is a *group linear regression model* (GLR), which can be written as the first model but with $x_{jk} \sim N\left(\mu_{xj}, \sigma_x\right)$ and $\varepsilon_{jk} \sim N\left(\mu_\varepsilon, \sigma_\varepsilon\right)$. In both models, an auxiliary variable, z_{dk}, was generated as $z_{dk} = \beta_o + \beta_d + \varepsilon_{dk}$, for $d = \left(\text{a,ab,b}\right)$ where $\beta_o = 200$ and $\varepsilon_{dk} \sim N\left(0, 350\right)$. For both the first and the second model, the simulation factors were as follows:

1. Sampling Designs
   a) Simple random samples from both frames.
   b) Stratified sample with equal allocation across five strata from frame A, and a simple random sample from frame B.
2. Domain means
   a) Small-differences in domain means, $\beta_a = 5$, $\beta_{ab} = 6$ and $\beta_b = 7$.
   b) Frame-different means, $\beta_a = 5$, $\beta_{ab} = 5$ and $\beta_b = 10$.
   c) Large-differences in domain means, $\beta_a = 5$, $\beta_{ab} = 10$ and $\beta_b = 15$.

3. Correlation between *y* and *x*

    a) $\rho_{xy} = 0.40$.

    b) $\rho_{xy} = 0.60$.

    c) $\rho_{xy} = 0.80$.

The correlation levels determined population model parameters (see Table 2). Both σ_x and σ_ε were deliberately manipulated to generate each correlation level. Since μ_{xj} does not contribute to the correlation, it is almost fixed across the correlation levels but is different across the six strata.

Table 2. Model parameters based on correlation levels between y_{jk} and x_{jk} .

Model parameters	$\rho_{xy} = 0.40$	$\rho_{xy} = 0.60$	$\rho_{xy} = 0.80$
CLR Model			
$x_{jk} \sim N(\mu_X, \sigma_X)$	$N(750,192)$	$N(780,288)$	$N(760,384)$
$\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$	$N(0,440)$	$N(0,384)$	$N(0,288)$
GLR Model			
$x_{1k} \sim N(\mu_{x1}, \sigma_x)$	$N(487,192)$	$N(500,288)$	$N(480,384)$
$x_{2k} \sim N(\mu_{x2}, \sigma_x)$	$N(618,192)$	$N(640,288)$	$N(620,384)$
$x_{3k} \sim N(\mu_{x3}, \sigma_x)$	$N(750,192)$	$N(780,288)$	$N(760,384)$
$x_{4k} \sim N(\mu_{x4}, \sigma_x)$	$N(881,192)$	$N(919,288)$	$N(900,384)$
$x_{5k} \sim N(\mu_{x5}, \sigma_x)$	$N(1013,192)$	$N(1059,288)$	$N(1039,384)$
$x_{6k} \sim N(\mu_{x6}, \sigma_x)$	$N(487,192)$	$N(500,288)$	$N(479,384)$
$\varepsilon_{jk} \sim N(\mu_\varepsilon, \sigma_\varepsilon)$	$N(0,440)$	$N(0,384)$	$N(0,288)$

Source: Own elaboration.

The simulation factors combined to form 36 simulation studies, 18 studies under each population model. One thousand replicates of initial samples of 1,000 cases each were run for each study, resulting in a standard error less than 60 for the difference in the biases between the FWE and JCE estimators.

To simulate a dual frame design within each simulation replicate, two equally allocated samples were independently drawn from both frames A and B, with $n_A = n_B = 500$. These samples were 'stacked' to form each dual frame sample, *s=1,…,1000.*

8.2. The first study: comparison estimators

For each of the 1000 samples generated for each of the 36 sets of simulation conditions, dual frame estimates were then calculated for each simulated dual frame sample. The FWE with $\theta = 0.5$ was the standard fixed weight dual frame estimator, $\hat{Y}_{FWE}$. That is, the base weights for the probability samples from frames A and B were adjusted using a composite factor $\theta = 0.5$. Three calibrated versions of the FWE estimator were also applied to simulated dual-frame sample data. For the calibrated versions of the FWE, besides the population size *N*, the dual frame adjusted base weights were calibrated to the auxiliary totals for three combinations of *x* and *z* (*x* only, *z* only and *x* and *z* together) resulting in the calibrated versions $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$.

For the JCEs, the base weights, d_k^A and d_k^B were used for each sample, and the auxiliary variables *x* and *z* were used to calibrate the base weights directly. Six versions of the JCE estimator were applied, each differing in the set of auxiliary population controls included in the joint calibration of the dual frame sample estimates. Controls to *x* and *z* singly or in combination are denoted by $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. Also, $\hat{Y}_{JCE.xH}$ was produced using the auxiliary variables *x* and $\mathbf{H} = (h_1, \ldots., h_6)$, where **H** is a vector of population group identifiers for the six design strata. Additionally, in conjunction with the primary calibration variables, *x*, population totals for the design domains, D = (*a, ab, b*), and frames, F = (*A,B*), were also used to calibrate the adjusted base weights resulting in two additional JCEs, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xF}$.

The biases in the JCE and the FWE estimates for each simulation specification were estimated as a difference between the average of the 1000 survey estimates $\hat{Y}_s, s = 1, \ldots, 1000$, and the population total *Y* from the synthetic finite population. The Relative Bias (RB) of parameter estimates was computed as $RB = \left(\left(\sum_{i \in 1000} \hat{Y}_i \big/ 1000\right) - Y\right) \times 100 \big/ Y$. Similarly, the Relative Root Mean Squared Error (RMSE) for each estimator was computed as $RMSE = \sqrt{\left(\sum_{i \in 1000} \left(\hat{Y}_i - Y\right)^2 \big/ 1000\right)} \times 100 \big/ Y$ for each simulation specification. We also calculated the RB and RMSE for the summation of the dual frame samples estimates, $\hat{Y}_{s_A} + \hat{Y}_{s_B}$. Although it is a biased estimator, this summation is used in the comparisons to indicate the reduction in bias resulted from the FWE and JCE estimators. Here, only results for the simple random sampling design are discussed. Simulation results for the stratified sampling design specification show the same patterns of results, consistent with Proposition 1.

8.3. The first study: results

Tables 3 to 6 summarize the results of the simulation study, comparing the RB and RMSE for the various FWE and JCE estimators. As shown in Tables 3 and 5, the standard estimator $\hat{Y}_{FWE}$ and its calibrated versions, $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, achieve unbiased estimates. Only under the GLM model in Table 5, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are subject to higher relative biases than $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively. Thus, under the GLR model in which the stratum-specific relationship of *y* to *x* and domain-specific relationship of *y* to *z* differs in a significant way, jointly calibrating 'stacked' samples directly by *z* or *x*, as in $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$, is not a satisfactory estimation method. However, we do see that the higher the correlation between *y* and *x*, the lower the relative biases in $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. The same patterns of results apply under the other domain mean distributions.

Under the GLM, adding stratum population controls to the calibration in $\hat{Y}_{JCE.xH}$ results in nearly unbiased estimates, regardless of the correlation between *y* and *x*. Also, adding the domain totals or the frame totals to the vector of calibration auxiliary variables, as in $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xF}$, achieves unbiased estimates, and yielded identical RB and RMSE values. Either under the CLM or the GLM model, the domain means have very little effect on the relative biases of the JCE estimators. The RMSEs in Tables 4 and 6 show the same patterns as the RBs. However, the higher the correlation between *y* and *x* the lower the RMSE in $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. The same patterns of results apply under the other domain mean distributions.

Table 3. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under simple sampling design.

Domain means	ρρ_{xy}	$\hat{Y}_{s_A}+\hat{Y}_{s_B}$	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
β_d =(5,6,7)	ρ_{xy} = 0.40	58.64	0	-0.01	0.01	-0.01	-0.03	-0.01	-0.03
β_d =(5,5,10)	ρ_{xy} = 0.40	58.73	0.03	0.02	0.06	0.06	0.08	0.06	0.08
β_d =(5,10,15)	ρ_{xy} = 0.40	58.6	-0.03	-0.04	-0.07	-0.05	-0.06	-0.05	-0.06
β_d =(5,6,7)	ρ_{xy} = 0.60	59.1	-0.04	-0.05	-0.08	-0.02	-0.04	-0.02	-0.04
β_d =(5,5,10)	ρ_{xy} = 0.60	59.17	0	-0.01	0	0.06	0.08	0.06	0.08
β_d =(5,10,15)	ρ_{xy} = 0.60	58.74	0.1	0.09	0.09	0.07	0.09	0.07	0.09
β_d =(5,6,7)	ρ_{xy} = 0.80	58.92	0.01	0	0.07	-0.05	-0.01	-0.05	-0.02
β_d =(5,5,10)	ρ_{xy} = 0.80	59.22	0.04	0.03	0.07	-0.02	0.02	-0.02	0.02
β_d =(5,10,15)	ρ_{xy} = 0.80	59.19	-0.09	-0.1	-0.12	-0.04	-0.09	-0.04	-0.09

Source: Own elaboration.

Table 4. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under simple sampling design.

Domain means	ρ_{xy}	$\hat{Y}_{S_A}+\hat{Y}_{S_B}$	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$
β_d =(5,6,7)	ρ_{xy} = 0.40	58.74	2.27	1.93	1.8	1.74	1.63	1.74	1.64
β_d =(5,5,10)	ρ_{xy} = 0.40	58.83	2.3	1.96	1.79	1.81	1.65	1.81	1.65
β_d =(5,10,15)	ρ_{xy} = 0.40	58.7	2.24	1.95	1.82	1.76	1.65	1.76	1.65
β_d =(5,6,7)	ρ_{xy} = 0.60	59.18	2.29	2	1.84	1.62	1.49	1.62	1.49
β_d =(5,5,10)	ρ_{xy} = 0.60	59.26	2.27	1.94	1.79	1.58	1.44	1.58	1.44
β_d =(5,10,15)	ρ_{xy} = 0.60	58.83	2.18	1.87	1.74	1.51	1.41	1.52	1.41
β_d =(5,6,7)	ρ_{xy} = 0.80	59.01	2.3	2.01	1.87	1.21	1.14	1.22	1.14
β_d =(5,5,10)	ρ_{xy} = 0.80	59.32	2.33	2.04	1.88	1.21	1.1	1.21	1.1
β_d =(5,10,15)	ρ_{xy} = 0.80	59.29	2.32	2.06	1.86	1.21	1.11	1.21	1.11

Source: Own elaboration.

Table 5. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under simple sampling design.

Domain means	ρ_{xy}	$\hat{Y}_{S_A}+\hat{Y}_{S_B}$	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xF}$
β_d = (5,6,7)	ρ_{xy} = 0.40	58.64	0.02	0.02	5.76	-0.03	3.8	-0.02	3.8	-0.08	-0.03	-0.03
β_d = (5,5,10)	ρ_{xy} = 0.40	58.73	0.13	0.13	5.82	0.14	3.82	0.14	3.82	0.07	0.13	0.13
β_d = (5,10,15)	ρ_{xy} = 0.40	58.6	0.07	0.07	5.73	0.08	3.82	0.09	3.81	0.04	0.08	0.08
β_d = (5,6,7)	ρ_{xy} = 0.60	59.1	0.07	0.07	6.06	0.07	3.37	0.07	3.36	0.07	0.07	0.07
β_d = (5,5,10)	ρ_{xy} = 0.60	59.17	0.05	0.05	6.11	0.11	3.4	0.11	3.4	0.09	0.10	0.10
β_d = (5,10,15)	ρ_{xy} = 0.60	58.74	-0.12	-0.11	5.83	-0.04	3.24	-0.04	3.25	-0.09	-0.05	-0.05
β_d = (5,6,7)	ρ_{xy} = 0.80	58.92	-0.13	-0.12	5.95	0.01	2.47	0.02	2.47	-0.04	0.01	0.01
β_d = (5,5,10)	ρ_{xy} = 0.80	59.22	0.02	0.03	6.15	-0.02	2.47	-0.01	2.47	-0.03	-0.02	-0.02
β_d = (5,10,15)	ρ_{xy} = 0.80	59.19	0.07	0.08	6.13	0.00	2.47	0.01	2.47	-0.04	0.00	0.00

Source: Own elaboration.

Table 6. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under simple sampling design.

Domain means	ρ$_{xy}$	$\hat{Y}_{s_A}$ +	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xF}$
β$_d$ = (5,6,7)	ρ$_{xy}$ = 0.40	58.74	2.44	2.45	6.19	2.22	4.3	2.22	4.31	2.15	2.19	2.19
β$_d$ = (5,5,10)	ρ$_{xy}$ = 0.40	58.83	2.47	2.47	6.23	2.17	4.28	2.17	4.28	2.09	2.14	2.14
β$_d$ = (5,10,15)	ρ$_{xy}$ = 0.40	58.7	2.38	2.45	6.17	2.2	4.33	2.2	4.33	2.11	2.14	2.14
β$_d$ = (5,6,7)	ρ$_{xy}$ = 0.60	59.18	2.29	2.32	6.42	1.81	3.77	1.8	3.76	1.73	1.76	1.76
β$_d$ = (5,5,10)	ρ$_{xy}$ = 0.60	59.26	2.35	2.42	6.49	1.92	3.82	1.92	3.82	1.83	1.88	1.88
β$_d$ = (5,10,15)	ρ$_{xy}$ = 0.60	58.83	2.34	2.35	6.21	1.83	3.66	1.83	3.67	1.76	1.8	1.8
β$_d$ = (5,6,7)	ρ$_{xy}$ = 0.80	59.01	2.33	2.38	6.34	1.44	2.8	1.44	2.8	1.37	1.41	1.41
β$_d$ = (5,5,10)	ρ$_{xy}$ = 0.80	59.32	2.43	2.53	6.57	1.47	2.81	1.47	2.82	1.39	1.43	1.43
β$_d$ = (5,10,15)	ρ$_{xy}$ = 0.80	59.29	2.42	2.46	6.53	1.39	2.79	1.39	2.79	1.33	1.36	1.36

Source: own elaboration

8.4. The second study: design

The same synthetic population and population models used in the first study have been used in the second study. The simulation factors are as the following:

1. Sampling Designs: Simple Sampling Design where simple random samples were selected from both frames.
2. Domain means: Large-difference domains' means where $\beta_a = 5$, $\beta_{ab} = 10$ and $\beta_b = 15$.
3. Correlation between y_{jk} and x_{jk}: The population correlation coefficient is $\rho_{xy} = 0.40$.
4. Misclassification mechanisms:
   a) The one-way OWOM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(ab,a)} = 0.1$ and $\gamma^{B(ab,b)} = 0.1$. This means that 10% of the sample A overlapping domain *ab* cases are misclassified in non-overlapping domain *a* and 10% of the sample B

overlapping domain *ab* cases are misclassified in non-overlapping domain *b*.

b) The one-way OWNM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(a,ab)} = 0.1$ and $\gamma^{B(b,ab)} = 0.1$. This means that 10% of the sample A non-overlapping domain *a* cases are misclassified in overlapping domain *ab* and 10% of the sample B non-overlapping domain *b* cases are misclassified in overlapping domain *ab*.

c) The two-way TWM misclassification mechanism, where the misclassification probabilities were $\gamma^{A(a,ab)} = 0.1$, $\gamma^{B(b,ab)} = 0.1$, $\gamma^{A(a,ab)} = 0.1$ and $\gamma^{B(b,ab)} = 0.1$.

These sets of simulation factors combine to form 6 simulation studies, 3 simulation studies for each population model. To simulate a dual frame design, within each simulation replicate, two equal-size samples were drawn separately from both frames A and B, where $n_A = n_B = 500$. These samples were 'stacked' to form dual frame sample *s*. Conditional on the misclassification mechanisms, the misclassified domains were generated.

8.5. The second study: comparison estimators

Besides the estimators used in the first study, more estimators have been calculated in the second study such as $\hat{Y}_{JCE.zH}$, $\hat{Y}_{JCE.xzH}$, $\hat{Y}_{JCE.D}$ and $\hat{Y}_{JCE.xzD}$.

8.6. The second study: results

Generally, as indicated in Tables 7 and 9, in the presence of domain misclassification, biases in $\hat{Y}_{FWE}$ are present. Under the CLR model, in Table 7, the standard estimator $\hat{Y}_{FWE}$ is affected by the misclassification error, whereas the proposed estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are not. Adding the calibration in the standard estimators $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$ reduces the misclassification bias and achieved relative biases comparable to the JCE estimators, $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$. Interestingly, adding the misclassified domain variable to the auxiliary variable vector in the JCE estimators, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xzD}$, does not result in misclassification-biased estimates as in $\hat{Y}_{FWE}$. Even calibrating only by the misclassified domains in $\hat{Y}_{JCE.D}$ results in almost unbiased estimates. Generally, the relative mean square errors, in Table 8, show the same patterns as the relative biases, in Table 7. However, RMSEs for $\hat{Y}_{JCE.z}$ and $\hat{Y}_{JCE.x}$ were slightly lower than RMSEs for $\hat{Y}^{cal}_{FWE.z}$ and $\hat{Y}^{cal}_{FWE.x}$, respectively.

Under the GLR model, in Table 9, the JCE estimators $\hat{Y}_{JCE.z}$, $\hat{Y}_{JCE.x}$ and $\hat{Y}_{JCE.xz}$ are subject to higher relative biases than $\hat{Y}^{cal}_{FWE.z}$, $\hat{Y}^{cal}_{FWE.x}$ and $\hat{Y}^{cal}_{FWE.xz}$, respectively. Adding the strata totals to the calibration in $\hat{Y}_{JCE.zH}$, $\hat{Y}_{JCE.xH}$ and $\hat{Y}_{JCE.xzH}$ resulted in reduced relative biases. Adding the misclassified domain variable to the auxiliary variable vector in the JCE estimators, $\hat{Y}_{JCE.D}$, $\hat{Y}_{JCE.xD}$ and $\hat{Y}_{JCE.xzD}$, does not result in misclassification-biased estimates as in $\hat{Y}_{FWE}$. The relative mean square errors show similar patterns to relative biases, as indicated in Table 10.

9. Discussion

The JCE proposed here is a new model-assisted design-based dual frame estimator that can achieve efficiency parallel to that of the standard dual frame estimators. In the simulation studies, the JCEs achieved RBs and RMSEs comparable to those for the standard FWEs. JCEs for point estimates are easier to apply than the FWEs in practice, because they do not require information about domain membership. They also can be computed using standard survey software.

In dual frame designs, two types of variables may affect the accuracy of the estimators. The first is the auxiliary variables **x** associated with the study variable *y*. The second is the variables associated with the sample design such as the design domains, *D*. Regardless of the relation between *y* and *D*, when accurate information about the design domains is available, adding it to the JCE auxiliary variable vector results in unbiased estimates of the population total. Adding domain (*D)* population totals to the auxiliary variable vector results in an estimator which is identical to the standard FWE dual frame estimator with $\theta = 0.5$. When a strong relationship exists between auxiliary variables, **z** and *D*, adding **z** to the JCE auxiliary variable vector results in reduced-biased estimates. When a strong association exists between **x** and *y*, adding **x** to the JCE auxiliary variable vector results in almost unbiased estimates, a result that can be attributed to the fact that adding **x** to the auxiliary variable vector results in a calibration model that closely matches the population model, and hence unbiased estimates.

Table 7. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassi-fication	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	5.02	-0.04	-0.07	0.00	-0.03	0.00	-0.03	-0.05	-0.02	-0.02
OWNM	- 2.46	0.02	-0.05	0.03	-0.03	0.02	-0.03	0.03	0.03	0.02
TWM	2.48	-0.12	-0.12	-0.10	-0.07	-0.10	-0.07	-0.09	-0.08	-0.07

Source: Own elaboration.

Table 8. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the CLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassi-fication	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	5.55	1.94	1.84	1.78	1.69	1.78	1.70	1.94	1.78	1.78
OWNM	3.31	1.93	1.79	1.77	1.63	1.77	1.63	1.94	1.79	1.79
TWM	3.43	1.92	1.82	1.74	1.64	1.74	1.64	1.90	1.72	1.72

Source: Own elaboration.

Table 9. Simulation RB (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassi-fication	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zH}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xzH}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	6.05	0.95	5.78	0.63	3.84	0.63	3.83	0.10	0.07	0.07	0.16	0.12	0.12
OWNM	-2.08	0.43	5.71	0.27	3.75	0.27	3.75	-0.02	-0.05	-0.05	0.01	-0.01	-0.01
TWM	3.96	1.34	5.74	0.91	3.80	0.91	3.80	0.07	0.06	0.06	0.14	0.11	0.11

Source: Own elaboration.

Table 10. Simulation RMSE (%) for the FWE and JCE estimators of $\hat{Y}$ estimated from the GLR model population under $\rho_{xy} = 0.40$ in the presence of the misclassification errors.

Misclassi-fication	$\hat{Y}_{FWE}$	$\hat{Y}^{cal}_{FWE.z}$	$\hat{Y}_{JCE.z}$	$\hat{Y}^{cal}_{FWE.x}$	$\hat{Y}_{JCE.x}$	$\hat{Y}^{cal}_{FWE.xz}$	$\hat{Y}_{JCE.xz}$	$\hat{Y}_{JCE.zH}$	$\hat{Y}_{JCE.xH}$	$\hat{Y}_{JCE.xzH}$	$\hat{Y}_{JCE.D}$	$\hat{Y}_{JCE.xD}$	$\hat{Y}_{JCE.xzD}$
OWOM	6.58	2.63	6.20	2.30	4.35	2.30	4.35	2.25	2.08	2.08	2.38	2.18	2.17
OWNM	3.15	2.43	6.13	2.19	4.25	2.19	4.25	2.24	2.07	2.07	2.33	2.13	2.13
TWM	4.68	2.75	6.16	2.34	4.31	2.35	4.31	2.24	2.06	2.06	2.30	2.10	2.10

Source: Own elaboration.

Generally, the performance of the JCE depends on the extent of agreement between the population model and the working model in the calibration. It depends to a lesser degree on the association between the auxiliary variables, including the domain data, and the study variable. When the auxiliary vector or the implicit calibration model more closely matches the population model, the JCEs yield almost unbiased estimates. When the models do not agree, the JCEs have a higher level of bias than the standard FWEs. Thus, the extent of the association between the study variable y and the auxiliary variable x is an important determinant factor in JCE performance.

The JCE ought to be preferred to the standard dual frame estimators. It only depends on calibrating pooled datasets to available auxiliary variables. Unlike the optimal dual frame estimators, the JCE yields only one weight variable to be used in estimation, assuming that an agreement between the population model and the working model for the most important study variables can be fulfilled. And the JCE can be easily extended to the multiple frame case.

In this paper, the domain misclassification was introduced as a form of the non-sampling error, which could affect the bias properties of the dual frame estimators. The effect of the domain misclassification exceeds its effect as a type of measurement or reporting error in the domain membership information. The misclassified domains may affect the standard dual frame estimators substantially. This is due to the fact that the standard dual frame estimators require accurate information about the domain membership. Based on this information, the adjustment factor is applied to the design weights for dual frame estimation.

We derived a general expression for the analytic bias that results when the standard dual frame estimators are applied to data with misclassified dual frame domains. The bias expression indicated that the correlation between the misclassification probabilities and the study variable y within each domain is an important determinant of the misclassification bias. Also, the expected total of the y variable for the misclassified cases within each domain is another determinant of the misclassification bias. Controlling these two determinants could be the key for reducing the misclassification bias in the standard dual frame estimators.

In addition to introducing the domain misclassification problem in this paper, the JCE was highlighted as a robust dual frame estimator to the domain misclassification error. The JCE does not necessarily need any information about the domain classification. Therefore, the misclassification problem does not affect the JCE estimates as long as the domain membership information was not added to the calibration auxiliary variable vector. Interestingly, adding the misclassified domains to the JCE auxiliary variable vector does not lead to substantially biased estimates, as long as the domains are misclassified at random. This is due to the fact that the effect of the misclassified domains in the context of the JCE is a measurement error effect.

Finally, in this paper, the JCE was introduced for dual frame estimation. However, in the future the JCE could be extended to be a general approach for combining data from multiple sources. For example, multiple datasets from different surveys could be combined to provide more accurate estimates for study variables that are commonly collected in these surveys.

Acknowledgements

This work was completed as part of the first author’s doctoral dissertation at the Program in Survey Methodology, University of Michigan. He thanks the members of his committee, Steven Heeringa, James Lepkowski, Sunghee Lee and Richard Valliant for their guidance and constructive comments that improved this paper.

APPENDICES

Proof of proposition 1

Where the calibration estimator in (5) is equivalent to the generalized JCE estimator in (14), the JCE can be written as

$$\hat{Y}_{JCE} = \sum\nolimits_{k\in U} \hat{y}_k + \sum\nolimits_{k\in s_A} d_k\left(y_k - \hat{y}_k\right) + \sum\nolimits_{k\in s_B} d_k\left(y_k - \hat{y}_k\right)$$

where $\hat{y}_k = \mathbf{x}_k'\hat{B}_s^{A,B}$

$$\hat{Y}_{JCE} = \sum\nolimits_{k\in U} \mathbf{x}_k'\hat{B}_s^{A,B} + \sum\nolimits_{k\in s_A} d_k y_k - \sum\nolimits_{k\in s_A} d_k\mathbf{x}_k'\hat{B}_s^{A,B} + \sum\nolimits_{k\in s_B} d_k y_k - \sum\nolimits_{k\in s_B} d_k\mathbf{x}_k'\hat{B}_s^{A,B}$$

$$\hat{Y}_{JCE} - Y = \sum\nolimits_{k\in U} \mathbf{x}_k'\hat{B}_s^{A,B} + \sum\nolimits_{k\in s_A} d_k y_k - \sum\nolimits_{k\in s_A} d_k\mathbf{x}_k'\hat{B}_s^{A,B} + \sum\nolimits_{k\in s_B} d_k y_k - \sum\nolimits_{k\in s_B} d_k\mathbf{x}_k'\hat{B}_s^{A,B} - \sum\nolimits_{k\in U} y_k$$

$$\begin{aligned}\hat{Y}_{JCE} - Y = {} & \sum\nolimits_{k\in U} \mathbf{x}_k'\hat{B}_s^{A,B} + \sum\nolimits_{k\in s_A} d_k y_k - \sum\nolimits_{k\in s_A} d_k\mathbf{x}_k'\hat{B}_s^{A,B} + \sum\nolimits_{k\in s_B} d_k y_k - \sum\nolimits_{k\in s_B} d_k\mathbf{x}_k'\hat{B}_s^{A,B} - \sum\nolimits_{k\in U} y_k \\ & - \sum\nolimits_{k\in U} \mathbf{x}_k'\mathrm{B}_U - \sum\nolimits_{k\in s_A} d_k\mathbf{x}_k'\mathrm{B}_U - \sum\nolimits_{k\in s_B} d_k\mathbf{x}_k'\mathrm{B}_U + \sum\nolimits_{k\in U} \mathbf{x}_k'\mathrm{B}_U + \sum\nolimits_{k\in s_A} d_k\mathbf{x}_k'\mathrm{B}_U + \sum\nolimits_{k\in s_B} d_k\mathbf{x}_k'\mathrm{B}_U\end{aligned}$$

where $e_k = y_k - \mathbf{x}_k'\mathrm{B}_U$ and $\mathrm{B}_U = \left(\sum\nolimits_{k\in U}\mathbf{x}_k y_k\right)\left(\sum\nolimits_{k\in U}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}$

$$\hat{Y}_{JCE} - Y = A + C$$

where

$$A = \sum\nolimits_{k\in s_A} d_k e_k + \sum\nolimits_{k\in s_B} d_k e_k - \sum\nolimits_{k\in U} e_k$$

$$C = \left(\sum\nolimits_{k\in U}\mathbf{x}_k' - \sum\nolimits_{k\in s_A} d_k\mathbf{x}_k' - \sum\nolimits_{k\in s_B} d_k\mathbf{x}_k'\right)\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right)$$

$$E\left(\hat{Y}_{JCE} - Y\right) = E(A) + E(C)$$

$$E(A) = \sum\nolimits_{k\in U_A} e_k + \sum\nolimits_{k\in U_B} e_k - \sum\nolimits_{k\in U} e_k = \sum\nolimits_{k\in U_{ab}} e_k$$

$$\begin{aligned}E(C) &= E\left(\sum\nolimits_{k\in U}\mathbf{x}_k' - \sum\nolimits_{k\in s_A} d_k\mathbf{x}_k' - \sum\nolimits_{k\in s_B} d_k\mathbf{x}_k'\right).E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right) \\ &= -\sum\nolimits_{k\in U_{ab}}\mathbf{x}_k'.E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right)\end{aligned}$$

By Taylor Linearization, the estimator $\hat{B}_s^{A,B}$ can be defined as

$$\begin{aligned}\hat{B}_s^{A,B} = {} & \mathrm{B}_U^{A,B} + \left(\sum\nolimits_{k\in U'}\mathbf{x}_k\mathbf{x}_k'\right)^{-1}\left(\sum\nolimits_{k\in s'} d_k\mathbf{x}_k y_k - \sum\nolimits_{k\in U'}\mathbf{x}_k y_k\right) \\ & - \sum\nolimits_{k\in U'}\mathbf{x}_k y_k\left(\sum\nolimits_{k\in U'}\mathbf{x}_k\mathbf{x}_k'\right)^{-2}\left(\sum\nolimits_{k\in s'} d_k\mathbf{x}_k\mathbf{x}_k' - \sum\nolimits_{k\in U'}\mathbf{x}_k\mathbf{x}_k'\right)\end{aligned}$$

where

$$\sum\nolimits_{k\in s'} d_k \mathbf{x}_k y_k = \sum\nolimits_{k\in s_A} d_k \mathbf{x}_k y_k + \sum\nolimits_{k\in s_B} d_k \mathbf{x}_k y_k$$

$$\sum\nolimits_{k\in s'} d_k \mathbf{x}_k \mathbf{x}'_k = \sum\nolimits_{k\in s_A} d_k \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{k\in s_B} d_k \mathbf{x}_k \mathbf{x}'_k$$

$$\sum\nolimits_{k\in U'} \mathbf{x}_k y_k = \sum\nolimits_{k\in U_A} \mathbf{x}_k y_k + \sum\nolimits_{k\in U_B} \mathbf{x}_k y_k = \sum\nolimits_{k\in U} \mathbf{x}_k y_k + \sum\nolimits_{k\in U_{ab}} \mathbf{x}_k y_k$$

$$\sum\nolimits_{k\in U'} \mathbf{x}_k \mathbf{x}'_k = \sum\nolimits_{k\in U_A} \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{k\in U_B} \mathbf{x}_k \mathbf{x}'_k = \sum\nolimits_{k\in U} \mathbf{x}_k \mathbf{x}'_k + \sum\nolimits_{k\in U_{ab}} \mathbf{x}_k \mathbf{x}'_k$$

$$\hat{B}_s^{A,B} = \left(\sum\nolimits_{k\in s'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in s'} d_k \mathbf{x}_k y_k\right)$$

$$\mathrm{B}_U^{A,B} = \left(\sum\nolimits_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U'} \mathbf{x}_k y_k\right)$$

$$E\left(\hat{B}_s^{A,B}\right) = \mathrm{B}_U^{A,B} = \mathrm{B}_U + \mathrm{B}_U^{A,B} - \mathrm{B}_U$$

$$E\left(\hat{B}_s^{A,B}\right) = \mathrm{B}_{\mathrm{U}} + \left(\sum\nolimits_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum\nolimits_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U} \mathbf{x}_k y_k\right)$$

$$E\left(\hat{B}_s^{A,B} - \mathrm{B}_U\right) = \left(\sum\nolimits_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum\nolimits_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U} \mathbf{x}_k y_k\right)$$

$$\therefore E(C) = -\sum\nolimits_{k\in U_{ab}} \mathbf{x}'_k \left(\left(\sum\nolimits_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum\nolimits_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U} \mathbf{x}_k y_k\right)\right)$$

Consequently, under dual frame design

$$E\left(\hat{Y}_{JCE} - Y\right) = \sum\nolimits_{k\in U_{ab}} e_k - \sum\nolimits_{k\in U_{ab}} \mathbf{x}'_k \left(\left(\sum\nolimits_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U'} \mathbf{x}_k y_k\right) - \left(\sum\nolimits_{k\in U} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U} \mathbf{x}_k y_k\right)\right)$$

$$= \sum\nolimits_{U_{ab}} \left(y_k - \mathbf{x}'_k \left(\sum\nolimits_{k\in U'} \mathbf{x}_k \mathbf{x}'_k\right)^{-1} \left(\sum\nolimits_{k\in U'} \mathbf{x}_k y_k\right)\right)$$

$$= \sum\nolimits_{k\in U_{ab}} \left(y_k - \mathbf{x}'_k \mathrm{B}_U^{A,B}\right)$$

$$\therefore B\left(\hat{Y}_{JCE}\right) = \sum\nolimits_{k\in U_{ab}} e_k^{A,B}$$

where $e_k^{A,B} = \left(y_k - \mathbf{x}'_k \mathrm{B}_U^{A,B}\right)$

Proof of proposition 2

Under the two-way TWM misclassification, where δ_k is a sampling indicator for observation *k*,

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = E_{pq}\left(\sum_{U_{ab}} \delta_k I_k^{ab,c} d_k y_k\right) + E_{pq}\left((\theta-1)\sum_{U_a} \delta_k I_k^{c,ab} d_k y_k - \theta \sum_{U_b} \delta_k I_k^{c,ab} d_k y_k\right)$$

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = \sum_{U_{ab}} \gamma_k^{ab,c} y_k + (\theta-1)\sum_{U_a} \gamma_k^{c,ab} y_k - \theta \sum_{U_b} \gamma_k^{c,ab} y_k$$

$$\sum_{U_{ab}} \gamma_k^{ab,c} y_k = \sum_{U_{ab}} \gamma_k^{ab,c} y_k + \bar{Y}_{ab} \sum_{U_{ab}} \gamma_k^{ab,c} - \bar{Y}_{ab} \sum_{U_{ab}} \gamma_k^{ab,c}$$

$$= \left(N_{ab} \sum_{U_{ab}} \gamma_k^{ab,c} y_k + N_{ab} \bar{Y}_{ab} \sum_{U_{ab}} \gamma_k^{ab,c} - N_{ab} \bar{Y}_{ab} \sum_{U_{ab}} \gamma_k^{ab,c}\right) \Big/ N_{ab}$$

$$= \left(N_{ab} \sum_{U_{ab}} \gamma_k^{ab,c} y_k + N_{ab} \bar{Y}_{ab} \sum_{U_{ab}} \gamma_k^{ab,c}\right) \Big/ N_{ab} - N_{ab} \bar{Y}_{ab} \bar{\gamma}^{ab,c}$$

$$= N_{ab} \left(\sum_{U_{ab}} \gamma_k^{ab,c} y_k + N_{ab} \bar{Y}_{ab} \bar{\gamma}^{ab,c} - \bar{Y}_{ab} \sum_{U_{ab}} \gamma_k^{ab,c} + \bar{\gamma}^{ab,c} \sum_{U_{ab}} y_k\right) \Big/ N_{ab} - N_{ab} \bar{Y}_{ab} \bar{\gamma}^{ab,c}$$

$$= N_{ab} \sum_{U_{ab}} \left(\gamma_k^{ab,c} - \bar{\gamma}^{ab,c}\right)\left(y_k - \bar{Y}_{ab}\right) \Big/ N_{ab} + N_{ab} \bar{Y}_{ab} \bar{\gamma}^{ab,c}$$

$$= N_{ab} \left(\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right) + \bar{Y}_{ab} \bar{\gamma}^{ab,c}\right)$$

where $\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right) = \sum_{U_{ab}} \left(\gamma_k^{ab,c} - \bar{\gamma}^{ab,c}\right)\left(y_k - \bar{Y}_{ab}\right) \Big/ N_{ab}$

where $\bar{Y}_{ab} = \sum_{U_{ab}} y_k \Big/ N_{ab}$ and $\bar{\gamma}^{ab,c} = \sum_{U_{ab}} \gamma_k^{ab,c} \Big/ N_{ab}$.

Similarly

$$\sum_{U_a} \gamma_k^{c,ab} y_k = N_a \left(\varsigma_a\left(\gamma_k^{c,ab}, y_k\right) + \bar{\gamma}_a^{c,ab} \bar{Y}_a\right)$$

and

$$\sum_{U_b} \gamma_k^{c,ab} y_k = N_b \left(\varsigma_b\left(\gamma_k^{c,ab}, y_k\right) + \bar{\gamma}_b^{c,ab} \bar{Y}_b\right)$$

where

$$\varsigma_a\left(\gamma_k^{c,ab}, y_k\right) = \sum_{U_a} \left(\gamma_k^{c,ab} - \bar{\gamma}_a^{c,ab}\right)\left(y_k - \bar{Y}_a\right) \Big/ N_a$$

$$\varsigma_b\left(\gamma_k^{c,ab}, y_k\right) = \sum_{U_b} \left(\gamma_k^{c,ab} - \bar{\gamma}_b^{c,ab}\right)\left(y_k - \bar{Y}_b\right) \Big/ N_b$$

where

$\bar{Y}_a = \sum_{U_a} y_k \Big/ N_a$, $\bar{Y}_b = \sum_{U_b} y_k \Big/ N_b$, $\bar{\gamma}_a^{c,ab} = \sum_{U_a} \gamma_k^{c,ab} \Big/ N_a$ and $\bar{\gamma}_b^{c,ab} = \sum_{U_b} \gamma_k^{c,ab} \Big/ N_b$.

$$Bias_{pq}\left(\hat{Y}_{mis}\right) = N_{ab}\left(\varsigma_{ab}\left(\gamma_k^{ab,c}, y_k\right) + \bar{\gamma}^{ab,c} \bar{Y}_{ab}\right) -$$

$$(1-\theta) N_a \left(\varsigma_a\left(\gamma_k^{c,ab}, y_k\right) + \bar{\gamma}_a^{c,ab} \bar{Y}_a\right) - \theta N_b \left(\varsigma_b\left(\gamma_k^{c,ab}, y_k\right) + \bar{\gamma}_b^{c,ab} \bar{Y}_b\right)$$

REFERENCES

BANKIER, M. D., (1986). Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys, Journal of the American Statistical Association, 81, 1074−1079.

BRICK, J. M., DIPKO, S., PRESSER, S., TUCKER, C., YUAN, Y., (2006). Nonresponse Bias in a Dual-frame Sample of Cell and Landline Numbers. Public Opinion Quarterly, 70, 780−793.

BRICK, J. M., BRICK P.D., DIPKO, S., PRESSER, S., TUCKER, C., YUAN, Y., (2007). Cell Phone Survey Feasibility in the U.S.: Sampling and Calling Cell Numbers versus Landline Numbers, Public Opinion Quarterly, 71:23−39.

BRICK, J. M., FLORES-CERVANTES, I., LEE, S., NORMAN, G., (2011). Nonsampling Errors in Dual-frame Telephone Surveys, Survey Methodology, Vol. 37, No. 1, pp. 1−12.

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

FULLER, W. A., BURMEISTER, L. F., (1972). Estimators for Samples Selected from Two Overlapping Frames, Proceedings of the Social Statistics Section of the American Statistical Association, 245−249.

HARTLEY, H. O., (1962). Multiple Frame Surveys, Proceedings of the Social Statistics Section of the American Statistical Association, 203–206.

HARTLEY, H. O., (1974). Multiple Frame Methodology and Selected Applications, Sankhya, Series C, 36, 99−118.

HORVITZ, D. G., THOMPSON, D. J., (1952). A Generalization of Sampling without Replacement from a Finite Universe, Journal of the American Statistical Association, 47, 663−685.

KALTON, G., ANDERSON, D. W., (1986). Sampling Rare Populations, Journal of the Royal Statistical Society, Ser. A 149, 65−82.

KENNEDY, C., (2007). Evaluating the Effects of Screening for Telephone Service in Dual-frame RDD Surveys. Public Opinion Quarterly 70:750–771.

LEPKOWSKI, J. M., (1991). Sampling the Difficult to Sample. Journal of Nutrition, 121, 416−423.

LEPKOWSKI, J. M., GROVES, R. M., (1986). A Mean Squared Error Model for Multiple Frame, Mixed Mode Survey Design. Journal of the American Statistical Association, 81, 930−937.

LINK, M. W., BATTAGLIA, M. P., FRANKEL, M. R., OSBORN, L., MOKDAD, A. H., (2006). Address-Based Versus Random-Digit Dialed Surveys: Comparison of Key Health and Risk Indicators, American Journal of Epidemiology, 164:1019−25.

LINK, M. W., BATTAGLIA, M. P., FRANKEL, M.R., OSBORN, L., MOKDAD, A. H., (2007). Reaching The U.S. Cell Phone Generation: Comparison of Cell Phone Survey Results With an Ongoing Landline Telephone Survey, Public Opinion Quarterly 71:814−839.

LINK, M. W., BATTAGLIA, M. P., FRANKEL, M. R., OSBORN, L., MOKDAD, A. H., (2008). A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys, Public Opinion Quarterly, 72, 6−27.

LINK, M. W., LAI, J. (2011). Cell Phone-Only Households and Problems of Differential Nonresponse Using an Address Based Sampling Design, Public Opinion Quarterly, 75(4), 613−635.

LOHR, S., (2011). Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames, Survey Methodology, 37, 197−213.

LOHR, S. L., RAO, J. N. K., (2000). Inference in Dual Frame Surveys, Journal of the American Statistical Association, 95, 271−280.

LOHR, S. L., RAO, J. N. K., (2006). Estimation in Multiple-Frame Surveys, Journal of the American Statistical Association, 101, 1019−1030.

MECATTI, F., (2007). A Single Frame Multiplicity Estimator for Multiple Frame Surveys, Survey Methodology, 33, 151−157.

RAO, J. N. K., WU, C., (2010). Pseudo-Empirical Likelihood Inference for Dual Frame Surveys, Journal of the American Statistical Association, 105, 1494−1503.

SÄRNDAL, C. E., SWENSSON, B., WRETMAN, J., (1992). Model-assisted Survey Sampling, New York: Springer-Verlag.

SINGH, A. C., MOHL, C. A., (1996). Understanding Calibration Estimators in Survey Sampling, Survey Methodology, 22, 107-115.

SKINNER, C. J., (1991). On the Efficiency of Raking Ratio Estimation for Multiple Frame Surveys, Journal of the American Statistical Association, 86, 779−784.

SKINNER, C. J., RAO, J. N. K., (1996). Estimation in Dual-Frame Surveys with Complex Designs, Journal of the American Statistical Association, 91, 349−356.

STUKEL, D. M., HIDIROGLOU, M. A., SÄRNDAL, C. E., (1996). Variance Estimation for Calibration Estimators: A Comparison of Jackknifing Versus Taylor Linearization, Survey Methodology, 22, 117−125.

SUDMAN, S., KALTON, G., (1986). New developments in the sampling of special populations. Annual Review of Sociology, 12, 401−429.

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 37–52

SOME CLASSES OF MODIFIED RATIO TYPE ESTIMATORS IN SAMPLE SURVEYS

A. K. P. C. Swain[1], **Manjula Das**[2]

ABSTRACT

In this paper some classes of modified ratio type estimators with additive and multiplicative adjustments made to the simple mean per unit estimator and classical ratio estimator are suggested to obtain more efficient ratio type estimators compared to the classical one. Their biases and mean square errors are obtained and compared with first order approximations.

Key words: ratio type estimator, simple random sampling, bias, mean square error, efficiency.

1. Introduction

In sample surveys it is usual practice to look for information on auxiliary variables which are either available from official records or can be collected inexpensively in the course of investigation. In the case of single auxiliary variable the ratio estimator and the regression estimator are two classical estimators making use of the auxiliary information to improve the efficiency of the finite population parameters such as population mean, total, variance, etc. Although simple to compute, the ratio estimator is always less efficient than the linear regression estimator in large samples.

But the theory of linear regression is not very much appropriate for the sample survey situations(Cochran,1953) and requires that the assumptions such as:
(a) existence of linearity of regression of y on x in the population;

(b) constancy of residual variance;

(c) infinite nature of population;

should be approximately satisfied, but are rarely satisfied in finite population sampling.

[1] Former Professor of Statistics, Utkal University, Bhubaneswar-751004, India.
E-mail:akpcs@rediffmail.com.

[2] Associate Professor, Department of Mathematics, ITER, SOA University, Bhubaneswar-751030, India.

This has motivated some research workers to look for different techniques to form ratio type estimators whose mean square errors approximate to that of the approximate mean square error of the linear regression estimate in large samples.Srivastava(1967) modified the ratio estimator with power transformation of the ratio of the population mean to the sample mean whose minimum mean square error to first approximation equals to that of the linear regression estimator.Srivastava(1971) proposed a class of estimators having minimum mean square error equal to that of the linear regression estimator,provided certain regularity conditions are satisfied. In this paper we make a variety of additive and multiplicative adjustments to the simple mean per unit estimator and classical ratio estimator so that their large sample mean square errors attain the minimum mean square bound of Srivastava's class of estimators, which is in fact the large sample mean square error of the linear regression estimator. The proposed classes of estimators are compared as regards their large sample biases.

Let $U=(U_1,U_2,......,U_N)$ be a finite population of N distinct and identifiable units. Let y and x denote the study variable and auxiliary variable respectively taking paired values (Y_i,X_i) on the unit $U_i(i=1,2,...,N)$ .Assume x to be positively correlated with y. Further, assume that y and x are positively measured.

Define $\bar{Y}=\frac{1}{N}\sum_{i=1}^{N}Y_i$, $\bar{X}=\frac{1}{N}\sum_{i=1}^{N}X_i$, $R=\frac{\bar{Y}}{\bar{X}}$

$$S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})^2 \quad ,S_x^2=\frac{1}{N-1}\sum_{i=1}^{N}(X_i-\bar{X})^2 \ ,$$

$$S_{yx}=\frac{1}{N-1}\sum_{i=1}^{N}(Y_i-\bar{Y})(X_i-\bar{X}),C_y^2=\frac{S_y^2}{\bar{Y}^2} \ ,C_x^2=\frac{S_x^2}{\bar{X}^2} \text{ and}$$

$C_{yx}=\frac{S_{yx}}{\bar{Y}\bar{X}}=\rho C_yC_x,\rho$ being the correlation coefficient between y and x.The population regression coefficient of y on x is defined as $\beta=\frac{S_{yx}}{S_x^2}$.

Let $u_1,u_2,....,u_n$ be a simple random sample s of size n units drawn without replacement from U .We observe paired values $(y_i,x_i),i=1,2,...,n$ on the sampled units.

Define $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$,

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2, s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

$s_{yx} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})$ and $r = \frac{\overline{y}}{\overline{x}}$.The sample regression coefficient of y on x is defined as $b = \frac{s_{yx}}{s_x^2}$.

To estimate the population mean $\overline{Y}$ of the study variable y,the classical ratio estimator $\hat{\overline{Y}}_R$ is defined by

$$\hat{\overline{Y}}_R = \overline{y}(\frac{\overline{X}}{\overline{x}}), \tag{1.1}$$

where $\overline{X}$ is assumed to be known in advance.

It is well known (Cochran,1953)that $\hat{\overline{Y}}_R$ is a biased estimate of the population mean $\overline{Y}$ with bias to $O(1/n)$ given by

$$Bias(\hat{\overline{Y}}_R) = \theta\overline{Y}(C_x^2 - \rho C_y C_x) = -\theta\overline{Y}(\frac{\beta}{R} - 1)C_x^2$$

$$= -\theta\overline{Y}(K-1)C_x^2 \text{ ,where } K = \frac{\beta}{R} \tag{1.2}$$

$\theta = (\frac{1}{n} - \frac{1}{N}), C_y$ and C_x being the coefficients of variation of y and x respectively,.

Further, up to terms of $O(1/n)$, the mean square error of $\hat{\overline{Y}}_R$ is given by

$$MSE(\hat{\overline{Y}}_R) = \theta\overline{Y}^2(C_y^2 + C_x^2 - 2\rho C_y C_x), \tag{1.3}$$

$\hat{\overline{Y}}_R$ is more efficient than $\overline{y}$

$$\text{if} \quad \rho > \frac{1}{2}(\frac{C_x}{C_y}) \tag{1.4}$$

Besides ratio method of estimation, linear regression method of estimation is another early method initiated by Watson (1937), making use of auxiliary information in sample surveys. The simple regression estimate of the population mean $\overline{Y}$ is given by

$$\overline{y}_{lr} = \overline{y} + b(\overline{X} - \overline{x}), \tag{1.5}$$

where b is the linear regression coefficient of y on x, calculated from the sample.

The mean square error of $\overline{y}_{lr}$ to $O(1/n)$
is given by

$$MSE(\overline{y}_{lr}) = \theta \overline{Y}^2 C_y^2 (1-\rho^2) \tag{1.6}$$

Both ratio and regression estimators are biased estimators and biases decrease with increase in the sample size. Comparing (1.3) and (1.6) it may be seen that in large samples $\hat{\overline{Y}}_R$ is always less efficient than $\overline{y}_{lr}$ unless the regression line of y on x passes through the origin, in which case they have equal efficiency.

Srivastava (1967) suggested a class of power transformation estimator

$$\hat{\overline{Y}}_{SR} = \overline{y}(\frac{\overline{X}}{\overline{x}})^{\alpha}, \tag{1.7}$$

where α is a real number to be suitably chosen. The optimum value of $MSE(\hat{\overline{Y}}_{SR})$,when optimized with respect to α, gives the expression given in (1.6).

Walsh (1970) suggested an alternative class of ratio-type estimator where x_i is transformed to z_i such that

$z_i = \alpha x_i + (1-\alpha)\overline{X}$

Hence, $\overline{z} = \alpha\overline{x} + (1-\alpha)\overline{X}$ and $\overline{Z} = \overline{X}$.

As such, modified ratio type estimator is formed as

$$\hat{\overline{Y}}_{WR} = \frac{\overline{y}}{\overline{z}}\overline{Z} = \frac{\overline{y}}{\alpha\overline{x} + (1-\alpha)\overline{X}}\overline{X} \tag{1.8}$$

To first order approximations of the optimum mean square error of $\hat{\overline{Y}}_{WR}$ is given by

$MSE(\hat{\overline{Y}}_{WR}) = \theta\overline{Y}^2 C_y^2 (1-\rho^2)$, as given in (1.6).

Srivastava (1971) proposed a generalized class of estimators given by

$$t_g = \overline{y}H(u) \tag{1.9}$$

where $u = \overline{x} / \overline{X}$ and $H(.)$ is a parametric function satisfying certain regularity conditions as given in Srivastava (1971), such as

(i) $H(1) = 1$

(ii) The first and second order derivatives of H with respect to u exist and are known constants at a given point $u=1$.

He also showed that the asymptotic mean square error of t_g cannot be reduced further than $\min.MSE(t_g)=\theta\bar{Y}^2C_y^2(1-\rho^2)$, which is the approximate mean square error of the linear regression estimator, which is the lower bound to mean square error of class of estimators t_g .Prabhu-Ajgaonkar(1993) has noted that an optimum estimator does not exist uniformly in the class t_g .

Srivastava (1980) defined another wider class of estimators as

$$t_w = H(\bar{y},u) \tag{1.10}$$

where $H(\bar{y},u)$ is a function of $\bar{y}$ and u, satisfying certain regularity conditions specified by him. He showed that asymptotic minimum mean square error of t_w cannot be reduced further than that given in (1.6).

Thus, ratio estimator $\hat{\bar{Y}}_R$, Srivastava's power transformation estimator $\hat{\bar{Y}}_{SR}$ and Walsh's estimator $\hat{\bar{Y}}_{WR}$ are the special cases of t_g .The wider class t_w includes regression estimator besides ratio estimator, power transformation estimator and many others.

Swain (2013) proposed a class of estimators

$$\hat{\bar{Y}}_{SWR} = \bar{y}\left[\alpha\left(\frac{\bar{X}}{\bar{x}}\right)^g + (1-\alpha)\left(\frac{\bar{x}}{\bar{X}}\right)^h\right]^\delta , \tag{1.11}$$

which is also a subclass of t_g and where α, g, h and δ are free real constants to be suitably chosen and also the asymptotic mean square error $\hat{\bar{Y}}_{SWR}$ is equal to asymptotic mean square error of the linear regression estimator given in (1.6). Both $\hat{\bar{Y}}_{SR}$ and $\hat{\bar{Y}}_{WR}$ are the special cases of $\hat{\bar{Y}}_{SWR}$, which can be further generalized as

$$\hat{\bar{Y}}^*_{SWR} = \bar{y}\left[\alpha\left(\frac{A\bar{X}+B}{A\bar{x}+B}\right)^g + (1-\alpha)\left(\frac{A\bar{x}+B}{A\bar{X}+B}\right)^h\right]^\delta ,$$

$$= \bar{y}\left[\alpha\left(\frac{\bar{X}+d}{\bar{x}+d}\right)^g + (1-\alpha)\left(\frac{\bar{x}+d}{\bar{X}+d}\right)^h\right]^\delta , \tag{1.12}$$

where $d = B/A$ and α, g, h, d and δ are free parameters to be suitably chosen.

We may arbitrarily specify any four of the aforesaid parameters and minimize the approximate mean square error with respect to the remaining one and the resulting mean square error equals the approximate mean square error of the linear regression estimator which is the lower bound to the mean square error of the class of estimators defined by t_g . To choose best estimator in this class the survey practitioner should select those set values for the unspecified parameters for which the first order bias is zero or approximately so.

In the following some adjustments are made to $\bar{y}$ and $\hat{\bar{Y}}_R$ to construct some classes of modified ratio type estimators to provide more efficient estimators of the population mean $\bar{Y}$, and the proposed classes of estimators, which are sub-classes of Srivastava’s (1971,1980) classes of estimators,are compared as regards their biases and mean square errors.

2. Proposed classes of estimators

Consider the following classes of estimators, where $H(u)$ is as defined by Srivastava (1971).

$$T_1 = \bar{y}\left[H(u)\right]^{\mu}$$

$$T_2 = \bar{y}\left[\frac{1}{\mu H(u) + (1-\mu)}\right]$$

$$T_3 = \hat{\bar{Y}}_R\left[H(u)\right]^{\mu}$$

$$T_4 = \frac{\hat{\bar{Y}}_R}{\mu H(u) + (1-\mu)}$$

$$T_5 = \bar{y} + \mu(1 - H(u))$$

$$T_6 = \hat{\bar{Y}}_R + \mu(1 - H(u))$$

Expanding $H(u)$ by the value 1 in the second order Taylor’s series we have

$$H(u) = H\left[1 + (u-1)\right] = H(1) + (u-1)(\frac{\partial H}{\partial u})_{u=1} + \frac{1}{2}(u-1)^2(\frac{\partial^2 H}{\partial u^2})_{u=1} + \ldots\ldots \quad (2.1)$$

Assuming $|u-1| < 1$, the higher order terms can be neglected and we write

$$T_1 = \bar{y}\left[1 + (u-1)H_1 + (u-1)^2 H_2 + \ldots.\right]^{\mu}, \quad (2.2)$$

where $H_1 = (\frac{\partial H}{\partial u})_{u=1}$ and $H_2 = \frac{1}{2}(\frac{\partial^2 H}{\partial u^2})_{u=1}$ denote the first and second order partial derivatives of H with respect to u and are the known constants.

Thus, we write

$$T_1 = \bar{Y}(1+e_0)(1+e_1H_1+e_1^2H_2+......)^{\mu}\,, \tag{2.3}$$

where $e_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}}$ and $e_1 = u-1$.

Expanding (2.3) in power series we have

$$T_1 = \bar{Y}(1+e_0)\left[1+\mu(H_1e_1+H_2e_1^2)+\frac{\mu(\mu-1)}{2}(H_1e_1+H_2e_1^2)^2+......\right].$$

To first order of approximations

$$MSE(T_1) = \theta\bar{Y}^2(C_y^2+\mu^2H_1^2C_x^2+2\mu H_1C_{yx}) \tag{2.4}$$

$$Bias(T_1) = B(T_1) = \theta\bar{Y}(\mu H_1C_{yx}+\mu H_2C_x^2+\frac{\mu(\mu-1)}{2}H_1^2C_x^2) \tag{2.5}$$

where $E(e_1^2) = \theta C_x^2$ and $E(e_0e_1) = \theta C_{yx}$.

Differentiating $MSE(T_1)$ with respect to μ and putting it equal to zero, we have optimum μ given by

$$\mu_{opt} = -\frac{K}{H_1}\,.$$

Thus, the optimum mean square error of T_1 obtained by substituting the optimum value of H_1 in (2.4) is given by

$$MSE(T_1) = \theta\bar{Y}^2C_y^2(1-\rho^2)\,, \tag{2.6}$$

which is equal to that of the large sample mean square error of the linear regression estimator given by (1.6).

Also, the bias of T_1 with optimum μ is given by

$$\text{Bias }(T_1) = \theta\bar{Y}K\left[-K-(\frac{H_2}{H_1})+\frac{(K+H_1)}{2}\right]C_x^2 \tag{2.7}$$

Proceeding as before we find to $O(1/n)$

$$MSE(T_2) = MSE(T_3) = MSE(T_4) = MSE(T_5) = MSE(T_6) = \theta\bar{Y}^2 C_y^2 (1-\rho^2)$$

$$Bias(T_2) = -\theta\bar{Y}K\frac{H_2}{H_1}C_x^2$$

$$Bias(T_3) = \theta\bar{Y}(1-K)\left[\ \left(\frac{H_2}{H_1} - \frac{H_1}{2}\right) + \frac{(1+K)}{2}\ \right]C_x^2$$

$$Bias(T_4) = \theta(1-K)\left(\frac{H_2}{H_1} + 1\right)C_x^2$$

$$Bias(T_5) = -\theta\bar{Y}K\left(\frac{H_2}{H_1}\right)C_x^2$$

$$Bias(T_6) = \theta\bar{Y}K\left(H_1 - \frac{H_2}{H_1}\right)C_x^2$$

The biases and mean square errors of different classes of estimators are summarized in Table 1.

Table 1. Biases and mean square errors

Classes of Estimators	μ_{opt}	Bias	MSE_{opt}
$T_1 = \bar{y}\left[H(u)\right]^{\mu}$	$-(\frac{K}{H_1})$	$\theta\bar{Y}K\left[-K - \frac{H_2}{H_1} + \frac{K+H_1}{2}\right]C_x^2$	$\theta\bar{Y}^2 C_y^2 (1-\rho^2)$
$T_2 = \bar{y}\left[\frac{1}{\mu H(u) + (1-\mu)}\right]$	$\frac{K}{H_1}$	$-\theta\bar{Y}K\left[\frac{H_2}{H_1}\right]C_x^2$	$\theta\bar{Y}^2 C_y^2 (1-\rho^2)$
$T_3 = \hat{\bar{Y}}_R\left[H(u)\right]^{\mu}$	$\frac{1-K}{H_1}$	$\theta\bar{Y}(1-K)(\frac{H_2}{H_1} - \frac{H_1}{2} + \frac{1+K}{2})C_x^2$	$\theta\bar{Y}^2 C_y^2 (1-\rho^2)$
$T_4 = \left[\frac{\hat{\bar{Y}}_R}{\mu H(u) + (1-\mu)}\right]$	$\frac{K-1}{H_1}$	$\theta\bar{Y}(1-K)\left[\frac{H_2}{H_1} + 1\right]C_x^2$	$\theta\bar{Y}^2 C_y^2 (1-\rho^2)$
$T_5 = \bar{y} + \mu\left[1 - H(u)\right]$	$\frac{K\bar{Y}}{H_1}$	$-\theta\bar{Y}K\left[\frac{H_2}{H_1}\right]C_x^2$	$\theta\bar{Y}^2 C_y^2 (1-\rho^2)$
$T_6 = \hat{\bar{Y}}_R + \mu\left[1 - H(u)\right]$	$\frac{(K-1}{H_1}$	$\theta\bar{Y}K\left[H_1 - \frac{H_2}{H_1}\right]C_x^2$	$\theta\bar{Y}^2 C_y^2 (1-\rho^2)$

3. Some special cases of proposed classes of estimators

By defining $H(u)$ differently we may generate different classes of estimators and some of them related to ratio and product estimators are given in Table 2.

Table 2. Estimators and their Biases excluding the common multiplier

Class of estimators	$H(u)=\bar{x}/\bar{X}$	*Bias*	$H(u)=\bar{X}/\bar{x}$	*Bias*
$T_1=\bar{y}[H(u)]^{\mu}$	$T_{11}=\bar{y}\left[\bar{x}/\bar{X}\right]^{\mu}$	$\frac{K(K-1)}{2}$	$T_{12}=\bar{y}\left[\bar{X}/\bar{x}\right]^{\mu}$	$\frac{K(K-1)}{2}$
$T_2=\bar{y}\left[\frac{1}{\mu H(u)+(1-\mu)}\right]$	$T_{21}=\bar{y}\left[\frac{1}{\mu(\bar{x}/\bar{X})+(1-\mu)}\right]$	0	$T_{22}=\bar{y}\left[\frac{1}{\mu(\bar{X}/\bar{x})+(1-\mu)}\right]$	K
$T_3=\hat{\bar{Y}}_R[H(u)]^{\mu}$	$T_{31}=\hat{\bar{Y}}_R\left[\bar{x}/\bar{X}\right]^{\mu}$	$\frac{K(K-1)}{2}$	$T_{32}=\hat{\bar{Y}}_R\left[\bar{X}/\bar{x}\right]^{\mu}$	$\frac{K(K-1)}{2}$
$T_4=\left[\frac{\hat{\bar{Y}}_R}{\mu H(u)+(1-\mu)}\right]$	$T_{41}=\left[\frac{\hat{\bar{Y}}_R}{\mu(\bar{x}/\bar{X})+(1-\mu)}\right]$	0	$T_{42}=\left[\frac{\hat{\bar{Y}}_R}{\mu(\bar{X}/\bar{x})+(1-\mu)}\right]$	$1-K$
$T_5=\bar{y}+\mu[1-H(u)]$	$T_{51}=\bar{y}+\mu\left[1-(\bar{x}/\bar{X})\right]$	0	$T_{52}=\bar{y}+\mu\left[1-(\bar{X}/\bar{x})\right]$	K
$T_6=\hat{\bar{Y}}_R+\mu[1-H(u)]$	$T_{61}=\hat{\bar{Y}}_R+\mu\left[1-(\bar{x}/\bar{X})\right]$	K	$T_{62}=\hat{\bar{Y}}_R+\mu\left[1-(\bar{X}/\bar{x})\right]$	0

We find (Table 2) the first order biases of T_{21}, T_{41}, T_{51} and T_{62} vanish, having the same approximate mean square error as that of the linear regression estimator. Since the optimum value of μ is a usually unknown parametric function $K=\frac{\beta}{R}$, we estimate it by its consistent estimator $\hat{K}=\frac{b}{r}$ from the sample.

Thus, the estimators T_{21}, T_{41}, T_{51} and T_{62} with estimated values of K are given by

$$\hat{T}_{21}=\frac{\bar{y}}{(\frac{b}{r})(\bar{x}/\bar{X})+(1-\frac{b}{r})}$$

$$\hat{T}_{41} = \frac{\hat{\bar{Y}}_R}{\left[(\frac{b-r}{r})(\bar{x}/\bar{X}) + (2-\frac{b}{r}) \right]}$$

$$\hat{T}_{51} = \bar{y} + b(\bar{X} - \bar{x})$$

$$\hat{T}_{62} = \hat{\bar{Y}}_R + \bar{y}(1-\frac{b}{r})(1-\frac{\bar{X}}{\bar{x}}) = \bar{X}\left[r + (r-b)(1-\frac{\bar{X}}{\bar{x}}) \right]$$

3.1. Bias and mean square error of $\hat{T}_{21}, \hat{T}_{41}, \hat{T}_{51}$ and $\hat{T}_{62}$

(i)
$$\hat{T}_{21} = \frac{\bar{y}}{\frac{b}{r}(\bar{x}/\bar{X}) + (1-\frac{b}{r})}$$

Or alternatively,
$$\hat{T}_{21} = \frac{\bar{y}}{\bar{x}}\bar{X}\left[\frac{r}{b + (r-b)(\bar{X}/\bar{x})} \right] \tag{3.1}$$

Define

$$e_0 = \frac{\bar{y}-\bar{Y}}{\bar{Y}}, e_1 = \frac{\bar{x}-\bar{X}}{\bar{X}}, e_2 = \frac{s_{yx}-S_{yx}}{S_{yx}}, e_3 = \frac{s_x^2 - S_x^2}{S_x^2}$$

Expanding $\hat{T}_{21}$ using binomial series expansion with assumptions $|e_1| < 1$ and $|e_3| < 1$ for all possible samples and keeping terms up to second degree we have

$$\hat{T}_{21} = \bar{Y}\left[1 + (e_0 - \frac{\beta}{R}e_1) + \frac{\beta}{R}(e_1e_3 - e_1e_2) + \frac{\beta}{R}(\frac{\beta}{R}-1)e_1^2 \right] + \ldots$$

To first order approximations, that is to $O(1/n)$,

$$Bias(\hat{T}_{21}) = \bar{Y}\left[(\frac{\beta}{R}-1)(\frac{\beta}{R})\frac{V(\bar{x})}{\bar{X}^2} + \frac{\beta}{R}(\frac{Cov(s_x^2,\bar{x})}{S_x^2\bar{X}} - \frac{Cov(s_{yx},\bar{x})}{S_{yx}\bar{X}}) \right] \tag{3.2}$$

where $Cov(s_x^2, \bar{x}) = \frac{N(N-n)}{(N-1)(N-2)}\frac{\mu_{03}}{n}$ and

$$Cov(s_{yx}, \bar{x}) = \frac{N(N-n)}{(N-1)(N-2)}\frac{\mu_{12}}{n}$$

with $\mu_{rs} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{Y})^r (x_i - \overline{X})^s, i = 1,2,3,4,....$ (see Sukhatme, Sukhatme and Asok,1984)

$$MSE(\hat{T}_{21}) = \theta \overline{Y}^2 C_y^2 (1-\rho^2) \tag{3.3}$$

Under bivariate normality of (y, x) or for symmetrical populations,

$$Bias(\hat{T}_{21}) = \frac{1}{n}\overline{Y}\left[(\frac{\beta}{R})(\frac{\beta}{R}-1)\right]C_x^2 \tag{3.4}$$

(ii) $$\hat{T}_{41} = \frac{\hat{\hat{Y}}_R}{(\frac{b}{r}-1)(\overline{x}/\overline{X}) + (2-\frac{b}{r})} = \frac{\overline{y}}{(\frac{b}{r}-1)(\overline{x}/\overline{X})^2 + (2-\frac{b}{r})(\overline{x}/\overline{X})} \tag{3.5}$$

On expansion $\hat{T}_{41} = \overline{Y} + \overline{Y}\left[(e_0 - \frac{\beta}{R}e_1) + \frac{\beta}{R}(e_3 e_1 - e_2 e_1) + (\frac{\beta}{R}-1)^2 e_1^2\right] + ...$

$$Bias(\hat{T}_{41}) = \theta\overline{Y}(\frac{\beta}{R}-1)^2 C_x^2 + \overline{Y}\frac{\beta}{R}\left[\frac{Cov(s_x^2, \overline{x})}{S_x^2 \overline{X}} - \frac{Cov(s_{yx}, \overline{x})}{S_{yx}\overline{X}}\right] \tag{3.6}$$

$$MSE(\hat{T}_{41}) = \theta\overline{Y}^2 C_y^2 (1-\rho^2) \tag{3.7}$$

Under bivariate normal populations or for symmetrical populations

$$Bias(\hat{T}_{41}) = \frac{1}{n}\overline{Y}(\frac{\beta}{R}-1)^2 C_x^2 \tag{3.8}$$

(iii) $$\hat{T}_{51}) = \overline{y} + b(\overline{X} - \overline{x}) \tag{3.9}$$

Expanding $\hat{T}_{51}$ using Binomial series with assumptions $|e_1| < 1$ and $|e_3| < 1$ we have

$$\hat{T}_{51} = \overline{Y} + \overline{Y}\left[(e_0 - \frac{\beta}{R}e_1) + \frac{\beta}{R}(e_3 e_1 - e_2 e_1)\right] + ...$$

Thus, $Bias(\hat{T}_{51}) = \overline{Y}\left[\frac{\beta}{R}(\frac{Cov(s_x^2, \overline{x})}{S_x^2 \overline{X}} - \frac{Cov(s_{yx}, \overline{x})}{S_{yx}\overline{X}})\right]$ (Sukhatme et al., 1984)

(3.10)

$$\text{and } MSE(\hat{T}_{51}) = \theta \bar{Y}^2 C_y^2 (1-\rho^2) \qquad (3.11)$$

Under Bivariate normality $Bias(\hat{T}_{51})$ vanishes.

(iv)
$$\hat{T}_{62} = \hat{\bar{Y}}_R + \hat{\mu}_{opt}(1-\frac{\vec{X}}{\bar{x}}) = \bar{X}\left[r + (r-b)(1-\frac{\bar{X}}{\bar{x}}) \right] \qquad (3.12)$$

Expanding $\hat{T}_{62}$ using binomial series expansion with assumptions $|e_1| < 1$ and $|e_3| < 1$ for all possible samples and keeping terms up to second degree we have

$$\hat{T}_{62} = \bar{Y} + \bar{Y}\left[(e_0 - \frac{\beta}{R}e_1) + \frac{\beta}{R}(e_1e_3 - e_1e_2) + (\frac{\beta}{R}-1)e_1^2 \right]$$

$$E(\hat{T}_{62}) = \bar{Y} + \bar{Y}\left[(\frac{\beta}{R}-1)\frac{V(\bar{x})}{\bar{X}^2} + \frac{\beta}{R}(\frac{Cov(s_x^2,\bar{x})}{S_x^2\bar{X}} - \frac{Cov(s_{yx},\bar{x})}{S_{yx}\bar{X}}) \right]$$

$$Bias(\hat{T}_{62}) = \bar{Y}\left[(\frac{\beta}{R}-1)\frac{V(\bar{x})}{\bar{X}^2} + \frac{\beta}{R}(\frac{Cov(s_x^2,\bar{x})}{S_x^2\bar{X}} - \frac{Cov(s_{yx},\bar{x})}{S_{yx}\bar{X}}) \right] \qquad (3.13)$$

$$MSE(\hat{T}_{62}) = \theta \bar{Y}^2 C_y^2 (1-\rho^2) \qquad (3.14)$$

Under bivariate normality of (y, x) or for symmetrical populations

$Cov(s_x^2, \bar{x})$ and $Cov(s_{yx}, \bar{x})$ vanish and thus to $O(1/n)$

$$Bias(\hat{T}_{62}) = \frac{1}{n}\bar{Y}(\frac{\beta}{R}-1)C_x^2 \qquad (3.15)$$

3.2. Comparison of biases and mean square errors of $\hat{T}_{21}, \hat{T}_{41}, \hat{T}_{51}$ and $\hat{T}_{62}$

To first order of approximations, that is to $O(1/n)$

the mean square errors of $\hat{T}_{21}, \hat{T}_{41}, \hat{T}_{51}$ and $\hat{T}_{62}$ are equal to that of the linear regression estimator. Further,

(i) $\hat{T}_{21}$ is less biased than the regression estimator $\hat{T}_{51}$ if $|A+B| < |A|$

(ii) $\hat{T}_{41}$ is less biased than the regression estimator $\hat{T}_{51}$ if $|A+C| < |A|$

(iii) $\hat{T}_{62}$ is less biased than the regression estimator $\hat{T}_{51}$ if $|A+D| < |A|$

where $$A=\bar{Y}\left[\frac{\beta}{R}\left(\frac{Cov(s_x^2,\bar{x})}{S_x^2\bar{X}}-\frac{Cov(s_{yx},\bar{x})}{S_{yx}\bar{X}}\right)\right]$$

$$B=\theta\bar{Y}\left[\frac{\beta}{R}\left(\frac{\beta}{R}-1\right)\right]C_x^2$$

$$C=\theta\bar{Y}\left(\frac{\beta}{R}-1\right)^2C_x^2$$

$$D=\theta\bar{Y}\left(\frac{\beta}{R}-1\right)C_x^2$$

Under bivariate normality or for symmetrical populations

$$Bias(\hat{T}_{21})=\frac{1}{n}\bar{Y}\left(\frac{\beta}{R}\right)\left(\frac{\beta}{R}-1\right)C_x^2$$

$$Bias(\hat{T}_{41})=\frac{1}{n}\bar{Y}\left(\frac{\beta}{R}-1\right)^2C_x^2$$

$$Bias(\hat{T}_{51})=0$$

$$Bias(\hat{T}_{62})=\frac{1}{n}\bar{Y}\left(\frac{\beta}{R}-1\right)C_x^2$$

$$Bias(\hat{Y}_R)=-\frac{1}{n}\bar{Y}\left(\frac{\beta}{R}-1\right)C_x^2$$

$\hat{T}_{21}$ is less biased than $\hat{T}_{41}$, if $\frac{\beta}{R}<1/2$, and less biased than $\hat{T}_{62}$,if $\frac{\beta}{R}<1$.

Thus, $\hat{T}_{21}$ is less biased than both $\hat{T}_{41}$ and $\hat{T}_{62}$,if $\frac{\beta}{R}<1/2$.

$\hat{T}_{41}$ is less biased than $\hat{T}_{62}$ if $\left(\frac{\beta}{R}-1\right)^2<1$

Further , $\left|Bias\hat{\bar{Y}}_R\right|=\left|Bias\hat{T}_{62}\right|$

4. Numerical illustration

To estimate the total number of milch animals in 117 villages of zone 4 of Haryana state of India in 1977-78 a simple random sample of size 17 was selected. The number of milch animals in the survey (y) and the number of milch animals in the previous census (x) were observed for each village in the sample

(Singh and Chaudhary, 1986). The estimated values of approximate bias except the common multiplier are given in Table 3.

Table 3. Biases of estimators

Estimator	Absolute Bias excepting common multiplier
$\hat{T}_{21}$	0.01143
$\hat{T}_{41}$	0.01155
$\hat{T}_{51}$	0.01150
$\hat{T}_{62}$	0.01137

Comment: $\hat{T}_{62}$ is least biased among the competitors.

5. Conclusions

(i) Without assuming restrictive assumptions associated with the linear regression estimator ,the proposed modified ratio-type estimators $\hat{T}_{21}, \hat{T}_{41}$ and $\hat{T}_{62}$ are asymptotically as efficient as the linear regression estimator $\hat{T}_{51}(\overline{y}_{lr})$.

(ii) Under bivariate normality the first order bias of $\hat{T}_{51}$ is zero. Further, $\hat{T}_{21}$ is less biased than both $\hat{T}_{41}$ and $\hat{T}_{62}$ if $\frac{\beta}{R} < 1/2$, and $\hat{T}_{41}$ is less biased than $\hat{T}_{62}$ if $(\frac{\beta}{R} - 1)^2 < 1$

Further, $\hat{\overline{Y}}_R$ and $\hat{T}_{62}$ have same absolute bias

(iii) Numerical illustration shows that up to first order of approximations $\hat{T}_{62}$ is less biased than $\hat{T}_{21}, \hat{T}_{41}$ and $\hat{T}_{51}$ although the differences are marginal.

(iv) $H(u)$ may also be defined as exponential functions of u such as

$H(u) = Exp\left[\alpha(1-u)\right]$,where α is a real constant,

$H(u) = Exp\left[\dfrac{1-u}{1+u}\right]$

$H(u) = a^{1-u}$,where a is a non-zero positive real constant,etc.

Acknowledgement

The authors thank the esteemed referees for their helpful comments which helped improving the contents and presentation of results in the paper.

REFERENCES

COCHRAN, W. G., (1953). Sampling Techniques, Wiley, New York.

PRABHU-AJGAONKAR, S. G., (1993). Non-existence of an optimum estimator in a class of ratio estimators,Statistical Papers, 34, 1, 161−165.

SINGH, D., CHAUDHARY, F. S., (1986). Theory and Analysis of Sample Survey Designs, New Age International (P)Ltd. Publishers, New Delhi.

SRIVASTAVA, S. K., (1967). An estimator using auxiliary information in sample surveys, Bull. Cal. Stat Assoc., 16, 121−132.

SRIVASTAVA, S. K., (1971). Generalized estimator for mean of a finite population using multi auxiliary information. Journ. Amer. Stat. Assoc., 66, 404−407.

SRIVASTAVA, S. K., (1980). A class of estimator using auxiliary information in sample surveys. Canadian Journal of Statistics, 8, 253−254.

SUKHATME, P. V., SUKHATME, B. V., ASOK, C., (1984). Sampling theory of surveys with Applications, Asia Publishing House, New Delhi.

SWAIN, A. K. P. C., (2013). On some modified ratio and product type estimators-Revisited, Investigacion Operacional, Vol. 34, 1, 35−57.

WALSH, J. E., (1970). Generalization of ratio estimate of population total, Sankhya, A., 32, 141−46.

WATSON, D. J., (1937). The estimation of leaf areas, Journ. Ag. Sc., 27, 474.

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 53–64

IMPROVED SEPARATE RATIO AND PRODUCT EXPONENTIAL TYPE ESTIMATORS IN THE CASE OF POST-STRATIFICATION

Hilal A. Lone[1], **Rajesh Tailor**[2]

ABSTRACT

This paper addressed the problem of estimation of finite population mean in the case of post-stratification. Improved separate ratio and product exponential type estimators in the case of post-stratification are suggested. The biases and mean squared errors of the suggested estimators are obtained up to the first degree of approximation. Theoretical and empirical studies have been done to demonstrate better efficiencies of the suggested estimators than other considered estimators.

Key words: finite population mean, post-stratification, bias, mean squared error.

1. Introduction

The problem of post-stratification was first discussed by Hansen et al. (1953). Ige and Tripathi (1989) studied the properties of classical ratio and product estimators of population mean in the case of post-stratification. Chouhan (2012) studied the Bahl and Tuteja (1991) estimators in the case of post-stratification. Many researchers including Kish (1965), Fuller (1966), Raj (1972), Holt and Smith (1979), Agrawal and Pandey (1993), Lone and Tailor (2014), Jatwa (2014), Lone and Tailor (2015), Tailor et al. (2015) contributed significantly to this area of research.

Bahl and Tuteja (1991) envisaged a ratio and a product type exponential estimator of population mean in simple random sampling. Following Srivenkataramana (1980) and Bondyopadhyayh (1980), Lone and Tailor (2014, 2015) proposed dual to separate ratio and product type exponential estimators in the case of post-stratification.

Let us consider a finite population $U = (U_1, U_2, ..., U_N)$. A sample of size *n* is drawn from population U using simple random sampling without replacement.

[1] School of Studies in Statistics, Vikram University, Ujjain 456010, M.P, India.
E-mail: hilalstat@gmail.com.

[2] School of Studies in Statistics, Vikram University, Ujjain 456010, M.P, India.

After selecting the sample, it is observed which units belong to h^{th} stratum. Let n_h be the size of the sample falling in h^{th} stratum such that $\sum_{h=1}^{L} n_h = n$. Here, it is assumed that n is so large that the possibility of n_h being zero is very small.

Let y_{hi} be the observation on i^{th} unit that fall in h^{th} stratum for study variate y and x_{hi} be the observation on i^{th} unit that fall in h^{th} stratum for auxiliary variate x, then

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

$\overline{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{hi}$: h^{th} stratum mean of the auxiliary variate x,

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

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

In the case of post-stratification, the usual unbiased estimator of population mean $\overline{Y}$ is defined as

$$\bar{y}_{PS} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{1.1}$$

where

$W_h = \frac{N_h}{N}$ is the weight of the h^{th} stratum and $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$ is the sample mean of n_h sample units that fall in the h^{th} stratum.

Using the results from Stephen (1945), the variance of $\bar{y}_{PS}$ to the first degree of approximation is obtained as

$$Var(\bar{y}_{PS}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h S_{yh}^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1 - W_h)S_{yh}^2 \tag{1.2}$$

where $S_{yh}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(y_{hi} - \overline{Y}_h\right)^2$.

Separate ratio and product type estimators of population mean $\overline{Y}$ in the case of post-stratification are defined as

$$\hat{\overline{Y}}_{RPS} = \sum_{h=1}^{L} W_h \overline{y}_h \left(\frac{\overline{X}_h}{\overline{x}_h}\right) \tag{1.3}$$

and

$$\hat{\overline{Y}}_{PPS} = \sum_{h=1}^{L} W_h \overline{y}_h \left(\frac{\overline{z}_h}{\overline{Z}_h}\right). \tag{1.4}$$

Up to the first degree of approximation, biases and mean squared errors of the estimators $\hat{\overline{Y}}_{RPS}$ and $\hat{\overline{Y}}_{PPS}$ are obtained as

$$B\left(\hat{\overline{Y}}_{RPS}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L}\overline{Y}_h\left(C_{xh}^2 - \rho_{yxh}C_{xh}C_{yh}\right), \tag{1.5}$$

$$MSE\left(\hat{\overline{Y}}_{RPS}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[\sum_{h=1}^{L} W_h S_{yh}^2 + \sum_{h=1}^{L} W_h R_{1h}^2 S_{xh}^2 - 2\sum_{h=1}^{L} W_h R_{1h} S_{yxh}\right], \tag{1.6}$$

$$B\left(\hat{\overline{Y}}_{PPS}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L}\overline{Y}_h\, C_{yh}\, C_{zh}\, \rho_{yzh} \tag{1.7}$$

and

$$MSE\left(\hat{\overline{Y}}_{PPS}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[\sum_{h=1}^{L} W_h S_{yh}^2 + \sum_{h=1}^{L} W_h\, R_{2h}^2 S_{zh}^2 + 2\sum_{h=1}^{L} W_h\, R_{2h} S_{yzh}\right], \tag{1.8}$$

where $R_{1h} = \frac{\overline{Y}_h}{\overline{X}_h}$ and $R_{2h} = \frac{\overline{Y}_h}{\overline{Z}_h}$.

2. Improved separate ratio exponential type estimator

We suggest the improved separate ratio exponential type estimator for population mean $\overline{Y}$ in the case of post-stratification as

$$\hat{\overline{Y}}_{PS}^{(a_h)} = \sum_{h=1}^{L} W_h \overline{y}_h \exp\left(\frac{\overline{X}_h - \overline{x}_h}{\overline{X}_h + (a_h - 1)\,\overline{x}_h}\right), \tag{2.1}$$

where $a_h \geq 0$.

To obtain the bias and mean squared error of the suggested estimator $\hat{\bar{Y}}_{PS}^{(a_h)}$, we write

$$\bar{y}_h = \bar{Y}_h\left(1+e_{0h}\right),\ \ \bar{x}_h = \bar{X}_h\left(1+e_{1h}\right) \text{ such that}$$

$$E(e_{0h}) = E(e_{1h}) = 0,$$

$$E(e_{0h}^2) = \left(\frac{1}{nW_h} - \frac{1}{N_h}\right)C_{yh}^2,$$

$$E(e_{1h}^2) = \left(\frac{1}{nW_h} - \frac{1}{N_h}\right)C_{xh}^2\ ,$$

$$E(e_{0h}e_{1h}) = \left(\frac{1}{nW_h} - \frac{1}{N_h}\right)\rho_{yxh}C_{yh}C_{xh}\ .$$

Expressing (2.1) in terms of $e_{ih}'s$, we have

$$\hat{\bar{Y}}_{PS}^{(a_h)} = \sum_{h=1}^{L} W_h\bar{Y}_h\left(1+e_{0h}\right)\exp\left[-\frac{e_{1h}}{a_h}\left\{1+\left(\frac{a_h-1}{a_h}\right)e_{1h}\right\}^{-1}\right]$$

Now, by expanding the exponential function on the right-hand side, we get

$$\hat{\bar{Y}}_{PS}^{*\mathrm{Re}} = \sum_{h=1}^{L} W_h\bar{Y}_h\ \left(1+e_{0h}\right)\left[1-\frac{e_{1h}}{a_h}\left\{1+\left(\frac{a_h-1}{a_h}\right)e_{1h}\right\}^{-1} + \frac{1}{2}\frac{e_{1h}^2}{a_h^2}\left\{1+\left(\frac{a_h-1}{a_h}\right)e_{1h}\right\}^{-2} - ....\right]$$

$$\hat{\bar{Y}}_{PS}^{*\mathrm{Re}} = \sum_{h=1}^{L} W_h\bar{Y}_h\left(1+e_{0h}\right)\left[1-\frac{e_{1h}}{a_h}\left\{1-\left(\frac{a_h-1}{a_h}\right)e_{1h}+\left(\frac{a_h-1}{a_h}\right)^2e_{1h}^2\right\}+\frac{1}{2}\frac{e_{1h}^2}{a_h^2}\left\{1-2\left(\frac{a_h-1}{a_h}\right)e_{1h}\right\}\right]$$

$$\hat{\bar{Y}}_{PS}^{*\mathrm{Re}} = \sum_{h=1}^{L} W_h\bar{Y}_h\left(1+e_{0h}\right)\left[1-\frac{e_{1h}}{a_h}+\frac{e_{1h}^2}{a_h^2}\left(a_h-\frac{1}{2}\right)\right]$$

$$\left(\hat{\bar{Y}}_{PS}^{(a_h)} - \bar{Y}\right) = \sum_{h=1}^{L} W_h\bar{Y}_h\left[e_{0h} - \frac{e_{1h}}{a_h} + \left(a_h - \frac{1}{2}\right)\frac{e_{1h}^2}{a_h^2} - \frac{e_{0h}e_{1h}}{a_h}\right] \tag{2.2}$$

Now, taking expectation of both sides of (2.2), the bias of the suggested estimator $\hat{\bar{Y}}_{PS}^{(a_h)}$ to the first degree of approximation is obtained as

$$B(\hat{\bar{Y}}_{PS}^{(a_h)})=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{h=1}^{L}\frac{1}{a_h^2\,\bar{X}_h}\left[\left(a_h-\frac{1}{2}\right)R_{1h}\,S_{xh}^2-a_h S_{yxh}\right] \qquad (2.3)$$

Squaring both sides of (2.2) and then taking expectation, we get the mean squared error of the suggested estimator $\hat{\bar{Y}}_{PS}^{(a_h)}$ up to the first degree of approximation as

$$MSE\left(\hat{\bar{Y}}_{PS}^{(a_h)}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{h=1}^{L}W_h\left(S_{yh}^2+\frac{R_{1h}^2}{a_h^2}S_{xh}^2-2\frac{R_{1h}}{a_h}S_{yxh}\right) \qquad (2.4)$$

which is minimized for

$$a_h=\frac{R_{1h}}{\beta_h}=a_{ho}\quad (say), \qquad (2.5)$$

where $R_{1h}=\dfrac{\bar{Y}_h}{\bar{X}_h}$ and $\beta_h=\dfrac{S_{yxh}}{S_{xh}^2}$.

Putting (2.5) in (2.4), we get the minimum mean squared error of the estimator $\hat{\bar{Y}}_{PS}^{(a_h)}$ up to the first degree of approximation as

$$\min.MSE\left(\hat{\bar{Y}}_{PS}^{(a_h)}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{h=1}^{L}W_h\,S_{yh}^2\left(1-\rho_h^2\right), \qquad (2.6)$$

where $\rho_h=\dfrac{S_{yxh}}{S_{yh}\,S_{xh}}$.

Putting (2.5) in (2.1), we get the asymptotic optimum estimator (AOE) in the class of estimators $\hat{\bar{Y}}_{PS}^{(a_h)}$ as

$$\hat{\bar{Y}}_{PS}^{(a_{h0})}=\sum_{h=1}^{L}W_h\bar{y}_h\,\exp\left(\frac{\beta_h\,(\bar{X}_h-\bar{x}_h)}{\beta_h\bar{X}_h+(R_{1h}-\beta_h)\,\bar{x}_h}\right). \qquad (2.7)$$

3. Estimator based on estimated optimum

It is obvious that the estimator $\hat{\bar{Y}}_{PS}^{(a_{h0})}$ in (2.7) requires prior information of (R_{1h},β_h), which can be obtained easily from previous surveys. If the investigator is unable to guess the value of (R_{1h},β_h), the only alternative he is left with is to replace $(R_{1h}\,,\beta_h)$ in (2.7) by its consistent estimate $\hat{a}_h=\dfrac{\hat{R}_{1h}}{\hat{\beta}_h}$, where $\hat{\beta}_h=\dfrac{s_{yxh}}{s^2_{xh}}$ and $\hat{R}_{1h}=\dfrac{\bar{y}_h}{\bar{x}_h}$. Hence, the estimator based on estimated optimum is

$$\hat{\bar{Y}}^{(\hat{a}_{h0})}_{PS}=\sum_{h=1}^{L}W_h\bar{y}_h\exp\left(\frac{\hat{\beta}_h\,(\bar{X}_h-\bar{x}_h)}{\bar{y}_h+\hat{\beta}_h(\bar{X}_h-\bar{x}_h)}\right) \tag{3.1}$$

Up to the first degree of approximation, the mean squared error of the estimator $\hat{\bar{Y}}^{(\hat{a}_{h0})}_{PS}$ is given by

$$MSE\left(\hat{\bar{Y}}^{(\hat{a}_{h0})}_{PS}\right)=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{h=1}^{L}W_h\,S^2_{yh}\left(1-\rho^2_h\right) \tag{3.2}$$

which is the same as given in (2.6)

4. Efficiency comparisons of the suggested improved ratio exponential type estimator $\hat{\bar{Y}}^{(a_h)}_{PS}$ with $\hat{\bar{Y}}_{PS}$ and $\hat{\bar{Y}}_{RPS}$.

From (1.2), (1.6) and (2.4), it is observed that the suggested estimator $\hat{\bar{Y}}^{(a_h)}_{PS}$ would be more efficient than

(i) the usual unbiased estimator $\hat{\bar{Y}}_{PS}$ if

$$\sum_{h=1}^{L}\frac{R_{1h}}{a^2_h}W_h\left(R_{1h}S^2_{xh}-2\,a_hS_{yxh}\right)<0\ , \tag{4.1}$$

(ii) the usual separate ratio estimator $\hat{\bar{Y}}_{RPS}$ if

$$\sum_{h=1}^{L}W_h\left(R^2_{1h}S^2_{xh}\left\{\frac{1}{a^2_h}-1\right\}-2S_{yxh}R_{1h}\left\{\frac{1}{a_h}-1\right\}\right)<0\,. \tag{4.2}$$

5. Improved separate product exponential type estimator

Improved separate product exponential type estimator for population mean $\bar{Y}$ in the case of post-stratification is being suggested as

$$\hat{\bar{Y}}^{(b_h)}_{PS}=\sum_{h=1}^{L}W_h\bar{y}_h\exp\left(\frac{\bar{z}_h-\bar{Z}_h}{\bar{Z}_h+(b_h-1)\,\bar{z}_h}\right), \tag{5.1}$$

where $b_h \geq 0$.

The estimator $\hat{\bar{Y}}_{PS}^{(b_h)}$ in terms of $e's$ can be written as

$$\left(\hat{\bar{Y}}_{PS}^{(b_h)} - \bar{Y}\right) = \sum_{h=1}^{L} W_h \bar{Y}_h \left[e_{0h} + \frac{e_{2h}}{b_h} + \left(\frac{3}{2} - b_h\right)\frac{e_{2h}^2}{b_h^2} + \frac{e_{0h}e_{2h}}{b_h} \right] \tag{5.2}$$

Using the standard procedure, the bias and mean squared error of the suggested estimator $\hat{\bar{Y}}_{PS}^{(b_h)}$ are obtained as

$$B(\hat{\bar{Y}}_{PS}^{(b_h)}) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} \frac{W_h}{\bar{Z}_h}\left(\left(\frac{3}{2} - b_h\right)\frac{R_{2h}^2}{b_h^2} S_{zh}^2 + \frac{1}{b_h} S_{yzh}\right) \tag{5.3}$$

and

$$MSE\left(\hat{\bar{Y}}_{PS}^{(b_h)}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h \left(S_{yh}^2 + \frac{R_{2h}^2}{b_h^2} S_{zh}^2 + 2\frac{R_{2h}}{b_h} S_{yzh}\right) \tag{5.4}$$

which is minimized for

$$b_h = -\frac{R_{2h}}{\beta_h^*} = b_{ho} \ (say), \tag{5.5}$$

where $\rho_h^* = \dfrac{S_{yzh}}{S_{yh}\ S_{zh}}$ and $\beta_h^* = \dfrac{S_{yzh}}{S_{zh}^2}$.

Putting (5.5) in (5.4), we get the minimum mean squared error of the estimators $\hat{\bar{Y}}_{PS}^{(b_h)}$ to the first degree of approximation given as

$$\min . MSE\left(\hat{\bar{Y}}_{PS}^{(b_h)}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\sum_{h=1}^{L} W_h\, S_{yh}^2 \left(1 - \rho_h^{*2}\right) \tag{5.6}$$

Putting (5.5) in (5.1), we get the asymptotic optimum estimator (AOE) in the class of estimators $\hat{\bar{Y}}_{PS}^{(b_h)}$ as

$$\hat{\bar{Y}}_{PS}^{(b_{h0})} = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\beta_h^*(\bar{z}_h - \bar{Z}_h)}{\beta_h^* \bar{Z}_h - (R_{2h} + \beta_h^*)\,\bar{z}_h}\right), \tag{5.7}$$

with the same mean square as given in (5.6)

6. Estimator based on estimated optimum value of b_{ho}

If the value of (R_{2h}, β_h^*) is not known in advance, then it is advisable to replace them by its consistent estimate $(\hat{R}_{2h}, \hat{\beta}_h^*)$ computed from the sample values. Hence, the estimator based on estimated optimum is

$$\hat{\bar{\bar{Y}}}_{PS}^{(\hat{b}_{h0})} = \sum_{h=1}^{L} W_h \bar{y}_h \exp\left(\frac{\hat{\beta}_h^* (\bar{Z}_h - \bar{z}_h)}{\bar{y}_h + \hat{\beta}_h^* (\bar{z}_h - \bar{Z}_h)} \right) \tag{6.1}$$

The mean squared error of the estimator $\hat{\bar{\bar{Y}}}_{PS}^{(\hat{b}_{h0})}$ up to the first degree of approximation is given by

$$MSE\left(\hat{\bar{\bar{Y}}}_{PS}^{(\hat{a}_{h0})} \right) = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{h=1}^{L} W_h \, S_{yh}^2 \left(1 - \rho_h^{*2} \right) \tag{6.2}$$

which is the same as given in (5.6)

7. Efficiency comparisons of the suggested estimator $\hat{\bar{\bar{Y}}}_{PS}^{(b_h)}$ with $\hat{\bar{\bar{Y}}}_{PS}$ and $\hat{\bar{\bar{Y}}}_{PPS}$

From (1.2), (1.8) and (5.4), it is concluded that the suggested estimator $\hat{\bar{\bar{Y}}}_{PS}^{(b_h)}$ would be more efficient than

(i) the usual unbiased estimator $\hat{\bar{\bar{Y}}}_{PS}$ if

$$\sum_{h=1}^{L} W_h \frac{R_{2h}}{b_h^2} \left(R_{2h} S_{zh}^2 + 2 b_h S_{yzh} \right) < 0 \tag{7.1}$$

(ii) the usual separate product estimator $\hat{\bar{\bar{Y}}}_{PPS}$ if

$$\sum_{h=1}^{L} W_h \left(\left\{ \frac{1}{b_h^2} - 1 \right\} R_{2h}^2 S_{zh}^2 + 2 \left\{ \frac{1}{b_h} - 1 \right\} R_{2h} S_{yzh} \right) < 0 \,. \tag{7.2}$$

8. Empirical study

To judge the performance of the suggested estimators we are considering two natural population data sets, the descriptions of populations are given below:

Population I- [Source: National horticulture Board]

y : Productivity (MT/Hectare)

x : Production in ‘000 Tons and

z : Area in ‘000 Hectare

Constant	Stratum I	Stratum II
N_h	10	10
n_h	4	4
$\bar{Y}_h$	1.70	3.67
$\bar{X}_h$	10.41	289.14
$\bar{Z}_h$	6.32	80.67
S_{yh}	0.50	1.41
S_{xh}	3.53	111.61
S_{zh}	1.19	10.82
S_{yxh}	1.60	144.87
S_{yzh}	-0.05	-7.04
S_{xzh}	1.38	-92.02

Population II- [Source: Chouhan (2012)]

y : Snowy days

x : Rainy days and

z : Total annual sunshine hours

Constant	Stratum I	Stratum II
N_h	10	10
n_h	4	4
$\overline{Y}_h$	149.7	102.6
$\overline{X}_h$	142.8	91.0
$\overline{Z}_h$	1629.9	2035.9
S_{yh}	13.46	12.60
S_{xh}	6.09	6.57
S_{zh}	102.17	103.26
S_{yxh}	18.44	23.30
S_{yzh}	-1072.8	-655.25
S_{xzh}	-239.25	-240.45

Table 8.1. Percent Relative Efficiencies of $\hat{\bar{\bar{Y}}}_{PS}$, $\hat{\bar{\bar{Y}}}_{RPS}$, $\hat{\bar{\bar{Y}}}_{PPS}$, $\hat{\bar{\bar{Y}}}_{PS}^{(a_{h0})}$ and $\hat{\bar{\bar{Y}}}_{PS}^{(b_{h0})}$ with respect to $\hat{\bar{\bar{Y}}}_{PS}$

Estimators	**Percent Relative Efficiencies (PRE's)**	
	Population I	**Population II**
$\hat{\bar{\bar{Y}}}_{PS}$	100.00	100.00
$\hat{\bar{\bar{Y}}}_{RPS}$	593.50	98.72
$\hat{\bar{\bar{Y}}}_{PPS}$	116.84	176.97
$\hat{\bar{\bar{Y}}}_{PS}^{(a_{h0})}$	**643.41**	**106.82**
$\hat{\bar{\bar{Y}}}_{PS}^{(b_{h0})}$	**123.44**	**179.35**

9. Conclusion

Section 4 and 7 provides the conditions under which the suggested estimators $\hat{\bar{\bar{Y}}}_{PS}^{(a_{h0})}$ and $\hat{\bar{\bar{Y}}}_{PS}^{(b_{h0})}$ have fewer mean squared errors in comparison with usual unbiased estimator and separate ratio and product type estimators in the case of post-stratification. Table 8.1 shows that the suggested estimators $\hat{\bar{\bar{Y}}}_{PS}^{(a_{h0})}$ and $\hat{\bar{\bar{Y}}}_{PS}^{(b_{h0})}$ have higher percent relative efficiencies in comparison with usual unbiased estimator $\hat{\bar{\bar{Y}}}_{PS}$, separate ratio and product type estimators $\hat{\bar{\bar{Y}}}_{RPS}$ and $\hat{\bar{\bar{Y}}}_{PPS}$. Thus, the suggested estimators are recommended for use in practice for estimating the population mean when conditions obtained in section 4 and 7 are satisfied.

REFERENCES

AGRAWAL, M. C., PANDEY, K. B., (1993). An efficient estimator in post-stratification. Metron 51, 179−188.

BAHL, S., TUTEJA, R. K., (1991). Ratio and product type exponential estimators. Infor. & Optimiz. Sci., 12, 159−163.

BANDYOPADHYAY, S., (1980). Improved ratio and product estimators. Sankhya 12, C, 142, 45−49.

CHOUHAN, S., (2012). Improved estimation of parameters using auxiliary information in sample surveys. Ph. D. Thesis, Vikram University, Ujjain, M.P., India.

FULLER, W., (1966). Estimation employing post-strata. J. Amer. Statist. Assoc., 61, 1172−1183.

HANSEN, M. H., HURWITZ, W. N., MADOW, W. G., (1953). Sample survey methods and theory. Vol. I & II, John Wiley & Sons, New York.

HOLT, D., SMITH, T. M. F., (1979). Post-stratification. J. Roy. Statist. Soc., 142, A, 33−46.

IGE, A. F., TRIPATHI, T. P., (1989). Estimation of population mean using post-stratification and auxiliary information. Abacus, 18, 2, 265−276.

JATWA, N. K., (2014). Estimation of population parameters in presence of auxiliary information. Ph.D. Thesis, Vikram University, Ujjain, M.P. India.

KISH, L., (1965). Survey Sampling. John Wiley & Sons, New York.

LONE, H. A., TAILOR, R., (2015). Dual to Separate Product Type Exponential Estimator in Sample Surveys. J. Statist. Appl. Prob. Lett. 2, 1, 89−96.

LONE, H. A., TAILOR, R., (2014). Dual to Separate Ratio Type Exponential Estimator in Post-Stratification. J. Statist. Appl. Prob. 3, 3, 425−432.

RAJ, D., (1972). The design of sample surveys. McGraw Hill, New York.

SRIVENKATARAMANA, T., (1980). A dual to ratio estimator in sample surveys. Biometr., J. 67, 1, 199−204.

STEPHAN, F., (1945). The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate. Ann. Math. Statist. 16, 50−61.

TAILOR, R., JATWA, S. K., LONE, H. A., (2015). Dual to ratio and product type estimators in case of post-stratification. J. Mod. Appl. Statist. Meth. (in press).

http://nhb.gov.in/statistics/area-production-statistics.html (Official website of National Horticulture Board, India).

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 65–82

POLICY-ORIENTED INFERENCE AND THE ANALYST-CLIENT COOPERATION. AN EXAMPLE FROM SMALL-AREA STATISTICS

Nicholas T. Longford[1]

ABSTRACT

We show on an application to small-area statistics that efficient estimation is not always conducive to good policy decisions because the established inferential procedures have no capacity to incorporate the priorities and preferences of the policy makers and the related consequences of incorrect decisions. A method that addresses these deficiencies is described. We argue that elicitation of the perspectives of the client (sponsor) and their quantification are essential elements of the analysis because different estimators (decisions) are appropriate for different perspectives. An example of planning an intervention in a developing country's districts with high rate of illiteracy is described. The example exposes the deficiencies of the general concept of efficiency and shows that the criterion for the quality of an estimator has to be formulated specifically for the problem at hand. In the problem, the established small-area estimators perform poorly because the minimum mean squared error is an inappropriate criterion.

Key words: composition, empirical Bayes, expected loss, borrowing strength, exploiting similarity, shrinkage, small-area estimation.

1. Introduction

Survey methods have in the recent decades been greatly stimulated by the big-budget departments of national governments, such as social security, health care, education and employment, owing to their greater appreciation of the role of statistical information and inference in policy making. Developments in small-area estimation have responded to the demand for greater detail about the (administrative) divisions of a country, such as regions and districts. In a typical setting, a national survey is conducted, collecting information about the key variables, such as employment status, and an established set of background variables (age, sex, educational level, marital status, and the like), and estimates related to a key variable

[1] SNTL and Universitat Pompeu Fabra, Barcelona, Spain. E-mail: sntlnick@sntl.co.uk

are sought for each district. The districts are of varying sizes, and some of them are represented in the survey by small samples that on their own are not sufficient for estimating their key characteristics, usually percentages (e.g., the unemployment rate), with any appreciable precision.

The main advance in small-area estimation is the exploitation of similarity, also referred to as borrowing strength, in any aspect for which (auxiliary) data is available. Obvious examples of such data (and information) are values of the target variable observed in the other districts of the country, the background variables recorded in the survey, and information obtained from censuses and surveys. Somewhat less appreciated is the potential of variables prima facie related to the target variable and information from the previous years of the surveys in the same programme. For estimating subpopulation characteristics (e.g., for minorities), other subpopulations (or the complementary subpopulation) are often very effective auxiliaries. See Longford (2005) for examples. In particular, insisting on having the values of auxiliary variables for the entire population is extremely restrictive (Elbers, Lanjouw and Lanjouw, 2003).

In small-area estimation, as in other survey inference, efficient estimation is generally regarded as superior. A lot of the theory is concerned with deriving estimators that are efficient, or nearly so, sometimes in uncongenial circumstances, using models known not to be valid. Robust estimation is therefore regarded as invaluable. Estimation of standard errors of these estimators is also an important preoccupation. This research implies, often without a clinical statement to that effect, that the estimates obtained are best suited for a policy related to the target variable and that it will be well informed by efficient small-area estimators. The statistical analysis is concluded by the presentation of the estimates and the associated standard errors, and the remainder of the analysis is left to be dealt with by the policy maker. We show by example that this is a poor strategy and argue that the analytical skills of a statistician and the insight and other qualities of the policy maker have to be integrated much more closely.

The core of the problem is that the consequences of the (estimation) errors made have to be taken into account because they are in substantial discord with the default assumption of the (symmetric) quadratic loss. These consequences are difficult to elicit from experts and to quantify them, but that is hardly an excuse for applying methods that assume a particular structure of such consequences, especially when the assumed structure is in obvious discord with the client's perception. The analyst may not be aware of the consequences and of their relevance, and therefore would not communicate the (default) assumptions to the client. It is essential to sensitise the client to this issue because the solution to the problem is statistical, and there is a danger that when the client becomes aware of the issue he or she will seek a second-rate ad hoc solution not integrated with the original (incomplete) analysis.

In the next section, we outline the policy planned in an application and describe a survey that is intended to inform the policy. The policy is related to combating illiteracy in the districts of a developing country. Illiteracy is regarded as a major barrier to gaining employment and to economic development in general, but also to the spread of government information and, admittedly, of the governing party's political propaganda as well. In Section 3, we review the established methods for small-area estimation and highlight their deficiency which is then addressed in Section 4 by the proposed estimator, called *policy-related*. In Section 5, we compare by simulations the implementations of the policy using three estimators:

- direct estimators that no one would recommend (for small districts);
- composite (empirical Bayes) estimators that are generally regarded as superior;
- policy-related estimators constructed with the intent to minimise the total expected loss.

We show that for a wide range of perspectives and priorities of the decision maker the policy-related estimator is far superior to the other two estimators. The concluding section discusses the implications of this result on how small-area analysis should be conducted, and extends them to some general principles, including how official statistical institutes should operate.

Direct estimators use no auxiliary information; they are based only on the data for the target variable and the region concerned. Often they are standard survey-methods estimators (see, e.g., Särndal, Bengtsson and Wretman, 1992) restricted to the region. Empirical Bayes estimators exploit auxiliary information by means of a two-level regression model (Goldstein, 2002; Rao, 2003) and composite estimators (Longford, 1999 and 2005) combine direct and auxiliary estimators (or exact quantities) without a reference to a model. The policy-related estimator is developed in Longford (2013, Chapter 7, and 2015), where some technical details omitted from this article can be found. The method exploits auxiliary information, but requires also input about the consequences (ramifications) of the errors that may be committed. These errors are:

- *false negative*: failure to apply an intervention when it should have been applied;
- *false positive*: applying the intervention when it is not necessary.

Our interest is in settings in which these consequences are uneven. In the example of combating endemic illiteracy, a false negative has much more serious consequences (greater losses) than a false positive. Empirical Bayes methods and, more generally, methods that aim to minimize the mean squared error (MSE), are oblivious to such consequences.

Similar issues arise in medical screening (Longford, 2013, Chapter 6), where a false positive (incorrect labelling as diseased) is regarded as less serious an error than a false negative (failure to discover the disease), in (production) quality control, where a false claim of satisfying a standard is much more costly (in the long term) than the pursuit of further (unessential) improvements when the standard has already been achieved, and in the operation of warning systems (for epidemics, natural disasters, military or terrorist attacks, and the like), where false warnings may erode the credibility of the system, but a failure to warn is an unmitigated disaster.

2. Illiteracy in the districts of a country

We consider a developing country with an adult population (aged 16 or over) approaching 40 million and adult illiteracy rate of about 18%. The country has 72 districts, of population sizes (numbers of adults) varying from 50 000 to 1.8 million (the capital). Illiteracy tends to be more prevalent in rural districts. Its rates are smaller in the most populous districts which are mostly urban (large cities and their environs). However, some less populous districts are formed by single (smaller) towns and cities, and there the illiteracy rates tend to be smaller. Some of these towns are satellites of larger cities.

The Ministry of Education has appropriated funds for conducting a survey to study the illiteracy rate in the country. The results would then be used for implementing a particular policy aimed at the districts with illiteracy rates higher than 25%. The survey has several sponsors and subscribers with purposes different from small-area estimation, and so a compromise stratified sampling design is implemented, with the districts as the strata. Some clustering is applied within the strata, which is of marginal interest for our purposes, and its details are omitted. It is impossible to ensure that each stratum (district) would have a sample size sufficient for reliable direct estimation of its illiteracy rate.

Denote by θ_d the rate of illiteracy in district $d = 1,\ldots,D$, and by $\hat{\theta}_d$ and $\tilde{\theta}_d$ the respective direct and composite estimators of θ_d. The composite estimator is defined as the convex combination of the direct and (overall) national estimator,

$$(1-b_d)\,\hat{\theta}_d + b_d\,\hat{\theta}\,, \tag{1}$$

where the coefficient $b_d = 1/(1+n_d\,\omega)$ involves the ratio $\omega = \sigma_{\mathrm{B}}^2/\sigma^2$ of the between- and within-stratum variances and n_d is the sample size of the stratum. When the sampling weights are not constant within strata, n_d has to be replaced by the effective sample size. The variance σ^2 is estimated by pooling the within-district estimates of the variance and σ_{B}^2 is estimated by moment matching applied to the sum-of-squares statistic $\sum_d(\hat{\theta}_d - \hat{\theta})^2$.

As part of the policy, an intervention is designed and planned to be applied in districts in which $\theta_d > 0.25$. We assume that it will be applied to districts with $\hat{\theta}_d > 0.25$ or $\tilde{\theta}_d > 0.25$, or to districts that satisfy this inequality for a different estimator. At the beginning of the author's involvement, all parties involved agreed that the composite estimator $\tilde{\theta}_d$ should be used. The Research Department of the Ministry agreed that a simulation study would be conducted, mainly to assess the potential problems with districts that have extreme rates of illiteracy. The related methodological issue is discussed in Longford (2007), and it concerns mainly estimation of the standard errors and the claim that empirical Bayes and composite estimators are superior to direct estimators for every district. Also, the funding for the survey and the intervention come from the same budget, and so its split for the survey (data collection and analysis) and policy implementation was negotiated extensively, until it was agreed that a simulation study would inform this issue.

The simulation study, and the detailed discussion of its set-up, including the information on which it would be based, as well as the arguments about how to evaluate the results, brought to the fore the purpose of the survey, namely, allocation of funds to the districts. Here the established criterion of minimum MSE turned out to be irrelevant because in the Ministry's perspective, the evaluation should focus on the two types of error (false negative and false positive) in classifying the districts as

- deserving the intervention ($\theta_d > 0.25$), and

- not deserving the intervention ($\theta_d < 0.25$).

At first, this might suggest that hypothesis testing (HT) should be applied. However, this was also dismissed after a set of simulations when it became clear that HT is oblivious to the consequences of the two kinds of error. The following example disqualifies HT from almost any problem in which we have to decide how to proceed; as if the (null) hypothesis were valid, or not. Suppose incorrect omission of a district from intervention is five times as serious an error as incorrect inclusion. By a hypothesis test, with the conventional level of significance of $\alpha = 0.05$, we would come to a particular decision. Next, suppose incorrect omission is 50 times more serious than incorrect inclusion. With conventional statistical tools, and conventional operational mindset, we would apply the same hypothesis test, and come to the same decision. No procedure that always comes to the same conclusion in these two settings could possibly be appropriate. A decision (choice between two complementary options) is poorly founded if it is not influenced by the consequences of the two kinds of incorrect choices, unless the decision is always correct (or always incorrect) and entails no uncertainty. In the general problem of estimation, a similar dismissing argument is easy to formulate. The consequences of the errors that are committed in estimation have to be an important factor in how an estimator is constructed. Ignoring them is a licence for making poor decisions and, ultimately, rendering the statistical analysis irrelevant.

Eliciting information about the consequences of errors is, unfortunately, not a mainstream statistical activity. In practice, it can be surprisingly contentious, because many clients believe that ‘it is all in the data’, and the analyst’s task is to process the data and deliver an unambiguous verdict. Also, a client may suspect that the analyst wants to elicit the client’s perspective, priorities and goals, merely to fix up the results so as to superficially please the client, or to obtain confidential information that would later be disclosed to the client’s detriment. In the political and civil-service sphere, the value of information is well appreciated, but often leads to the practice of divulging it on a strictly need-to-know basis. This involves liberally placed controls and requirements for approvals (hurdles) across the layers of management that hinder and sometimes entirely disable the process of informing the statistical analysis. A change in the perspectives and priorities may appear as an embarrassment to the client. However, it is a responsible act when it responds appropriately to new information and circumstances. Elicitation is more commonly considered for prior distributions in Bayesian estimation (Garthwaite, Kadane and O’Hagan, 2005). The same principles apply to elicitation in our context, although much less of the experience is recorded in the literature.

An important element of elicitation is to put the clients (or experts) at ease by not rushing them to any quick decisions (or an uneasy consensus), and explaining that they are not expected to have answers ready at a moment's notice. In the example of combating illiteracy, the key question relates to the so-called *penalty ratio R* which quantifies how many times more costly is an error of one kind than of the other. There is no need to conclude with a single value *R*. It is more constructive to set (or declare) a plausible range of values of *R*. The key property of such a range is that any value outside it can be ruled out—that the client is satisfied that such a value is not realistic. Of course, it is advantageous to have as narrow a plausible range as possible, but its plausibility is an imperative. The wider the range, the greater the threat of an inferential *impasse*, when both decisions are plausible; one for certain plausible values of *R* and the other for the complement.

The rationale for reflecting in inferential statements different perspectives and priorities is hinted by Shen and Louis (1998). They coined the term 'triple-goal estimator' to highlight the need for different estimators for three distinct purposes in small-area estimation: estimating each district's population quantity, ranking the districts according to this quantity, and estimating the district-level distribution of these quantities. A compact summary of their conclusion is that efficiency and unbiasedness (of an estimator) are *fragile* properties. Fragility refers to the fact that these properties are not maintained by non-linear transformations, and even less so by some discontinuous ones, such as assessing whether a parameter is greater or smaller than a set threshold. We can paraphrase these conclusions by saying that optimality of an estimator is conditioned on the scale used for the error in estimation. By the same token, an estimator with minimum MSE may be far from optimal with a distinctly asymmetric loss function. Loss function is as important an input and has a similar nature as (informative) prior distribution in Bayesian analysis, where it is taken for granted that different priors lead to different posteriors and conclusions based on them. On this account, we have to dismiss the idea that the results of a respectable analysis, even in a frequentist paradigm, have to be 'objective' and applicable to a wide range of perspectives. Instead, they have to be client-specific, that is, responsive and sensitive to the client's perspective and value judgements.

3. Shrinkage estimation

Empirical Bayes estimators (Robbins, 1995; Carlin and Louis, 2000) are a general example of shrinkage estimators. Composite estimators (Longford, 1999) are also shrinkage estimators; in fact, they apply shrinkage directly, without the intermediation of a model, and thus absolve the analyst from the responsibilities related to the validity of the model, including the distributional assumptions. In small-area estimation, this feature is important because the analysts rarely have the freedom to choose what auxiliary information will be used; they have to operate with what is available and can neither present an excuse nor apportion the blame to anybody when the model is assessed not to fit well.

Shrinkage estimators pull a direct estimator $\hat{\theta}_d$ for district d toward a (national) focus $\hat{\theta}$; see equation (1). The amount of shrinkage (the strength of the pull) depends on the balance of the sampling variance of $\hat{\theta}_d$, equal to σ^2/n_d in the simplest setting, and the district level variance $\sigma_{\mathrm{B}}^2 = \mathrm{var}_d(\theta_d)$. The latter variance is defined for the districts and is not related to sampling; it is a population quantity. If σ_{B}^2 is very small, then the shrinkage is strong because the districts are very similar and the national estimator $\hat{\theta}$ is very useful for every district. In contrast, if σ_{B}^2 is very large, much less shrinkage takes place because θ may be far away from θ_d, and therefore $\hat{\theta}$ a poor estimator of θ_d; this is the case for a substantial fraction of the districts. Further, more shrinkage takes place for small districts (districts with large sampling variance of $\hat{\theta}_d$) and less for districts with more precise estimators $\hat{\theta}_d$.

Although this interpretation applies only to the simplest form of (univariate) shrinkage estimation, it offers some insights as to why it may be poorly suited for the Ministry's task. If a false positive has less serious consequences than a false negative then we should focus on the deserving districts, for which only errors of the former type are possible. In our example, these districts are in a minority because the threshold of $T = 25\%$ is far greater than the national rate of about 18%. Direct estimation for the deserving districts will result in an error if the estimation error is negative and $\hat{\theta}_d < 0.25 < \theta_d$. Positive or small negative estimation errors have no consequences because then both $\hat{\theta}_d$ and θ_d exceed the threshold. The likelihood that $\hat{\theta}_d$ is close to the national rate, $\theta \doteq 0.18$, is quite small for most deserving districts, because an error smaller than (more negative than) -0.07 is quite rare.

Shrinkage applied to the direct estimator in (1) pulls it toward the national rate, and therefore reduces it for nearly all deserving districts. As an aside, note that this implies that empirical Bayes estimation for a deserving district is biased, contrary to the acronym EBLUP (empirical Bayes linear *unbiased* predictor) used in the context of hierarchical linear models. There may be deserving districts with $\hat{\theta}_d > T > \tilde{\theta}_d$, for which direct estimation would lead to the appropriate decision (intervention), but the shrinkage estimation would yield the inappropriate decision. The opposite,

$\hat{\theta}_d < T < \tilde{\theta}_d$, is less likely to happen, because the focus of shrinkage is $\hat{\theta}$, an efficient estimator of θ, and θ is much smaller than T. Thus, efficient estimation contradicts good policy implementation. This is in accord with a clinical proposal described in Longford (2013, Chapter 7), in which shrinkage is applied, but with a different focus, and to an extent different from the empirical Bayes shrinkage.

4. Policy-related estimation

Before describing the proposed estimator, we give a minimum background to decision theory, which motivates it. Suppose our target is a quantity (parameter) ν, and let $\hat{\nu}$ be an estimator of ν. The estimator is unlikely to be without error; $\Delta\nu = \hat{\nu} - \nu \neq 0$. The conventional criterion for 'good' (efficient) estimation, the MSE, assigns the cost of $\Delta\nu^2 = (\hat{\nu} - \nu)^2$, and the efficient estimator is defined as the one that minimises the expectation $\mathrm{E}(\Delta\nu^2)$.

Suppose the cost is not symmetric. A simple adaptation of MSE is that the cost is $(\hat{\nu} - \nu)^2$ if $\hat{\nu} > \nu$, but it is $R(\hat{\nu} - \nu)^2$ if $\hat{\nu} < \nu$. That is, given a fixed absolute error $|\Delta\nu|$, understatement is R times more costly than overstatement. The penalty ratio R is positive, but may be smaller than unity. No generality is lost by having a factor (R) with only one of the squared errors, because multiplying both error functions by the same constant does not alter the nature of the costs; only their relative size matters, and it is very convenient that their ratio is a constant (R). An estimator may be optimal for $R = 1$, that is, for MSE as the criterion, but then it is not optimal for $R = 10$, nor for $R = 0.1$.

Further, suppose we are not interested in the value of ν as such, but merely want to establish whether ν is greater or smaller than a threshold T. In this setting, estimation is associated with no error if $\hat{\nu}$ and ν are on the same side of T—if both $\hat{\nu}$ and ν are greater than T, or both are smaller. Suppose one unit of loss is incurred if $\nu < T$ but $\hat{\nu} > T$ (incorrect inclusion, in the context of our study) and R units are lost if $\nu > T$ but $\hat{\nu} < T$ (incorrect omission).

This setting resembles HT, but if we wanted to apply it we would not know which case ($T \leq 0.25$ or $T \geq 0.25$) to declare as the hypothesis and which as the alternative. Even if we resolved this issue, we would not know how to act when the hypothesis is not rejected because in that case the hypothesis has not been accepted, merely we would have failed to find evidence against it. On the one hand, we are well aware of this; on the other hand, we liberally abuse this wisdom because the correct conclusion that we are ignorant about the relation of ν and T is unacceptable.

A hypothesis test can provide evidence for its alternative, by concluding that there is a probabilistic contradiction with the hypothesis. But in the absence of such a contradiction, it does not provide any evidence for the hypothesis. Continuing the analysis, a business agenda, or any other plan as if the hypothesis were valid, where there is no support for it, is a common logical inconsistency that does no favours to the image of any scientist.

In the decision-theoretical approach (Lindley, 1985; DeGroot, 2004), we evaluate the expected loss with the two options we have, to conclude A, that $\nu < T$, or B, that $\nu > T$, and choose the option that is associated with smaller expected loss. The evaluations are somewhat more complex than in HT, but they are a small price to pay for better allocation of our own money (assuming that we all are taxpayers) or, in general, for tailoring the solution closer to the clients' perspective, priorities and goals.

The policy-related estimator is developed from the following considerations. Let

$$\tilde{\theta}_d^* = (1-b_d)\,\hat{\theta}_d + b_d F_d\,, \tag{2}$$

where b_d is the shrinkage coefficient and F_d is the focus of shrinkage, set separately for each district. In empirical Bayes and composite estimation, $F_d \equiv \hat{\theta}$; see (1). In our problem, one might contemplate $F_d \equiv T$. Both proposals lead to poor solutions. The coefficients b_d and F_d are determined by two conditions:

- *equilibrium at T*: the choice between options A and B is immaterial for a district with $\theta_d = T$;
- minimum MSE.

The solution is derived in the Appendix.

The second condition is somewhat out of line with our general arguments, and is included to obtain a unique solution that is tractable. Thus, our proposal is not optimal; we have only empirical evidence that it is far superior to both direct and composite estimation. Further, decisions based on composite estimation are poorer than on direct estimation.

5. Simulations

For a simulation study, we form a computer version of the country, with its districts and within-district rates of illiteracy. We define a sampling design: stratified simple random sampling with the districts as the strata and sampling fractions slightly greater in the smallest districts than in the largest, with some variation to make the setting more realistic. The overall sample size is 20 000, and the average within-district sample sizes are in the range 21 – 1030. In the sampling design, these sample sizes are not fixed and involve some moderate randomness. The population rates of illiteracy within the districts are in the range 2 – 29%, set by a random process in which some prior information is used. These rates are fixed across replications. The rates are smaller in the largest and some of the smallest districts, and are highest for a few mid-size districts. The population is fixed (not altered) in the replications of a simulation study; the sampling process is the sole source of randomness. However, we conduct a large number of simulation studies, with different populations that are plausible in the considered setting, to check that the findings are replicated across studies.

A replication of the simulation comprises drawing a sample from the (fixed, artificially generated) population, and applying the three estimators, direct (D), composite (C) and policy-related (P). The errors of the two kinds are then summarised for each estimator (applied in 72 districts) and the losses added up. The population size of a district is reflected in these summaries by multiplying the loss, when an error in classifying the district is committed, by its population size (in millions). From $M = 1000$ replications, we obtain 1000 triplets of losses and compare their averages. This exercise is repeated for several values of R, which influence only the policy-related estimator; the direct and composite estimators do not depend on R. However, R is a factor in evaluating the average loss even for the direct and composite estimators. More detail can be obtained by evaluating the average losses within the two groups of districts: those deserving the intervention (15 districts) and the complement (57 districts).

Table 1 presents the results of one set of 1000 replications. For each replication, we record the numbers of districts with errors of the two kinds, the total population in them and the total of the losses scaled by the district-level population sizes. These summaries are evaluated for the three estimators and several penalty ratios R, indicated at the left-hand margin. For example, for $R = 10$, the policy-related estimator (P) generates false negatives (F–) for 1.484 districts on average (out of 15) and false positives (F+) for 12.346 districts (out of 57), that is, 13.830 in total. Their respective populations (numbers of adults) are 0.278 and 4.419 million on average, 4.697 million in total. Judging by these two totals, the policy-related estimator is inferior to both the direct and composite estimators, which have errors in approximately $4.4 + 5.3 = 9.7$ and $7.4 + 2.4 = 9.8$ districts, respectively, both involving only 2.9 million adults. However, on the criterion of expected loss, the direct and composite estimators are far inferior to the policy-related estimator. The former two have expected losses 0.495 and 0.711, respectively, whereas the expected loss of the policy-related estimator is only 0.283. The policy-related estimator has greater expected loss on false positives, but the other estimators have excessive expected losses on false negatives.

Table 1: The average number of districts, population and the loss associated with incorrect decisions; 1000 replications.

		Districts		Population		Loss		
R	Estimator	F–	F+	F–	F+	F–	F+	Total
1	P	4.380	5.312	1.035	1.923	0.040	0.057	0.097
	D	4.395	5.295	1.029	1.899	0.043	0.062	0.105
	C	7.387	2.447	1.649	1.247	0.068	0.029	0.097
5	P	1.926	10.072	0.418	3.581	0.077	0.139	0.216
	D	4.395	5.295	1.029	1.899	0.217	0.062	0.279
	C	7.387	2.447	1.649	1.247	0.341	0.029	0.370
10	P	1.484	12.346	0.278	4.419	0.097	0.186	0.283
	D	4.395	5.295	1.029	1.899	0.433	0.062	0.495
	C	7.387	2.447	1.649	1.247	0.682	0.029	0.711
25	P	0.704	15.162	0.138	5.532	0.123	0.256	0.379
	D	4.395	5.295	1.029	1.899	1.083	0.062	1.145
	C	7.387	2.447	1.649	1.247	1.705	0.029	1.734
50	P	0.606	17.314	0.102	6.414	0.162	0.317	0.479
	D	4.395	5.295	1.029	1.899	2.165	0.062	2.227
	C	7.387	2.447	1.649	1.247	3.410	0.029	3.439

Notes: The estimators are: P: policy-related; D: direct; C: composite. The components of loss are: F–: false negative; F+: false positive.

For greater penalty ratios R, the expected loss of the policy-related estimator is greater, but for the other two estimators it increases at a much faster rate. For $R = 50$, their expected losses are 0.47, 2.23 and 3.44. For $R = 1$, the policy-related and composite estimators have the same expected loss, 0.097, and the direct estimator has a slightly higher expected loss, 0.105. However, even for slightly higher R, the policy-related estimator has the smallest expected loss, followed by the direct and composite estimators. Note that $R = 1$ is not equivalent to MSE, because no loss is incurred when the appropriate decision is made, even when the estimate differs from the target substantially, so long as it is in the direction in which the decision is not altered.

In the implementation in the statistical language for computing and graphics R (R Core Development Team, 2012), such a simulation takes about 10 seconds, and so a wide variety of settings can be explored. In some situations, fewer than 1000 replications would suffice for comparisons with a high level of certainty, but the saving in computing is negligible. The key to efficient processing of the results is a partially automated assessment of the results and their compact tabular and graphical presentation.

In particular, the simulations can be re-run with different sample sizes, to explore whether the Ministry's funds could be allocated to the conduct of the survey and the implementation of the programme more effectively. A greater sample size requires higher expenditure, but the allocation of the remainder is then associated with smaller expected losses. It turns out that a sample size around $n = 35\,000$ would lead to a near-optimal split of the resources, although the calculation involved is based on some further assumptions which make some of the existing assumptions more onerous.

The issue of possibly insufficient funds can be explored similarly. Although it uncovers some weaknesses of the policy-related estimator, it remains far superior to its established competitors. The problem is that by erring on the side of inappropriate inclusion in the programme, the expenditure on its implementation increases, and the awards to the selected districts have to be reduced.

The composite and policy-related estimators have their multivariate versions in which auxiliary information from other variables is exploited. In further simulations, we generated such information as the same survey from the previous year. The bivariate composite estimator has substantially smaller expected loss than the univariate composite estimator, but it is still inferior in some settings to the direct estimator. The improvement of the policy-related estimator, based on a multivariate version of the shrinkage in equation (2), is somewhat more modest, but it remains far superior to the direct and composite estimators, except for R very close to unity, such as $R = 1.05$. Even when we get the value of R wrong, say by 50% in either direction, that is, we conduct the estimation with R, but evaluate the losses with $1.5R$ or $R/1.5$, the policy-related estimator is superior to the direct estimator, although the expected losses are appreciably greater than they would be with the assumed value of R. Thus, a modicum of uncertainty about R is acceptable, but there are obvious rewards for its more precise specification by a narrower plausible range.

In summary, Table 1 shows that no single estimator or method is superior by all criteria, so it is essential to specify the criterion that best describes the perspective and priorities of the client. The penalty ratio R is a key quantity in this regard. The policy-related estimation allows some leeway, and it suffices to declare a range of plausible values of R. In the application, this range was set to $(20,30)$, and the survey design was based on $R = 25$. In more detailed exploration of the results, we may find strong points of the composite estimator but on the principal criterion of minimum expected loss it is a total failure; it is inferior even to the direct estimator.

6. Analysts and clients

National statistical institutes and their principal clients, the national government departments and agencies, have over the recent decades negotiated a code of conduct that would ensure the efficient functioning of the institutes without the client exercising any undue influence on the outcomes of the assigned tasks. Transparency and unbiasedness of the institutes, sometimes interpreted as absence of any political influence, are highly prized by the public and are generally regarded as essential elements of the proper use of statistics by government agencies. This mode of operation is appropriate for the production of inferential statements that are effective without any uncertainty, such as national unemployment rates (based on national Labour Force Surveys), consumer price indices, and the like.

Our example of planning an intervention in the districts, for which the uncertainty about key quantities is nontrivial and has to be reckoned with, indicates that such a detached mode of cooperation between the client and the analyst is not conducive to good practice of statistics. Much closer cooperation and integration of the two sets of activities, analytical and decision making, is required. The process of elicitation implies such an integration, even though it is contrary to the current trends in which transparent detachment of the two parties is paramount. The analyst has to be privy to the details of how the estimates or other inferential statements are going to be applied, and under what conditions, to actually choose an appropriate estimator and, more generally, a format of the inferential statement and the assessment of its quality (the expected loss).

The argument that the client could deal with the decision-theoretical aspects of the inferential task without the analyst's assistance does not hold water. These evaluations entail nontrivial optimisation that is firmly in the remit of computational statistics, and for which the national institutes are, or should be, equipped much better than the client. These evaluations cannot be replaced by postprocessing the results obtained by established methods and presented in a standard format.

We see the resolution of this problem in altering the professional ethos in statistics and bringing it much closer to the standard of the (corporate) legal profession. Their standard of ‘representing the best interests of the client’ should be translated to the statistical profession as

> representing the best interests as regards data and information, their collection and all intermediate processes leading to decisions based on them.

Transparency is not disregarded in this process, because the loss functions, the quantified versions of the government priorities and perspectives, can and should be declared openly, as a matter of course by a transparent government. Uncertainty about them is not necessarily a sign of poor management or lack of control over the processes in the remit of the client (the government). On the contrary, it may be an indication of its integrity. Denial of such uncertainty is a sign of poor understanding of its relevance in statistical inference, sometimes combined with misplaced concerns about the false image of the government as an omniscient body.

Acknowledgements

This paper was presented at the First South European Survey Methodology Conference in Barcelona, Spain, in December 2013. By agreement with the author, the identity of the client (country and its government) cannot be disclosed. The person representing the client has approved the text of this paper. His and his colleagues’ cooperation and understanding for the authors’ academic perspective are appreciated. Constructive comments of the journal referees are acknowledged.

REFERENCES

CARLIN, B. P., LOUIS, T. A. (2009). *Bayes and Empirical Bayes Methods for Data Analysis*, 3rd ed., Boca Raton, FL: Chapman and Hall/CRC.

DEGROOT, M. H. (2004). *Optimal Statistical Decisions*, New York: McGraw-Hill.

ELBERS, C., LANJOUW, J. O., LANJOUW, P. (2003). Micro-level estimation of poverty and inequality. *Econometrica*, 71, 355–364.

GARTHWAITE, P. H., KADANE, J. B., O'HAGAN, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100, 680–701.

GOLDSTEIN, H. (2002). *Multilevel Statistical Models*, 3rd ed., London: E. Arnold.

LINDLEY, D. V. (1985). *Making Decisions*, Chichester, UK: Wiley.

LONGFORD, N. T. (1999). Multivariate shrinkage estimation of small-area means and proportions. *Journal of the Royal Statistical Society* Series A, 162, 227–245.

LONGFORD, N. T. (2005). *Missing Data and Small-Area Estimation*, New York: Springer-Verlag.

LONGFORD, N. T. (2007). On standard errors of model-based small-area estimators. *Survey Methodology*, 33, 69–79.

LONGFORD, N. T. (2013). *Statistical Decision Theory*, New York: Springer-Verlag.

LONGFORD, N. T. (2015). Policy-related small-area estimation. *South African Journal of Statistics* 49, 105–119. Also available as Working Paper 1427, Departament d'Economia i Empresa, Universitat Pompeu Fabra, Barcelona, Spain, http://www.econ.upf.edu/en/research/onepaper.php?id=1427.

R CORE DEVELOPMENT TEAM (2012). R: *A language for statistical computing and graphics*, Vienna, Austria: Foundation for Statistical Computing.

RAO, J. N. K. (2003). *Small Area Estimation*, New York: Wiley.

ROBBINS, H. (1995). An empirical Bayes approach to statistics. *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, pp. 157–164.

SÄRNDAL, C.-E., SWENSSON, B., WRETMAN, J. (1992). *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SHEN, W., and LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society* Series B, 60, 455–471.

Appendix

This appendix derives the policy-related estimator. Suppose $\hat{\theta}_d$ is normally distributed with expectation γ_d and variance v_d^2; $\hat{\theta}_d$ may be biased for θ_d. Denote $\tilde{z} = (\gamma_d - T)/v_d$ and $\tilde{z}^\dagger = (\gamma_d - \theta_d)/v_d$. We use ϕ and Φ for the density and the distribution function of the standard normal distribution.

For the piecewise constant loss, the expected losses for the false positives (when $\theta_d < T$) and false negatives (when $\theta_d > T$) are

$$\begin{aligned} Q_+ &= \frac{1}{v_d}\int_T^{+\infty} \phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y = \Phi(\tilde{z}_d) \\ Q_- &= \frac{R}{v_d}\int_{-\infty}^{T} \phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y = R\left\{1-\Phi(\tilde{z}_d)\right\}, \end{aligned}$$

respectively. The corresponding identities for the piecewise quadratic loss function are

$$\begin{aligned} Q_+ &= \frac{1}{v_d}\int_T^{+\infty} (y-\theta_d)^2\,\phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y \\ &= v_d^2\left\{\left(1+\tilde{z}_d^{\dagger 2}\right)\Phi(\tilde{z}_d) + \left(2\tilde{z}_d^{\dagger}-\tilde{z}_d\right)\phi(\tilde{z}_d)\right\} \\ Q_- &= \frac{R}{v_d}\int_{-\infty}^{T} (y-\theta_d)^2\,\phi\left(\frac{y-\gamma_d}{v_d}\right) \mathrm{d}y \\ &= Rv_d^2\left[\left(1+\tilde{z}_d^{\dagger 2}\right)\left\{1-\Phi(\tilde{z}_d)\right\} - \left(2\tilde{z}_d^{\dagger}-\tilde{z}_d\right)\phi(\tilde{z}_d)\right\}. \end{aligned}$$

These identities are obtained by integration by parts. Under the condition of equilibrium at T, $\tilde{z}_d = \tilde{z}_d^\dagger$, and we have the equations

$$\begin{aligned} \Phi(\tilde{z}_d) &= \frac{R}{R+1} \\ (R+1)\left\{\left(1+\tilde{z}_d^2\right)\Phi(\tilde{z}_d) + \tilde{z}_d\,\phi(\tilde{z}_d)\right\} &= R\left(1+\tilde{z}_d^2\right) \end{aligned}$$

for the respective constant and quadratic loss. The former equation has an explicit solution for $\tilde{z}_d$, while the latter is solved by the Newton method. Denote the solution of the relevant equation by $\tilde{z}_d^*$. The results in Table 1 are based on the quadratic loss. The MSE of $\tilde{\theta}_d$ is

$$\mathrm{MSE}\left(\tilde{\theta}_d;\theta_d\right) = (1-b_d)^2\,v_d^2 + b_d^2\,(F_d-\theta_d)^2\,.$$

We replace the square $(F_d - \theta_d)^2$ by its average over the districts d, to eliminate the target θ_d. We refer to this operation as averaging. It results in the identity

$$\mathrm{aMSE}\left(\tilde{\theta}_d;\theta_d\right) = (1-b_d)^2\,\nu_d^2 + b_d^2\left\{\sigma_{\mathrm{B}}^2 + (F_d-\theta)^2\right\}.$$

Equilibrium at $\theta_d = T$ is satisfied when

$$F_d = T + \frac{|1-b_d|}{b_d}\,\tilde{z}_d^*\nu_d\,;$$

then the estimator is

$$\tilde{\theta}_d = (1-b_d)\,\hat{\theta}_d + b_d T + \tilde{z}_d^*\,|1-b_d|\,\nu_d$$

and its aMSE is

$$\begin{aligned}\mathrm{aMSE}\left(\tilde{\theta}_d;\theta_d\right) &= (1-b_d)^2\left(1+\tilde{z}_d^{*2}\right)\nu_d^2 + b_d^2\left\{\sigma_{\mathrm{B}}^2 + (T-\theta)^2\right\} \\ &\quad + 2b_d\,|1-b_d|\,(T-\theta)\tilde{z}_d^*\nu_d\,.\end{aligned}$$

This quadratic function of b_d attains an extreme when

$$b_d^* = \frac{\nu_d^2\left(1+\tilde{z}_d^{*2}\right) - \mathrm{sign}\left(1-b_d^*\right)\,(T-\theta)\tilde{z}_d^*\nu_d}{\nu_d^2 + \sigma_{\mathrm{B}}^2 + \left\{\tilde{z}_d^*\nu_d - \mathrm{sign}\left(1-b_d^*\right)\,(T-\theta)\right\}^2}\,,$$

and it can be shown that one of the two solutions is the unique minimum. For further details, see Longford (2015).

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 83–96

APPLICATION OF BOX-JENKINS METHOD AND ARTIFICIAL NEURAL NETWORK PROCEDURE FOR TIME SERIES FORECASTING OF PRICES

Abhishek Singh[1], **G. C. Mishra**[2]

ABSTRACT

Forecasting of prices of commodities, especially those of agricultural commodities, is very difficult because they are not only governed by demand and supply but also by so many other factors which are beyond control, such as weather vagaries, storage capacity, transportation, etc. In this paper time series models namely ARIMA (Autoregressive Integrated Moving Average) methodology given by Box and Jenkins has been used for forecasting prices of Groundnut oil in Mumbai. This approach has been compared with ANN (Artificial Neural Network) methodology. The results showed that ANN performed better than the ARIMA models in forecasting the prices.

Key words: forecasting, feed forward network, ARIMA, ANN.

1. Introduction

Price forecasting is very essential for planning and development. Therefore, it has become pertinent to develop methods which help the policy makers to have some idea about the prices of commodities in the future. There are various approaches to forecast prices such as using econometric methods which use economic theory and cause and effect relationships to forecast prices of essential commodities. These approaches require a large amount of information regarding different variables which may lead to various types of errors. The time series approach to forecasting is an approach which relies on the assumption that the past pattern in a time series will be repeated in the future and this information can be used to forecast prices. There are many methods for analyzing a time series but one of the most simple and benchmark method is that of Box and Jenkins (1970) which is popularly known as ARIMA methodology. De Gooijer and Hyndman

[1] Asstt. Prof., Department of Farm Engineering; Institute of Agricultural Sciences; Banaras Hindu University, Varanasi, India-221005. E-mail: asbhu2006@gmail.com.

[2] Prof., Department of Farm Engineering; Institute of Agricultural Sciences; Banaras Hindu University, Varanasi, India-221005. E-mail: gcmishrabhu@gmail.com.

(2006) provided an excellent review of time series methods in forecasting. Numerous studies have shown that this univariate method is very effective when compared to some other multivariate methods like linear regression and vector autoregressive models. The problem with ARIMA methodology is that it assumes a linear structure of the process the realization of which is a particular times series, which is often not correct. The other important aspect is that ARIMA methodology is only suitable under the assumption that the time series is stationary. To overcome this limitation of the ARIMA methodology, Artificial Neural Networks (ANN) have also been used to forecast the prices as shown by Kohzadi Nowrouz et al. (1996), Tang et al. (1991) and Zoua et al. (2007). This is because Artificial Neural Networks do not make any assumption about the process from which a particular time series has generated. Therefore, Artificial Neural Networks effectively cover both linear and non-linear processes, stationary as well as non-stationary time series. Neural Networks are now being used in wide domain of studies in areas as diverse as finance, medicine, engineering, geology and physics. This tremendous success of the Artificial Neural Networks can be attributed to some of its distinct character such as its power to model extremely complex function, in particular the non-linear functions. They can also handle the problem of parsimony in linear models. Combination of forecasts also increases the forecasting abilities of different methods as suggested in studies by Newbold et al. (1974), Zhang (2003). With the availability of sophisticated software, fitting of non-linear equations with the help of non-parametric methods has evolved to a new level. Neural Networks have been effective at forecasting and prediction in a variety of scenarios, Adya et al. (1998). Chen et al. (1992) and Park et al. (1991) found that for forecasting electric load ANN was better than traditional approaches. Tang et al. (1991) used ANN for forecasting car sales and airline passenger data and reported that ANN outperformed Box-Jenkins approach, both for short-term and long-term forecasting. Agricultural processes are affected by typical factors which are unique to this sector and prediction of prices of agricultural commodities is very difficult because they are not only governed by demand and supply also by so many other factors which are beyond control such as weather vagaries, storage capacity, transportation, etc. The performance of ANN in agricultural scenario is relatively less explored.

Therefore, in this paper time series of prices of Groundnut oil in Mumbai from January 1994 to July 2010 has been analyzed with both the ARIMA methodology and artificial neural networks and the forecasting abilities of both the models have been compared.

Rest of the paper is organized as follows - in Section 2 the traditional univariate time series approach to forecasting is described. In Section 3 the neural network architecture that is designed for this study is discussed. Section 4 discusses the evaluation methods for comparing the two forecasting approaches. Data and forecast procedure are discussed in Section 5. Moreover, section 5 shows the results obtained from the ARIMA and the Artificial Neural Network estimations. Section 6 contains a comparison of applied statistical measures and conclusions.

2. Auto Regressive Integrated Moving Average (ARIMA) time series model

Introduced by Box and Jenkins (1970), the ARIMA model has been one of the most popular approaches for forecasting. In the ARIMA model, the estimated value of a variable is supposed to be a linear combination of the past values and the past errors. Generally, a non-seasonal time series can be modelled as a combination of past values and errors, which can be denoted as ARIMA (p,d,q) which is expressed in the following form:

$$X_t = \theta_0 \ +\Phi_1 X_{t-1}+ \Phi_2 \ X_{t-2} +\ldots\ldots\ldots.+ \Phi_p X_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} \ -\ldots\ldots.- \theta \ e_{t-q} \qquad (1)$$

where X_t and e_t are the actual values and random error at time t, respectively, Φ_i (i = 1,2,…….,p) and θ_j (j = 1,2,……,q) are model parameters. p and q are integers and often referred to as orders of autoregressive and moving average polynomials respectively. Random errors e_t are assumed to be independently and identically distributed with mean zero and the constant variance, σ_e^2. Similarly, a seasonal model is represented by **ARIMA (p,d,q) x (P,D,Q)** model, where P denotes number of seasonal autoregressive (SAR) terms, D denotes number of seasonal differences, Q denotes number of seasonal moving average (SMA) terms. Basically, this method has three phases: model identification, parameters estimation and diagnostic checking.

The ARIMA model is basically a data oriented approach that is adapted from the structure of the data itself.

3. Artificial Neural Network (ANN) model

Neural Networks are simulated networks with interconnected simple processing neurons which aim to mimic the function of the brain central nervous system. ANN closely mimics functioning of the brain so its architecture is similar to that of the brain. A biological neuron has three types of components, namely dendrite, soma and axon. The dendrite accepts signals from other neurons which are electrical impulses transmitted through a synaptic gap with the help of certain chemical processes. A biological network is a collection of many biological neurons. Similarly, ANN is characterized by its architecture, i.e. the pattern of connections between the neurons, the method of determining the weights of the connections i.e. training or learning algorithm and its activation function. Mcculloch and Pitts (1943) for the first time proposed the idea of the artificial neural network but because of the lack of computing facilities they were not in much use until the back propagation algorithm was discovered by Rumelhart et al. in 1986. Neural networks are good at input and output relationship modelling even for noisy data. The greatest advantage of a neural network is its ability to model complex non-linear relationship without a priori assumptions of the nature of the relationship. Apart from this, artificial neural networks can also be used for classification problems as was shown by Ripley (1994).

The ANN model performs a non-linear functional mapping from the past observations (X_{t-1}, $X_{t-2,}$.………., X_{t-p}) to the future value X_t i.e.

$$X_t = f\,(X_{t-1}\,, X_{t-2,}\ldots\ldots\ldots.., X_{t-p,}w\,) + e_t \tag{2}$$

where w is a vector of all parameters and f is a function determined by the network structure and connection weights.

Training of the Neural Network is an essential factor for the success of the neural networks and among the several learning algorithms available, back propagation has been the most popular and most widely implemented learning algorithm of all neural networks paradigms. The important task of the ANN modelling for a time series is to choose an appropriate number of hidden nodes, q, as well as the dimensions of the input vector p (the lagged observations). However, in practice the choices of q and p are difficult.

4. Criteria for comparing the prediction accuracy of ARIMA and ANN procedures

Different criteria will be used to make comparisons between the forecasting ability of the ARIMA time series models and the neural network models. The first criterion is the absolute mean error (AME). It is a measure of average error for each point forecast made by the two methods. AME is given by

$$AME = \left({}^{1}/_{T}\right)\sum|P_t - A_t| \tag{3}$$

The second criterion is the mean absolute percent error (MAPE). It is similar to AME except that the error is measured in percentage terms, and therefore allows comparisons in units which are different.

The third criterion is the mean square error (MSE) which measures the overall performance of a model. The formula for MSE is

$$\text{MSE} = \left({}^{1}/_{T}\right)\sum(P_t - A_t)^2 \tag{4}$$

where P_t is the predicted value for time t, A_t is the actual value at time t and T is the number of predictions and the fourth criterion is RMSE which is the square root of MSE.

5. Results

Monthly cash prices of groundnut oil in Mumbai from April 1994 to July 2010 are used to test the prediction power of the two approaches. Data are obtained from the official Website of Ministry of Consumer Affairs. An ARIMA model was estimated. The model was then used to forecast on its respective three month out-of-sample set.

In the case of the neural networks, the time series was divided into a training, testing, and a validation (out-of-sample) set. The out-of-sample period was identical to the ARIMA model. SPSS (Statistical package for social sciences) was used to analyze the data and to carry out the calculations.

5.1. ARIMA time series results

Data is first differenced in order to remove the trend and the ARIMA estimated. For estimating the ARIMA model the three stages of modelling as suggested by Box and Jenkins namely identification, estimation and diagnostic checking were undertaken. Identification was done after examining the autocorrelation function and the partial autocorrelation function. After that, estimation of the model was done by the least square method. In the diagnostic checking phase the model residual analysis was performed.

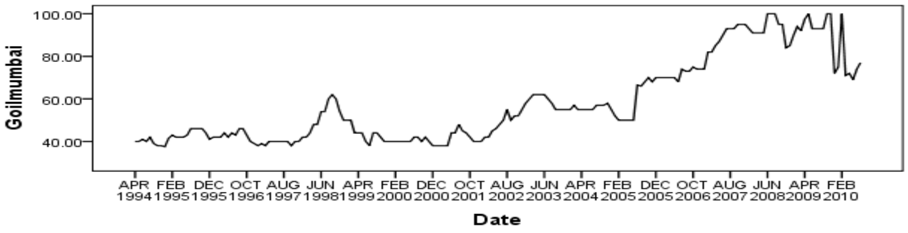


Figure 1. The time plot of prices of the Groundnut oil in Mumbai

In Figure 1 the time plot prices of the Groundnut oil in Mumbai is given. By looking at the graph it can be inferred that the series is not stationary because the mean of the time series is increasing with the increase in time. However, to confirm this autocorrelation function was also observed.

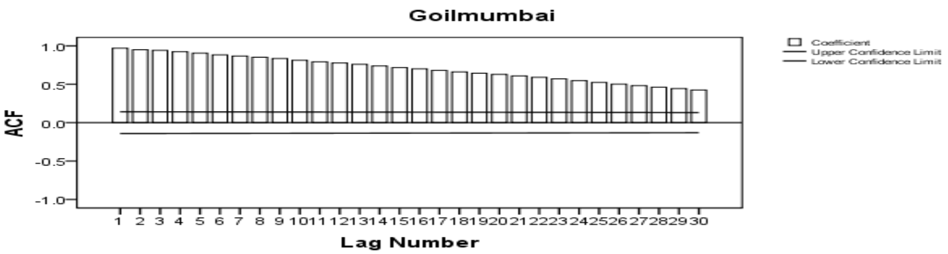


Figure 2. The autocorrelation function of the time series

In Figure 2 the autocorrelation function of the time series is shown. It certainly shows that the series is not stationary because autocorrelation coefficient does not cut off to statistical insignificance enough quickly which is caused by the fact that autocorrelations are significantly greater than the ± 2/√N confidence limits at 5% level of significance up to the 30th lag. To make the series stationary it was differenced.

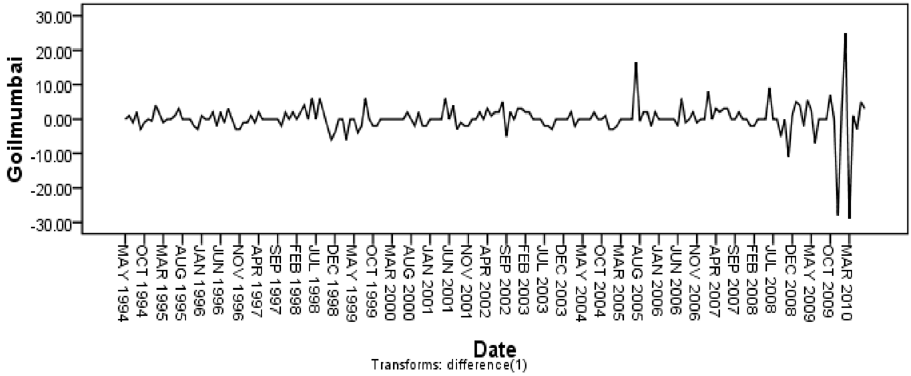


Figure 3. The time plot of the differenced series

In Figure 3 the time plot of the differenced series is given. It clearly shows that the series has now become mean stationary. However, it is not a variance stationary since the variance of the data around the mean of the differenced series in the end is greater than the rest of the series. Therefore, log transformation of the data was done.

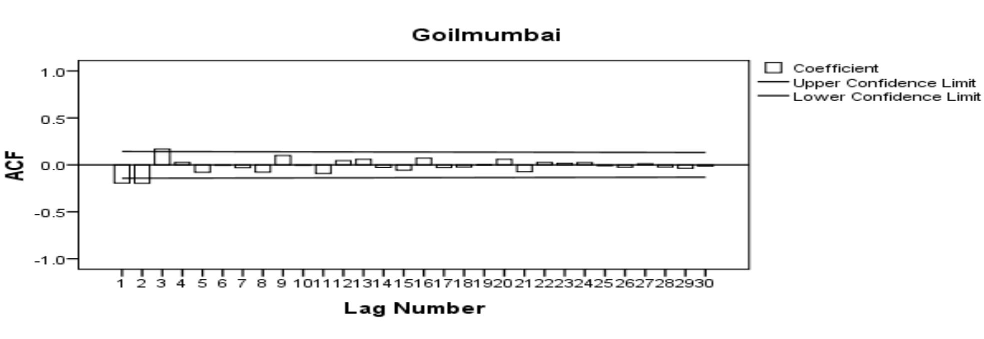


Figure 4. Autocorrelation function (ACF) of the differenced series

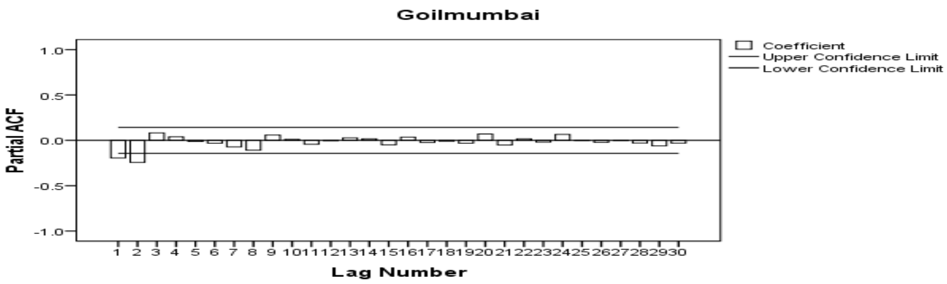


Figure 5. Partial autocorrelation function (PACF) of the differenced series

In Figure 4 and Figure 5 autocorrelation function (ACF) and partial autocorrelation function (PACF) of the differenced series are shown. Various ARIMA models were fitted using the expert modeler option in the SPSS software and after going through these stages the ARIMA (0,1,0) (1,0,1) model was found to be the best among the family of ARIMA models. ARIMA model parameters and model fit statistics are given in Table 1. The estimates of both the AR seasonal Lag 1 and MA seasonal Lag 1 were found to be statistically significant.

Table 1. ARIMA model parameters and model fit statistics

	Estimate	SE	t	Sig	Model Fit Statistics	
Differencing	1				Stationary R Squared	0.041
					R Squared	0.951
AR Seasonal Lag 1	0.990	0.091	10.841	0.00	RMSE	4.327
					MAPE	3.707
MA Seasonal Lag 1	0.953	0.231	4.127	0.00	MAE	2.215
					Normalized BIC	2.985

At the diagnostic checking stage residuals were examined and the autocorrelation coefficients were found to be non-significant (Figure 6).This shows that the model is satisfactory.

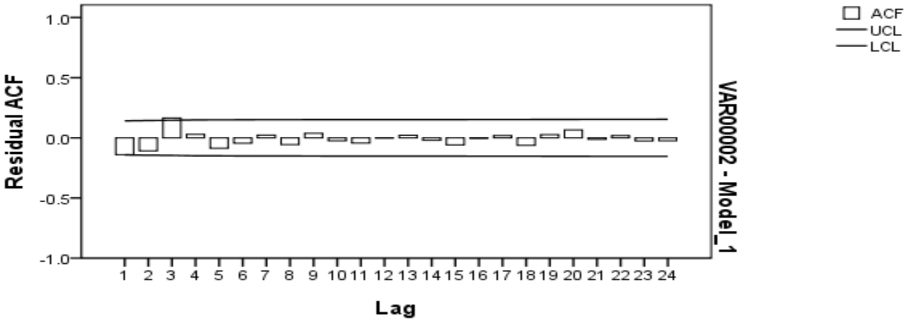


Figure 6. Autocorrelation function of the residuals

5.2. Neural network results

A feed forward neural network was fitted to the data, where values of the time series at first, second and third lag were taken as independent variables and the value to be forecasted was the dependent variable. The data was divided into 3 sets training, testing and hold out. In Table 2 it is shown that 81.6% observations were used for training, 16.8% for testing and 1.5% for forecasting. The training set was used for the estimation of the weights in the neural network and then predictions were made in the testing set. On the basis of the error in the testing set, the weights of the neural network were again adjusted to minimize the errors in the testing set.

Table 2. Case processing summary

		N	Percent
Sample	Training	160	81.6%
	Testing	33	16.8%
	Holdout	3	1.5%
Valid		196	100.0%
Excluded		0	
Total		196	

The information about the neural network architecture is given in Table 3. It shows that the network has an input layer, a single hidden layer and an output layer. In the hidden layer there is 1 unit and the activation function used is the hyperbolic tangent.

Table 3. Network architecture

Input layer	Covariates No. of units Rescaling methods of covariates	Lag1, Lag2, Lag3 3 Standardized
Hidden Layers	No. of hidden layers No. of units in hidden layers Activation Function	1 1 Hyperbolic tangent
Output Layer	Dependent variables Number of units Rescaling methods for scale dependents Activation function Error function	1 1 Standardized Identity Sum of squares

The architecture of the network has been shown in Figure 7. Light colour lines show weights greater than zero and the dark colour lines show weight less than zero.

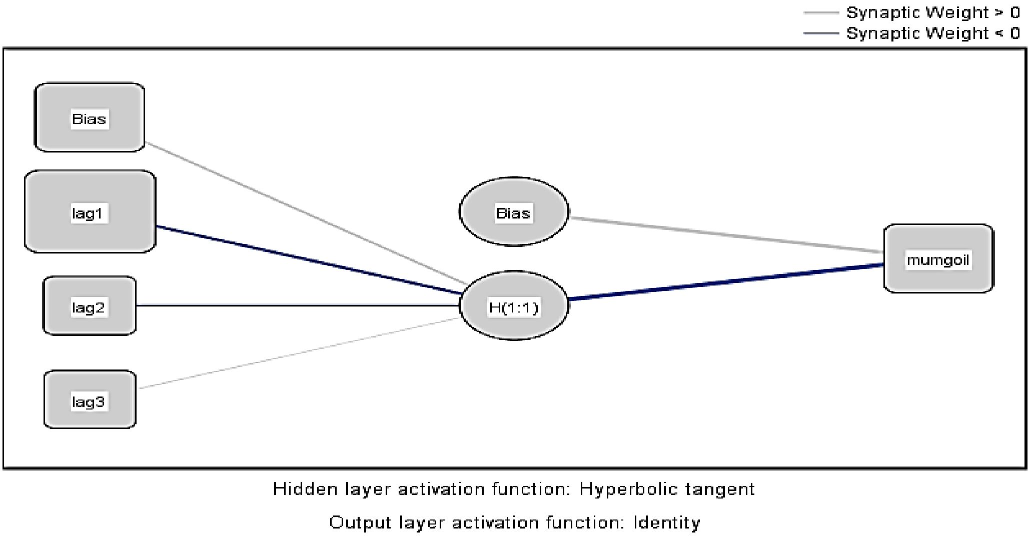

Synaptic Weight > 0
Synaptic Weight < 0
Bias
lag1
lag2
lag3
Bias
H(1:1)
mumgoil
Hidden layer activation function: Hyperbolic tangent
Output layer activation function: Identity

Figure 7. The architecture of the network

The training summary and the fit statistics for the training, testing and the holdout sets are given in Table 4.

Table 4. Model summary

Training	Sum of squares error Relative error Stopping rule used	3.480 0.044 Maximum number of epochs(100000) Exceeded
Testing	Sum of squares error Relative error	7.253 1.048
Holdout	Relative error	0.291

The estimates of the weights and bias are given in Table 5. The results in this table show the value of weights from input to the hidden layer and from the hidden layer to the output layer. H (1:1) means hidden layer 1 and 1St neuron. The weight attached to the neuron from bias is .354, from lag 1 is -.356 from lag 2 is -.045 and from lag 3 is .042.

The weights from the hidden layer to the output layer for bias 1.024 and from 1st neuron in the hidden layer to the output is -3.188.

Table 5. Parameter estimates

Predictor		Predicted	
		Hidden Layer 1	Output Layer
		H(1:1)	mumgoil
Input Layer	(Bias)	.354	
	lag1	-.356	
	lag2	-.045	
	lag3	.042	
Hidden Layer 1	(Bias)		1.024
	H(1:1)		-3.188

The observed values and the predicted graph in the Figure 8 show that except for few outliers it is a straight line.

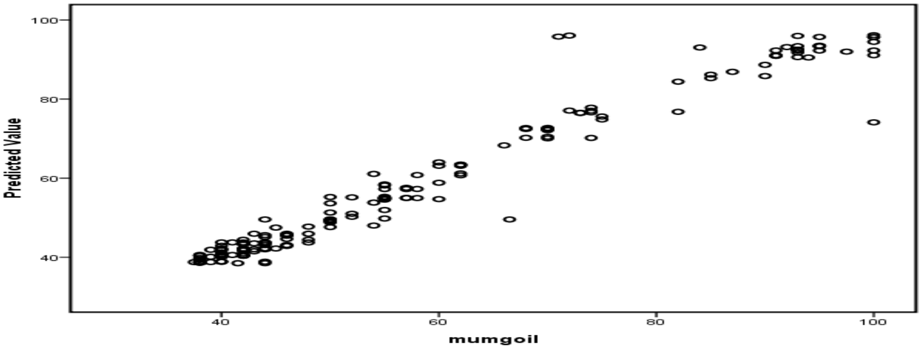


Figure 8. The observed values and the predicted values

It indicates almost one to one correspondence among the observed and predicted values. Hence, it can be inferred that the performance of ANN is satisfactory.

The residual and predicted chart (Figure 9) also shows that the residual does not follow a definite pattern and therefore is not correlated. If there is no dependence among the residuals then they can be regarded as observations of independent random variables and show that the ANN is satisfactory.

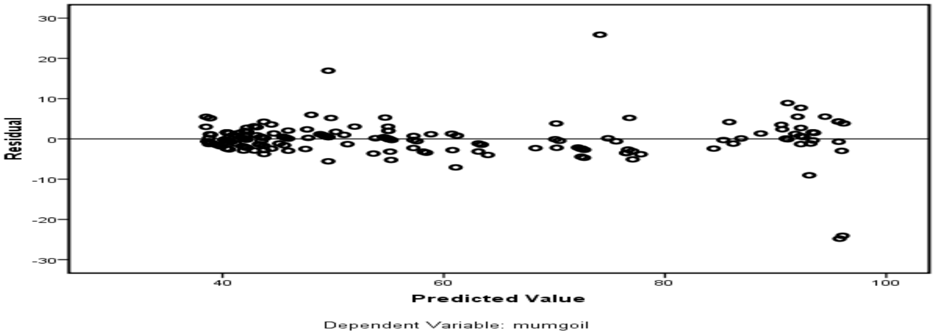


Figure 9. The residual and the predicted

6. Comparison of the accuracy of models and conclusions

The ARIMA and ANN models were compared for their forecasting capabilities with the help of RMSE and MSE. The results are shown below in Table 6.

The one step ahead forecast for May 2010 (69) was best predicted by ANN model (70.63) followed by the forecast of the ARIMA model (72.63).

The two step ahead forecast for June 2010 (74) was best predicted by ARIMA model (73.24) followed by the forecast of the ANN model (71.46).

The three steps ahead forecast for July 2010 (77) was best predicted by ANN model (76.36) followed by the forecast of the ARIMA model (74.91)

Overall, the forecast by ANN model was found to be the best predicted with MAPE (2.21), RMSE (3.09), MSE (9.52) followed by the forecast by the ARIMA model with MAPE (3.00), RMSE (4.26), MSE (18.12).

Table 6. Comparison of the accuracy of models

Observed		Predicted	
		ARIMA	ANN
May 2010	69	72.63	70.63
June 2010	74	73.24	71.46
July2010	77	74.91	76.36
	MSE	18.12	9.52
	RMSE	4.26	3.09
	MAPE	3.00	2.21

Artificial neural networks performed considerably better than the ARIMA models showing the forecasting ability and accuracy of this approach. The mean squared error (MSE), root mean square error (RMSE) and mean absolute percent error (MAPE) were all lower on average for the neural network forecast than for the ARIMA. The reason the neural network model performed better than the ARIMA may be because the data shows chaotic behaviour, which cannot be fully captured by the linear ARIMA model. Finally, the neural network results conform to the theoretical proofs that a feed forward neural network with only one hidden layer can precisely and satisfactorily approximate any continuous function.

REFERENCES

ADYA, M., COLLOPY, F., (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. Journal of Forecasting, 17, 481−495.

BOX, G. E. P., JENKINS, G. M., (1976). Time Series Analysis: Forecasting and Control. revised ed. Holden-Day, San Francisco.

CHEN, S. T., YU, D. C., MOGHADDAMJO, A. R., (1992). Weather sensitive short-term load forecasting using nonfully connected artificial neural network. IEEE Transactions on Power Systems, 7, 3,1098−1105.

DE GOOIJER, J. G., HYNDMAN, R. J., (2006). 25 years of time series forecasting. International Journal of Forecasting. Elsevier, Vol. 22(3), 443−473.

KOHZADI, N., BOYD, M. S., KAASTRA, I., KERMANSHAHI, B. S., (1996). A comparison of artificial neural network and time series models for forecasting commodity price Neurocomputing, 10, 169−18.

MCCULLOCH, W. S., PITTS, W., (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115−133.

NEWBOLD, P., GRANGER C. W. J., (1974). Experience with forecasting univariate time series and the combination of forecasts. Journal of the Royal Statistical Society, A 137, 131−165.

PARK, D. C., EL-SHARKAWI, M. A., MARKS II, R. J., ATLAS, L. E., DAMBORG, M. J., (1991). Electric load forecasting using an artificial neural network, IEEE Transactions on Power Systems, 62, 442−449.

RIPLEY, B., (1994). Neural Networks and Related Methods for Classification (with discussion). Journal of the Royal Statistical Society, B, 56, 409−456.

RUMELHART, D. E., HINTON, G. E., WILLIAMS, R. J., (1986). Learning Internal Representations by Error Propagation, in Parallel Distributed Processing: Exploration in the Microstructure of Cognition. Cambridge, MA: MIT Press., Vol. 1, 318−362.

TANG, Z., DE ALMEIDA, C., FISHCWICK, P. A., (1991). Time series forecasting using neural networks vs. Box Jenkins methodology. Simulation,57, 5, 303−310.

Website of Ministry of Consumer Affairs:
http://fcainfoweb.nic.in/PMSver2/Reports/Report_Menu_web.aspx

ZHANG, G. P., (2003). Times series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159−75.

ZOUA, H. F., XIAA, G. P., YANGC, F. T., WANGA, H. Y., (2007). An investigation and comparison of artificial neural network and time series models for Chinese food grain price forecasting. Neurocomputing, 70, 2913−2923.

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 97–110

CLASSIFICATION PROBLEMS BASED ON REGRESSION MODELS FOR MULTI-DIMENSIONAL FUNCTIONAL DATA

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

ABSTRACT

Data in the form of a continuous vector function on a given interval are referred to as multivariate functional data. These data are treated as realizations of multivariate random processes. We use multivariate functional regression techniques for the classification of multivariate functional data. The approaches discussed are illustrated with an application to two real data sets.

Key words: multivariate functional data, functional data analysis, multivariate functional regression, classification.

1. Introduction

Much attention has been paid in recent years to methods for representing data as functions or curves. Such data are known in the literature as functional data (Ramsay and Silverman (2005)). Applications of functional data can be found in various fields, including medicine, economics, meteorology and many others. In many applications there is a need to use statistical methods for objects characterized by multiple features observed at many time points (doubly multivariate data). Such data are called multivariate functional data. The pioneering theoretical work was that of Besse (1979), in which 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). In this paper we focus on the problem of classification via regression for multivariate functional data. Functional regression models have been extensively studied; see for example James (2002), Müller 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: wolynski@amu.edu.pl.

Stadmüller(2005), Reiss and Ogden (2007), Matsui et al. (2008) and Li et al. (2010). Various basic classification methods have also been adapted to functional data, such as linear discriminant analysis (Hastie et al. (1995)), logistic regression (Rossi et al. (2002)), penalized optimal scoring (Ando (2009)), $k$nn (Ferraty and Vieu (2003)), SVM (Rossi and Villa (2006)), and neural networks (Rossi et al. (2005)). Moreover, the combining of classifiers has been extended to functional data (Ferraty and Vieu (2009)).

In the present work we adapt multivariate regression models to the classification of multivariate functional data. We focus on the binary classification problem. There exist several techniques for extending the binary problem to multi-class classification problems. A brief overview can be found in Krzyśko and Wołyński (2009). The accuracy of the proposed methods is demonstrated using biometrical examples. Promising results were obtained for future research.

2. Classification problem

The classical classification problem involves determining a procedure by which a given object can be assigned to one of K populations based on observation of p features of that object.

The object being classified can be described by a random pair $(\boldsymbol{X}, Y)$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)' \in \mathbf{R}^p$ and $Y \in \{0, 1, \ldots, K-1\}$.

The optimum Bayesian classifier then takes the form (Anderson (1984)):

$$d(\boldsymbol{x}) = \arg \max_{k \in \{0,1,\ldots,K-1\}} \mathrm{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}).$$

We shall further consider only the case $K = 2$. Here

$$d(\boldsymbol{x}) = \begin{cases} 1, & \mathrm{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) \geqslant \mathrm{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x}); \\ 0, & \mathrm{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) < \mathrm{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x}). \end{cases}$$

We note that

$$\mathrm{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \mathrm{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = r(\boldsymbol{x}),$$

where $r(\boldsymbol{x})$ is the regression function of the random variable Y with respect to the random vector $\boldsymbol{X}$.

Hence

$$d(\boldsymbol{x}) = \begin{cases} 1, & r(\boldsymbol{x}) \geqslant 1/2; \\ 0, & r(\boldsymbol{x}) < 1/2. \end{cases}$$

3. Functional data

We now assume that the object being classified is described by a p-dimensional random process $\boldsymbol{X} = (X_1, X_2, ..., X_p)' \in L_2^p(I)$, where $L_2(I)$ is the Hilbert space of square-integrable functions.

Let $\boldsymbol{x}$be the realization of the random process $\boldsymbol{X}$. Moreover, assume that the k th component of the vector $\boldsymbol{x}$ can be represented by a finite number of orthonormal basis functions $\{\varphi_b\}$

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

where $c_{k0}, c_{k1}, \ldots, c_{kB_k}$ are the unknown coefficients.

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

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

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

Then, the vector of the continuous function $\boldsymbol{x}$ at point $t$can be represented as

$$\boldsymbol{x}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c}. \tag{2}$$

We can estimate the vector $\boldsymbol{c}$ on the basis of n independent realizations $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ of the random process $\boldsymbol{X}$ (functional data).

Typically data are recorded at discrete moments in time. Let x_{kj} denote an observed value of the feature X_k, $k = 1, 2, \ldots p$ at the jth time point t_j, where $j = 1, 2, ..., J$. Then our data consist of the pJ pairs (t_j, x_{kj}). These discrete data can be smoothed by continuous functions x_k and I is a compact set such that $t_j \in I$, for $j = 1, ..., J$.

Details of the process of transformation of discrete data to functional data can be found in Ramsay and Silverman (2005) or in Górecki et al. (2014).

4. Regression analysis for functional data

We now consider the problem of the estimation of the regression function $r(\boldsymbol{x})$.

Let us assume that we have an n-element training sample

$$\boldsymbol{x}(t) = \boldsymbol{\Phi}(t)\boldsymbol{c}. \tag{3}$$

where $\boldsymbol{x}_i \in L_2^p(I)$ and $y_i \in \{0,1\}$.

Analogously as in section 3, we assume that the functions $\boldsymbol{x}_i$ are obtained as the result of a process of smoothing n independent discrete data pairs (t_j, x_{kij}), $k = 1, \ldots, p$, $j = 1, ..., J$, $i = 1, ..., n$.

Thus the functions $\boldsymbol{x}_i$ at point t have the following representation:

$$\boldsymbol{x}_i(t) = \boldsymbol{\Phi}(t)\boldsymbol{c}_i,\ i = 1, 2, \ldots, n. \tag{4}$$

4.1. Multivariate linear regression. We take the following model for the regression function:

$$r(\boldsymbol{x}) = \beta_0 + <\boldsymbol{\beta}, \boldsymbol{x}> = \beta_0 + \int_I \boldsymbol{\beta}'(t)\boldsymbol{x}(t)dt.$$

We seek the unknown parameters in the regression function by minimizing the sum of squares

$$S(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \int_I \boldsymbol{\beta}'(t)\boldsymbol{x}_i(t)dt)^2.$$

We assume that the functions $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$ have the representation (4). We adopt an analogous representation for the p-dimensional weighting function $\boldsymbol{\beta}$, namely

$$\boldsymbol{\beta}(t) = \boldsymbol{\Phi}(t)\boldsymbol{d}, \tag{5}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$,and $\boldsymbol{d} = (d_{10}, \ldots, d_{1B_1}, \ldots, d_{p0}, \ldots, d_{pB_p})'$.

Then

$$\begin{aligned} \int_I \boldsymbol{\beta}'(t)\boldsymbol{x}_i(t)dt &= \int_I \boldsymbol{d}'\boldsymbol{\Phi}'(t)\boldsymbol{\Phi}(t)\boldsymbol{c}_i dt \\ &= \boldsymbol{d}' \int_I \boldsymbol{\Phi}'(t)\boldsymbol{\Phi}(t)dt\boldsymbol{c}_i = \boldsymbol{d}'\boldsymbol{c}_i, \quad i = 1, 2, \ldots, n. \end{aligned}$$

Hence

$$S(\beta_0, \boldsymbol{\beta}) = S(\beta_0, \boldsymbol{d}) = \sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{d}'\boldsymbol{c}_i)^2.$$

We define $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ and

$$\boldsymbol{Z} = \begin{bmatrix} 1 & \boldsymbol{c}_1' \\ 1 & \boldsymbol{c}_2' \\ \vdots & \vdots \\ 1 & \boldsymbol{c}_n' \end{bmatrix}, \quad \boldsymbol{\gamma} = \begin{bmatrix} \beta_0 \\ \boldsymbol{d} \end{bmatrix}.$$

Then

$$S(\beta_0, \boldsymbol{\beta}) = S(\boldsymbol{\gamma}) = (\boldsymbol{y} - \boldsymbol{Z\gamma})'(\boldsymbol{y} - \boldsymbol{Z\gamma}).$$

Minimizing the above sum of squares leads to the choice of a vector $\boldsymbol{\gamma}$ satisfying

$$\boldsymbol{Z}'\boldsymbol{Z\gamma} = \boldsymbol{Z}'\boldsymbol{y}. \tag{6}$$

Provided the matrix $\boldsymbol{Z}'\boldsymbol{Z}$ is non-singular, equation (6) has the unique solution

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}. \tag{7}$$

In the case of functional data we may use the smoothed least squares method (Ramsay and Silverman (2005)), that is, we minimize the sum of squares in the form

$$S(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \beta_0 - \int_I \boldsymbol{\beta}'(t)\boldsymbol{x}_i(t)dt)^2 + \lambda \int_I [L\boldsymbol{\beta}(t)]'[L\boldsymbol{\beta}(t)]dt,$$

where L denotes the linear differential operator. Assuming $L\boldsymbol{\beta} = D^2\boldsymbol{\beta}$, we obtain

$$L\boldsymbol{\beta}(t) = D^2(\boldsymbol{\Phi}(t)\boldsymbol{d}) = D^2(\boldsymbol{\Phi}(t))\boldsymbol{d}.$$

Thus

$$\int_I [L\boldsymbol{\beta}(t)]'[L\boldsymbol{\beta}(t)]dt = \boldsymbol{d}' \int_I [D^2\boldsymbol{\Phi}(t)]'[D^2\boldsymbol{\Phi}(t)]dt\, \boldsymbol{d}.$$

We define

$$\boldsymbol{R} = \int_I [D^2\boldsymbol{\Phi}(t)]'[D^2\boldsymbol{\Phi}(t)]dt.$$

Hence

$$\int_I [L\boldsymbol{\beta}(t)]'[L\boldsymbol{\beta}(t)]dt = \boldsymbol{d}'\boldsymbol{R}\boldsymbol{d} = \boldsymbol{\gamma}'\boldsymbol{R}_0\boldsymbol{\gamma},$$

where

$$\boldsymbol{R}_0 = \begin{bmatrix} 0 & \boldsymbol{0}' \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}.$$

Then

$$S(\beta_0, \boldsymbol{\beta}) = S(\boldsymbol{\gamma}) = (\boldsymbol{y} - \boldsymbol{Z\gamma})'(\boldsymbol{y} - \boldsymbol{Z\gamma}) + \lambda\boldsymbol{\gamma}'\boldsymbol{R}_0\boldsymbol{\gamma}.$$

Minimizing the above sum of squares leads to the choice of a vector $\boldsymbol{\gamma}$ satisfying the equation

$$(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{R}_0)\boldsymbol{\gamma} = \boldsymbol{Z}'\boldsymbol{y}.$$

The equation thus obtained has the form

$$\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{R}_0)^{-1}\boldsymbol{Z}'\boldsymbol{y}. \tag{8}$$

From this we obtain the following form for the estimator of the regression function for the multivariate functional data:

$$\hat{r}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\boldsymbol{d}}'\boldsymbol{c},$$

where $\hat{\boldsymbol{\gamma}} = (\hat{\beta}_0, \hat{\boldsymbol{d}})'$ is given by the formula (7) or (8).

4.2. Functional logistic regression. We adopt the following logistic regression model for functional data:

$$r(\boldsymbol{x}) = \frac{\exp(\beta_0 + < \boldsymbol{\beta}, \boldsymbol{x} >)}{1 + \exp(\beta_0 + < \boldsymbol{\beta}, \boldsymbol{x} >)} = \frac{\exp(\beta_0 + \int_I \boldsymbol{\beta}'(t)\boldsymbol{x}(t)dt)}{1 + \exp(\beta_0 + \int_I \boldsymbol{\beta}'(t)\boldsymbol{x}(t)dt)}. \tag{9}$$

Using the representation of the function $\boldsymbol{x}$ given by (2) and the weighting function $\boldsymbol{\beta}$ given by (5) we reduce (9) to a standard logistic regression model in the form

$$r(\boldsymbol{x}) = \frac{exp(\beta_0 + \boldsymbol{d}'\boldsymbol{c})}{1 + exp(\beta_0 + \boldsymbol{d}'\boldsymbol{c})}.$$

To estimate the unknown parameters of the model, we use the training sample $\mathcal{L}_n$ and the analogous representation for the functions $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$ given by (4).

Thus we obtain the following form for the estimator of the regression function

$$\hat{r}(\boldsymbol{x}) = \frac{\exp(\hat{\beta}_0 + \hat{\boldsymbol{d}}'\boldsymbol{c})}{1 + \exp(\hat{\beta}_0 + \hat{\boldsymbol{d}}'\boldsymbol{c})}.$$

4.3. Local linear regression smoothers. We consider the problem of nonparametric estimation of a regression function $r(\boldsymbol{x})$ from a sample (3).

Let $\boldsymbol{x}_0$ be a fixed and known point in the space $L_2^p(I)$.

Using Taylor series, we can approximate $r(\boldsymbol{x})$, where $\boldsymbol{x}$ is close to a point $\boldsymbol{x}_0$, as follows:

$$r(\boldsymbol{x}) \approx r(\boldsymbol{x}_0) + < \frac{\partial r(\boldsymbol{x}_0)}{\partial \boldsymbol{x}_0}, \boldsymbol{x} - \boldsymbol{x}_0 >= \beta_0 + < \boldsymbol{\beta}, \boldsymbol{x} - \boldsymbol{x}_0 >, \tag{10}$$

where

$$\beta_0 = r(\boldsymbol{x}_0), \quad \boldsymbol{\beta} = \frac{\partial r(\boldsymbol{x}_0)}{\partial \boldsymbol{x}_0}.$$

This is a local polynomial regression problem in which we use the data to estimate the polynomial which best approximates $r(\boldsymbol{x})$ in a small neighborhood around the point $\boldsymbol{x}_0$, i.e. we minimize it with respect to β_0 and $\boldsymbol{\beta}$ in the function

$$S(\beta_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \int_I \boldsymbol{\beta}'(t)(\boldsymbol{x}_i(t) - \boldsymbol{x}_0(t))dt \right)^2 K(\frac{1}{h}||\boldsymbol{x}_i - \boldsymbol{x}_0||).$$

This is a weighted least squares problem where the weights are given by the kernel functions $K(||\boldsymbol{x}_i - \boldsymbol{x}_0||/h)$.

Analogously as in the previous sections, suppose that the vector functions $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are in the same space, i.e.

$$\boldsymbol{x}_i(t) = \boldsymbol{\Phi}(t)\boldsymbol{c}_i, \; i = 1, 2, \ldots, n, \quad \boldsymbol{\beta}(t) = \boldsymbol{\Phi}(t)\boldsymbol{d}.$$

Then

$$\int_I \boldsymbol{\beta}'(t)(\boldsymbol{x}_i(t) - \boldsymbol{x}_0(t))dt = \boldsymbol{d}'(\int_I \boldsymbol{\Phi}'(t)\boldsymbol{\Phi}(t)dt)(\boldsymbol{c}_i - \boldsymbol{c}_0) = \boldsymbol{d}'(\boldsymbol{c}_i - \boldsymbol{c}_0),$$

$$||\boldsymbol{x}_i - \boldsymbol{x}_0|| = \sqrt{(\boldsymbol{c}_i - \boldsymbol{c}_0)'(\boldsymbol{c}_i - \boldsymbol{c}_0)}, \; i = 1, 2, \ldots, n.$$

The least squares problem is then to minimize the weighted sum-of-squares function

$$S(\beta_0, \boldsymbol{\beta}) = S(\beta_0, \boldsymbol{d}) = \sum_{i=1}^{n} (y_i - \beta_0 - \boldsymbol{d}'(\boldsymbol{c}_i - \boldsymbol{c}_0))^2 K(\frac{1}{h}\sqrt{(\boldsymbol{c}_i - \boldsymbol{c}_0)'(\boldsymbol{c}_i - \boldsymbol{c}_0)})$$

with respect to the parameters β_0 and $\boldsymbol{d}$.

It is convenient to define the following vectors and matrices:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{Z} = \begin{bmatrix} 1 & (\boldsymbol{c}_1 - \boldsymbol{c}_0)' \\ 1 & (\boldsymbol{c}_2 - \boldsymbol{c}_0)' \\ \vdots & \vdots \\ 1 & (\boldsymbol{c}_n - \boldsymbol{c}_0)' \end{bmatrix},$$

$$\boldsymbol{\gamma} = \begin{bmatrix} \beta_0 \\ \boldsymbol{d} \end{bmatrix}, \quad \boldsymbol{W} = \begin{bmatrix} K_1 & 0 & \ldots & 0 \\ 0 & K_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & K_n \end{bmatrix},$$

where

$$K_i = K(\frac{1}{h}\sqrt{(\boldsymbol{c}_i - \boldsymbol{c}_0)'(\boldsymbol{c}_i - \boldsymbol{c}_0)}),\ i = 1, 2, \ldots, n.$$

The least squares problem is then to minimize the function

$$S(\beta_0, \boldsymbol{\beta}) = S(\boldsymbol{\gamma}) = (\boldsymbol{y} - \boldsymbol{Z\gamma})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{Z\gamma}) \,.$$

The solution is

$$\hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{y}$$

provided $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z}$ is a non-singular matrix.

As in the case of multivariate functional linear regression model we can also include an additional smoothing component. Then, we seek the unknown parameter $\boldsymbol{\gamma}$ by minimizing the sum of squares

$$S(\gamma) = (\boldsymbol{y} - \boldsymbol{Z}\gamma)' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{Z}\gamma) + \lambda\gamma'\boldsymbol{R}_0\gamma.$$

Provided the matrix $\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \lambda\boldsymbol{R}_0$ is non-singular we have the unique solution

$$\hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \lambda\boldsymbol{R}_0)^{-1}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{y}.$$

The $r(\boldsymbol{x}_0)$ is than estimated by the fitted intercept parameter (i.e. by $\hat{\boldsymbol{\beta}}_0$) as this defines the position of the estimated local polynomial curve at the point $\boldsymbol{x}_0$. By varying the value of $\boldsymbol{x}_0$, we can build up an estimate of the function $r(\boldsymbol{x})$ over the range of the data.

We have

$$\hat{r}(\boldsymbol{x}_0) = \boldsymbol{e}'\hat{\gamma},$$

where the vector $\boldsymbol{e}$ is of the length $B_1 + \cdots + B_p + p + 1$ and has a 1 in the first position and 0's elsewhere.

4.4. Nadaraya-Watson kernel estimator. In Section 4.3 we approximated the regression function $r(\boldsymbol{x})$ using Taylor series. In the approximation (10) let us take into account only the first term, i.e.

$$r(\boldsymbol{x}) \approx r(\boldsymbol{x}_0) = \beta_0.$$

Then

$$S(\beta_0) = \sum_{i=1}^{n}(y_i - \beta_0)^2 K(\frac{1}{h}||\boldsymbol{x}_i - \boldsymbol{x}_0||).$$

Minimizing the above sum of squares leads to the kernel estimator of the regression function $r(\boldsymbol{x})$ of the form

$$\hat{r}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} y_i K_i}{\sum_{i=1}^{n} K_i},$$

where $K_i = K(\frac{1}{h}\sqrt{(\boldsymbol{c}_i - \boldsymbol{c}_0)'(\boldsymbol{c}_i - \boldsymbol{c}_0)})$, $i = 1, 2, \dots, n$.

This gives us a well-known kernel estimator proposed by Nadaraya and Watson (1964).

5. Examples

Experiments were carried out on two data sets, these being labelled data sets whose labels are given. The data sets originate from Olszewski (2001).The *ECG* data set uses two electrodes (Figure 1) to collect data during one heartbeat. Each heartbeat is described by a multivariate time series (MTS) sample with two variables and an assigned classification of normal or abnormal. Abnormal heartbeats are representative of a cardiac pathology known as supraventricular premature beat. The *ECG* data set contains 200 MTS samples, of which 133 are normal and 67 are abnormal. The length of an MTS sample is between 39 and 152.

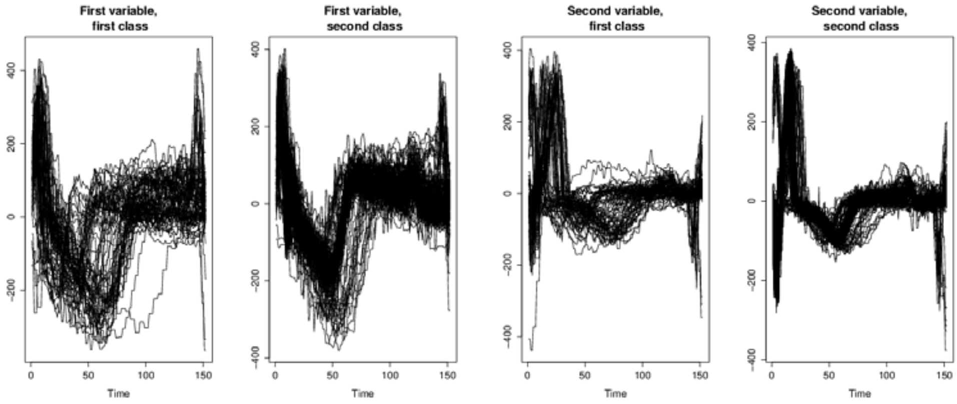


Figure 1. Variables of the extended *ECG* data set.

The *Wafer* data set uses six vacuum-chamber sensors (Figure 2) to collect data while monitoring an operational semiconductor fabrication plant. Each wafer is described by an MTS sample with six variables and an assigned classification of normal or abnormal. The data set used here contains 327 MTS samples, of which 200 are normal and 127 are abnormal. The length of an MTS sample is between 104 and 198.

The multivariate samples in the data sets are of different lengths. For each data set, the multivariate samples are extended to the length of the longest multivariate sample in the set (Rodriguez et al. (2005)). We extend all variables to the same length. For a short univariateinstance x with length J, we extend it to a long instance x_{ex} with length J_{max} by setting

$$x_{ex}(t_j) = x(t_i), \quad \textsf{for} \quad i = \left\lceil \frac{j-1}{J_{\mathsf{max}}-1}(J-1)+0.5 \right\rceil \quad (j = 1, 2, \ldots, J_{\mathsf{max}}).$$

Some of the values in a data sample are duplicated in order to extend the sample. For instance, if we wanted to extend a data sample of length 75 to a length of 100, one out of every three values would be duplicated. In this way, all of the values in the original data sample are contained in the extended data sample.

For the classification process, we used the classifiers described above. For each data set we calculated the classification error rate using the leave-one-out cross-validation method (LOO CV). Table 1 contains the results of the classification error rates (in %).

Table 1. Classification error (in %)

Model	**ECG**	**Wafer**
Multivariate functional linear regression	11.50	0.59
Functional logistic regression	11.50	0.17
Local linear regression smoothers	16.50	0.67
Nadaraya-Watson kernel estimator	20.50	10.64

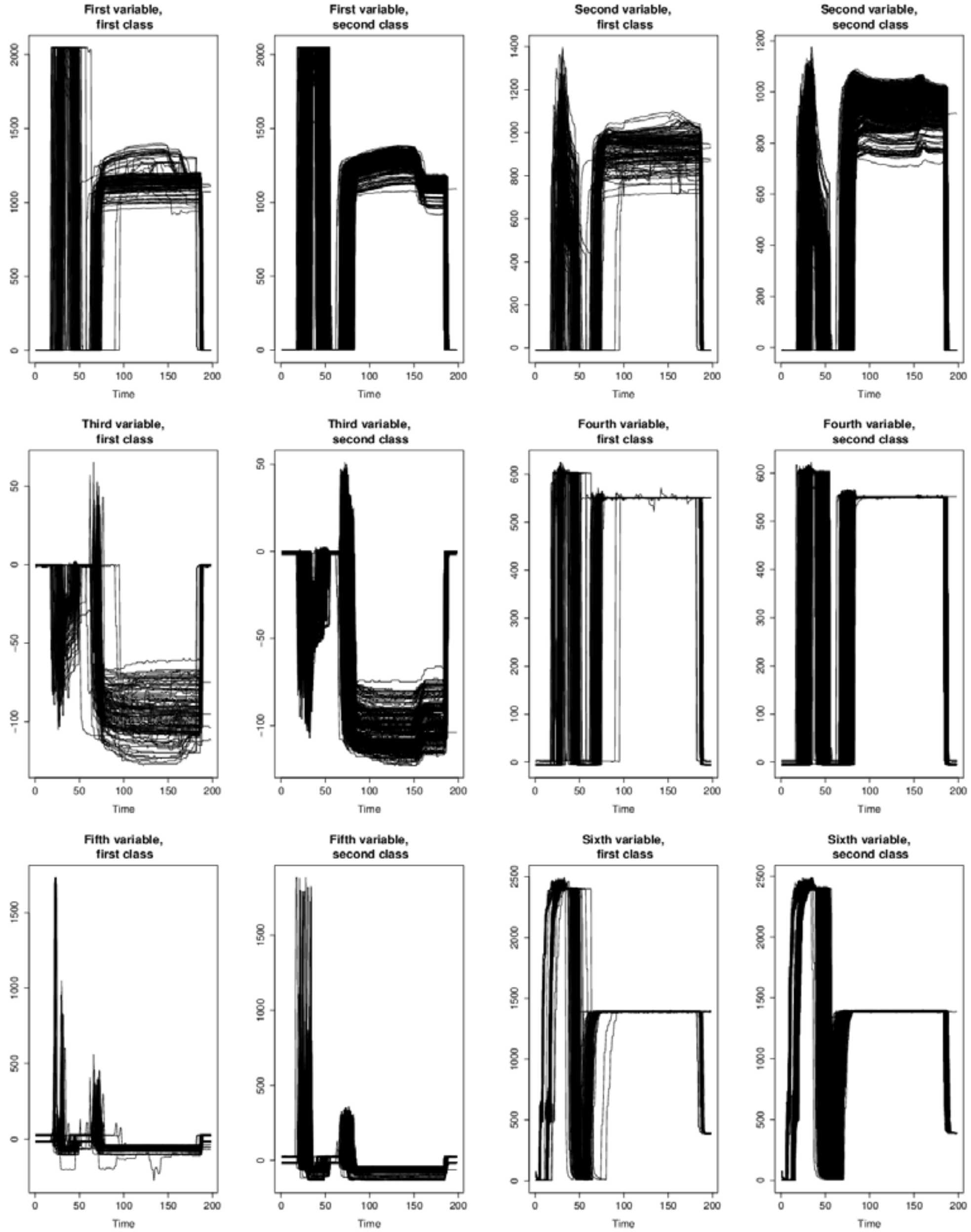


Figure 2. Variables of the extended *Wafer* data set.

From Table 1 we see that the *ECG* data set is difficult to recognize. None of the four regression methods can deal with it well. In contrast, the data set *Wafer* is easily recognizable. For this set of data definitely the best results are given by a functional logistic regression. We also see a big difference between the local linear regression smother, and the Nadaraya-Watson kernel estimator.

6. Conclusion

This paper develops and analyzes methods for constructing and using regression methods of classification for multivariate functional data. These methods were applied to two biometrical multivariate time series. In the case of these examples it was shown that the use of multivariate functional regression methods for classification 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 other regression methods, such as partial least squares regression – PLS (Wold (1985)), least absolute shrinkage and selection operator – LASSO (Tibshirani (1996)), or least-angle regression – LARS (Efron et al. (2004)), to the multivariate functional case. This will be the direction of our future research.

REFERENCES

ANDERSON, T. W., (1984). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

ANDO, T., (2009). Penalized optimal scoring for the classification of multi-dimensional functional data. Statistcal Methodology 6, 565–576.

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

EFRON, B., HASTIE, T., JOHNSTONE, I., TIBSHIRANI, R., (2004). Least Angle Regression. Annals of Statistics 32(2), 407–499.

FERRATY, F., VIEU, P., (2003). Curve discrimination. A nonparametric functional approach. Computational Statistics & Data Analysis 44, 161–173.

FERRATY, F., VIEU, P., (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer, New York.

FERRATY, F., VIEU, P., (2009). Additive prediction and boosting for functional data. Computational Statistics & Data Analysis 53(4), 1400–1413.

GÓRECKI, T., KRZYŚKO, M., (2012). Functional Principal Components Analysis. In: J. Pociecha and R. Decker (Eds.): Data analysis methods and its applications. C. H. Beck, Warszawa, 71–87.

GÓRECKI, T, KRZYŚKO, M., WASZAK, Ł., WOŁYŃSKI, W., (2014). Methods of reducing dimension for functional data. Statistics in Transition new series 15, 231–242.

HASTIE, T. J., TIBSHIRANI, R. J., BUJA, A., (1995). Penalized discriminant analysis. Annals of Statistics 23, 73–102.

JAMES, G. M., (2002). Generalized linear models with functional predictors. Journal of the Royal Statistical Society 64(3), 411–432.

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

KRZYŚKO, M., WOŁYŃSKI, W., (2009). New variants of pairwise classification. European Journal of Operational Research 199(2), 512–519.

MATSUI, H., ARAKI, Y., KONISHI, S., (2008). Multivariate regression modeling for functional data. Journal of Data Science 6, 313–331.

MÜLLER, H. G., STADMÜLLER, U., (2005). Generalized functional linear models. Annals of Statistics 33, 774–805.

NADARAYA, E. A., (1964). On Estimating Regression. Theory of Probability and its Applications 9(1), 141–142.

OLSZEWSKI, R. T., (2001). Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data. Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA.

RAMSAY, J. O., SILVERMAN, B. W., (2005). Functional Data Analysis. Springer, New York.

REISS, P. T., OGDEN R. T., (2007). Functional principal component regression and functional partial least squares. Journal of the American Statistcal Assosiation 102(479), 984–996.

ROSSI, F., DELANNAYC, N., CONAN-GUEZA, B., VERLEYSENC, M., (2005). Representation of functional data in neural networks. Neurocomputing 64, 183–210.

ROSSI, F., VILLA, N., (2006). Support vector machines for functional data classification. Neural Computing 69, 730–742.

ROSSI, N., WANG, X., RAMSAY, J. O., (2002). Nonparametric item response function estimates with EM algorithm. Journal of Educational and Behavioral Statistics 27, 291–317.

RODRIGUEZ, J. J., ALONSO, C. J., MAESTRO, J. A., (2005). Support vector machines of intervalbased features for time series classification. Knowledge-Based Systems 18, 171–178.

SAPORTA, G., (1981). Méthodes exploratoires d'analyse de données temporelles, thèse de doctorat d'état es sciences mathématiques soutenue le 10 juin 1981, Université Pierre et Marie Curie.

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

TIBSHIRANI, R., (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58(1), 267–288.

WATSON, G. S., (1964). Smooth regression analysis. Sankhya – The Indian Journal of Statistics, Series A 26(4), 359–372.

WOLD, H., (1985). Partial least squares. In: S. Kotz, and N.L. Johnson (Eds.): Encyclopedia of statistical sciences vol. 6, Wiley, New York, 581–591.

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 111–136

STOCHASTIC GOALS IN FINANCIAL PLANNING FOR A TWO-PERSON HOUSEHOLD

Radosław Pietrzyk[1], **Paweł Rokita**[2]

ABSTRACT

In household financial planning two types of risk are typically being taken into account. These are life-length risk and risk connected with financing. In addition, also various types of events of insurance character, like health deterioration, are sometimes taken into account. There are, however, no models addressing stochastic nature of household financial goals. The last should not be confused with modelling factors that influence performance of financing the goals, which is a popular research topic. The problem of modelling goals themselves is, in turn, not so well explored. There are two main characteristics that describe a goal: magnitude and time. At least for some goals one or both of these characteristics may show a stochastic nature. This article puts forward a proposition of working goal time and magnitude into a household financial plan and taking their distributions into account when optimizing the plan. A model of two-person household is used. The decision variables of the optimization task are consumption-investment proportion and division of household investments between household members.

Key words: financial goals, personal finance, intertemporal choice, financial plan optimization, stochastic goals.

1. Introduction

The aim of this article is to present a concept of a household financial plan optimization model that allows for stochastic character of household goals, in respect of goal realization time and magnitude.

The model assumes that the household maximizes its value function, which is composed of expected discounted utilities of consumption and bequest.

For the financial plan optimization procedure, the value function plays the role of a goal function. This article, however, is not meant to propose an optimization technique. It is rather intended to discuss a concept of how to formulate the problem. Optimization of a financial plan with a number of

[1] Wroclaw University of Economics. E-mail: radoslaw.pietrzyk@ue.wroc.pl.
[2] Wroclaw University of Economics. E-mail: pawel.rokita@ue.wroc.pl.

dynamic stochastic factors is described from operational research perspective by Konicz, et al. (2014), for instance. The last builds on the results by Gayer et al. (2009) and Richard (1975). Also, a multi-person household case was analyzed (Bruhn and Steffensen, 2014). A comprehensive introduction to stochastic programming in finance may be found in the positions by Ziemba (2003) and Vickson and Ziemba (2006). A more general summary of the concept and methods of stochastic programming models is presented by Ruszczyński and Shapiro (2003).

As it has already been mentioned, the main focus is to take account of uncertainty about goal realization, but not in the sense of the question if the household is able to afford them, but rather in the sense of stochastic nature of such goal characteristics as time and magnitude of the need (in financial terms) the goal is expected to satisfy. The subject-matter of this article does not cover the risk of financing. For example, if an investment is meant as a future source of financing for the goal, the market risk of assets used as part of the investment is not in the scope of interest here. Also, if a credit is planned for the realization of the goal, the interest rate risk connected with the credit is not the issue to be discussed in this article.

The model belongs to the area of life-cycle financial planning in personal finance. It uses the basic concepts present in this field of research. In the literature on life-long financial planning a lot of research has been done in the areas of consumption optimization and dynamic portfolio optimization. The first current builds on Modigliani, Brumberg (1954), Ando, Modigliani (1957) and Yaari (1965), whereas the dynamic asset allocation research – on the works by Merton (1969, 1971) and Richard (1975). Further research in life cycle planning with dynamic stochastic properties of asset prices was done by Cox and Huang (1989). An important issue in personal finance is also a trade-off between life insurance and capital-based protection against unexpected events in the area of survival process (Ibbotson et al. 2005, Huang et al., 2008). Also, the problem of two-person household was tackled by personal finance researchers. Kotlikoff and Spivak (1981) investigated the influence of longevity risk sharing between household members (a married couple) on the demand for annuities. Hurd (1999) constructed a two-person generalization of the classical (Yaari 1965) life-cycle model and presented an analytical solution to the consumption optimization problem for a couple. Brown and Poterba (2000) analysed advantages of joint annuities suited to the needs of married couples, with reference to the longevity-risk-sharing effect discussed earlier by Kotlikoff and Spivak (1981). Generally speaking, the current state of the art covers many important aspects of consumption transfer in life-cycle financial planning in the sense by Bodie et al. (2008), that is – in the two directions: between periods (by means of saving and investments) and between optimistic and pessimistic scenarios (by means of insurance, but also risk sharing). The last (considering scenarios) is a way of addressing a more general problem, namely – uncertainty. In this area, the following stochastic factors influencing the financial plan were taken into account: rates of return on investment, length of life (also in a bivariate case), health condition, labour income, and sometimes also damage to physical property. What remains a rather unexplored field is the way financial goals themselves may behave.

It is much more popular to take into consideration the risk of having insufficient means to finance a goal on a specified date (it refers to pre-financing and post-financing) than the risk that, for instance, the time when the household really needs to accomplish this goal will be shifted in time. This article presents a proposition that may serve as a starting point for filling in this gap. For example, an important risk factor influencing the performance of the household financial plan is the time of a child's birth. This may of course differ from a planned or desired time. There is, however, no research on statistical properties of this source of uncertainty in the literature. And, consistently, in personal finance, it does not belong to the set of risk factors that are modelled using statistical methods as part of financial planning. Starting a discussion about conditional distribution of a child's birth, under the condition of planned time, seems a natural step further in the development of personal finance research area.

It seems necessary to explain that the proposition put forward in this article provides a general framework within which more detailed models may be developed. For instance, assumed interdependencies between household financial goals may be in fact different in details, but the provided example is sufficient to outline the general idea. Moreover, for some stochastic elements of the model, only putative properties are discussed, without even suggesting types, nor even general families, of parametric distributions that might be used.

The paper is organized in the following way. Section 2 shortly sums up the basic version of the model, upon which the current concept is developed. Types of financial goals are discussed in section 3. In this section also some definitions and assumptions are presented. Sections 4 and 5 are devoted to the main subject of this article, namely – stochastic character of the goals and interdependencies between them (section 4) and the way these properties may be reflected in value function of the household (section 5). The last section contains conclusions.

2. An outline of the model

The model developed here is based on its basic version presented by Feldman, Pietrzyk and Rokita (2014a, 2014c) and Pietrzyk and Rokita (2014). This is a discrete time, two-person household, life-long financial plan model. In the basic version it assumed only two financial goals: retirement and bequest. Consumption is optimized in the life cycle of the household, both in accumulation phase and in retirement.

The dates of death of the two persons are the only risk factors in the basic version of the model. Plans differ in respect of risk. Some are very immune against even very large deviations of dates of death from their expected values, some other are very sensitive. They are exposed to premature-death risk and longevity risk. In the two-person case, premature-death risks play not less important role than longevity (and this statement remains valid even if the household has no bequest motive).

The only two goals of the basic model are set in different ways. Whereas retirement is set explicitly as a goal to be accomplished, in terms of time and magnitude, the bequest motive is merely declared, only in order to pass the information to the value function of the household if utility of residual wealth is to be calculated or not.

It is assumed that the retirement capital accumulated until retirement date of a given person is fully spent on a life annuity assigned to this person. The household has, yet, a choice whether to invest in building retirement capital of the first, the second or both persons in any proportions.

One of the main features of a two-person household, as compared to a single-individual case, is life-time risk sharing. It allows for building plans with the so-called "partial retirements" – compare *Full-Partial*, *Partial-Full* and *2×Partial* retirement as defined by Feldman, Pietrzyk and Rokita (2014a). The possibility of investing in less than *2×Full* retirement in the sense by Feldman et al. (2014a) broadens the spectrum of possible proportions between consumption and investments the household may choose.

A deterministic growth pattern of consumption, which is usually a constant growth rate in real terms, is assumed. This constitutes the *basic path of consumption*. Upon this, some additional consumption may be put on. In extended versions of the model, going beyond the two-goal framework, it is usually the result of the realization of some goals (like, for instance, new consumption structure connected with a child in the household).

It is also assumed that at the starting time of the plan the household invests the whole part of income which is not consumed (there is no surplus generated). Since the basic version of the model does not assume any other financial goals than retirement, the investment is here understood as investing for retirement. In subsequent periods, because of differences in income and consumption growth rates, there may be some additional unconsumed and uninvested surplus. Saying that the surplus is not invested is a mental shortcut. More precisely – one cannot assume in advance that the surplus will be invested at a high rate (unplanned investments that, in addition, need to be very liquid, because they are used to cover some intermittent shortfalls).

The surplus cumulated until a moment may be used to smoothen consumption path in the next periods. But it must be pointed out that if there is a goal of creating a safety cash reserve to play a similar role (not discussed in this work as a separate type of goal), then this is not treated as a part of a cumulated surplus. It may be created from the surplus, but if it is done on purpose, the reserve disappears from the account of the surplus. Creating it should be treated as goal realization.

The plan is optimized by maximization of the value function of the household, given by equation (2) in section 5. The decision variables are:

- consumption-investment proportion (given by consumption rate in the first period),
- division of investment between persons (given by first-period proportion of Person 1 investment).

All parameters of the model like incomes, returns on investments, growth rates, macroeconomic parameters, etc., are assumed to be revised on a regular basis (at least once a year as recommended). Each plan revision session includes new optimization. This allows avoiding the need of making long-term forecasts. All parameters are assumed to be valid for the whole planning period, but they are updated every year. Plan revisions include also corrections in goal time and magnitude. No modifications of goal structure are made automatically as a result of optimization. They may be introduced only by the decision of the household. The only variables that are changed in the optimization procedure are the two aforementioned decision variables.

As it has already been mentioned, any changes in time or size of the goals due to lack of financing, too risky financing, or any other reasons that are not intrinsic to the goals themselves, are not the subject of the analysis in this work.

3. Financial goals

Besides securing some acceptable life standard for all household members throughout the whole life of the household, households tend to realize yet some ambitions and dreams that, if given some planning rigor to, may be called life objectives. Some of these objectives may be expressed in financial terms and included into a quantitative model as *financial goals*. The aim of a financial plan is maximization of the value function of the household and fulfilling at the same time all constraints, including accomplishment of financial goals. In this article an attempt is made to take into account a stochastic character of time and magnitude with which the goals are realized. Defining the terms “financial goal” and “realization of financial goal” first seems necessary to avoid confusion. It has also to be specified which financial goals will be considered in the discussion, and finally, which of them are to be formally included into the model.

3.1. Definitions and assumptions

> According to the definition of the household by Zalega (2007), a household is *an autonomous economic entity distinguished according to the criterion of individual property, making decisions about consumption on the basis of its preferences and existing constraints*.

Here, in this paper a two-person household is considered. It is understood as a household with two decision makers, called also *main members*, who (at least as far as their predetermination at the moment of plan preparation is concerned) intend to remain members of the household until its end. This does not really exclude cases with more or less members (comp. Feldman, Pietrzyk and Rokita, 2014a, 2014c; Pietrzyk and Rokita, 2014) if one or two main household members are distinguished.

It is assumed that households maximize utility of consumption throughout their whole life cycle. The household has also some life objectives that its members want to accomplish in certain time and to a certain extent. These objectives are kinds of constraints for the consumption utility maximization. But the true time and cost of realization of the objectives is not certain indeed. Here, the focus is on the uncertainty inherent in the objectives themselves, not in the tools of their realization like financing.

From the point of view of the model discussed here the life objectives may be divided into two groups: financial and non-financial. The model is intended to include in a household life-long financial plan those that are of financial nature. They will be further called *financial goals* (comp. ***Definition 1***).

Definition 1. Financial goal

> To provide financial means for covering a negative cash flow or a series of negative cash flows of a substantial value, prearranged as to time and magnitude and isolated from the basic path of consumption (comp. explanation of the term "basic path of consumption" in section 2).

Satisfaction resulting from realization of the financial goals is here identified with the utility of consumption. The financial goals may, of course, give also other kinds of satisfaction to the household, but this is not taken into account in the model.

The most important financial goals may include, for example: purchase of a house, leaving a bequest, covering costs of bringing up children and covering costs of their education.

Even though the main goal of the household is to guarantee a satisfactory life standard to all household members throughout the whole life cycle, vast part of which is realized by means of preserving a desired level of the basic path of consumption. This is why the goal is not distinguished as one of financial goals.

As to the non-financial objectives, they influence decisions of the households in many respects, including financial decisions, but they themselves are not expressed in financial terms. These objectives may cover, for instance, finding a suitable partner, or just broadly understood self-fulfilment in personal and professional area. The realization of non-financial objectives may indirectly modify cash flows of the household, because it may change propensity to consume, motivation to save and invest, etc.

As it has already been mentioned, only financial goals are considered in the model. The model is intended to allow for stochastic character of goal realization in two respects: time and magnitude. Goal realization is understood here as specified by the ***Definition 2***:

Definition 2. Goal realization

A negative cash flow or series of negative cash flows connected with a financial goal.

It is important to avoid confusion between the three related terms:

(a) life objective,
(b) financial goal of providing required financial means at a given time,
(c) goal realization.

A practical difference between these terms may be easily explained by the example of having-a-child objective. Is the time of a child's birth a stimulant, destimulant or nominant in the sense of its influence on the value function of the household? Of course, the time at which the household is (in financial sense) ready to bring up children – comp. (b) financial goal – is a destimulant (the earlier the better). But the time when the child is born and the period of higher consumption starts – comp. (c) goal realization – is a nominant (the closer to the planned time the better). And finally, whether the time of birth is a stimulant, destimulant or nominant from a general life situation perspective (a) is really hard to say and this piece of work does not even attempt to answer this question.

Five types of goals are distinguished in the model. They are treated in different manners. There is a set of characteristics according to which the goals may be grouped. They are: time of realization, goal magnitude in monetary terms, number of cash flows needed to realize the goal. Another important feature is the role of the financial goal in the household wealth. Realization of some goals contributes to the household wealth (like buying a house), whereas other are by their nature more similar to consumption (put differently, their realization is just consumption, but isolated from its basic path).

The distinguished types are:

- Type I – Child(ren),
- Type II – Retirement,
- Type III – House (residential real estate bought in order to live there, not as kind of investment),
- Type IV – Endowment,
- Type V – Bequest.

It is important to emphasize that in this model any investments are treated as tools supporting financing realization of some goals of the household. Investments and goals are two separated categories. Making an investment cannot be a goal. Nevertheless, there is a strong link between the first and the second. Namely, if there are any investments in the model, they are assigned to some corresponding goals. This is, however, not a bijective relation. Two or more investments may be used to support financing of one goal, and also one investment may be used to provide financing of two or more goals (Feldman, Pietrzyk, Rokita, 2014b). Reservations are expressed about the fact that if a purchase of an asset (financial or non-financial) is treated as an investment, then this purchase is not a goal itself. It is rather used to finance some goals. The inverse does not hold, because realizing some types of goals like buying a house for own residential needs, may contribute to the total wealth of the household and then be used to finance some other goals (bequest, for example). In more details the issue is explained in the subsection 3.3.

3.2. Goal comparison

The five types of goals are described below in terms of the following criteria:

- Interpretation – general practical interpretation of the goal,
- Realization – how goal realization is expressed,
- Time and Magnitude of realization – determinants of the main characteristics, namely – time and magnitude (whether they are random or deterministic, if the household controls them or they are out of control by the household, whether they depend on some other factors and if the dependence is of stochastic or deterministic nature),
- Contribution to household wealth – if the goal realization contributes to household wealth or is just a particular kind of consumption (isolated from the basic path of consumption but not different in its nature from consumption),
- Utility – when and for how long (periods/cash flows) utility of goal realization is measured.

The comparison of goal types is presented in the Table 1.

Table 1. Goal types and their characteristics

Goal type	Interpretation	Realization	Time of realization	Magnitude of realization	Contri-bution to household wealth	Utility
Type I (Child)	Being able to provide for additional consumption needs throughout the period when the child is a household member (usu. from birth to becoming independent)	Additional consumption. May be expressed as a percentage or absolute increase in consumption, prevailing throughout the period when the child remains in the household	Main characteristic. Stochastic, but the household declares some planned time	(To a vast extent) a derivative of general standard of life assumed by the household	None	Prolonged. Utility of consumption in many periods.
Type II (Retirement)	Providing financial means for maintaining in the retirement period the	A series of consumption expenditures in retirement period that	Start: legally determined (retirement age) unless the person	At the discretion of the household, declaring	None	Prolonged. Utility of consumption in many periods.

Goal type	Interpretation	Realization	Time of realization	Magnitude of realization	Contri-bution to household wealth	Utility
	standard of life of the pre-retirement phase	are in excess of what can be covered by retirement income from public pension system	dies before her or his retirement age. End: driven by biological factors (survival in the retirement phase of the life cycle)	some required level of consumption in retirement.		
Type III (House)	Providing financial means that make it possible to buy a residential real estate for one own needs (not treated as an investment) on a planned date	Cash outflow connected with the purchase	Decision of the household (but in the model it is dependent also on the time of a child's birth – comp. section 4)	Decision of the household, but exposed to real-estate price risk (in addition, in the model it is dependent also on the number of children – comp. section 4)	Substantial	Here, in the model: a) utility of additional consumption, at the moment of purchase – where the additional consumption is a single negative cash flow, b) contributes to utility of bequest
Type IV (Endowment)	All kinds of donations to a variety of entities, from own children who have become independent (e.g., for university education or first flat), through charity donations, to hobbyist's sponsoring of projects, events or people	Cash flow or a series of cash flows transferred to other entities	At the discretion of the household	At the discretion of the household (or may depend, for instance, on commodity prices if donation of a physical asset is the aim of the household)	None	In the same way as utility of consumption. Goal realization is treated as additional consumption

Goal type	Interpretation	Realization	Time of realization	Magnitude of realization	Contri-bution to household wealth	Utility
Type V (Bequest)	Treated in a different way than other types. Interpreted as wealth at the moment of household end. This includes cumulated surplus and other assets. Particular role in the bequest mass is played by a house or flat bought as realization of type III goal.	Treated in a different way than other types. The bequest is just the wealth of the household passed to the descendants.	Not defined in terms of required magnitude and planned time. Instead, the household declares just bequest motive and chooses the value of its parameter (comp. section 5, subsection 5.1) Time of realization is determined by conditions of biological nature, which is the time of death of both household members. Magnitude of type V goal depends in the model on the household decision, but in a different way than it is for other goals. The declared bequest motive parameter just passes to the value function of the household the information on how important utility of bequest is to the household members.		The bequest is the residual wealth of the household	Utility of bequest (separate function)

3.3. Special cases – goals contributing to household wealth and resulting utility issues

As it has already been mentioned (comp. subsection 3.1), the realization of a goal may have the effect of becoming a kind of investment. This is when goal realization consists in buying an asset that is of a durable nature.

This situation is encountered in the case of the Type II goal, that is – a house. One of the most important characteristics of this goal is that it contributes to household wealth. The real estate remains a part of household fixed assets for a long period (often until the household end). If the house or flat is sold earlier and a new one is bought instead, the new purchase may be treated as a next phase of realization of the same goal or the next goal. There is, yet, a technical question that needs to be answered before accepting such interpretation. As it has already been explained, utility in the model is only utility of consumption and utility of bequest. Utility of consumption attached to the goals is understood as utility of goal realization, where goal realization is a negative cash flow. This is logically consistent because the negative cash flows for goal realization and basic path of consumption (also a series of negative cash flows) are identical in their nature (albeit magnitudes and regularity may differ). In the case of goals that do not contribute to household wealth, no further explanation to that is needed. But if realization of the goal is connected with a purchase of some component of fixed assets, treating it just as consumption is controversial. It is treated in this way in the model, but to avoid a have-one's-cake-and-eat-it-too paradox a concept of *negative consumption* is introduced, as described by ***Definition 3***.

Definition 3. Negative consumption (also:* inverted consumption, deconsumption*)

> If a component of fixed assets of the household is purchased as part of goal realization (and thus the cash outflow connected with the purchase enters utility of consumption), then the cash inflow from selling this component at some later time is also taken into account in the utility of consumption as consumption with negative sign, and is called *negative consumption* (or *inverted consumption* or *deconsumption*).

Changing a flat or house into a new one is, thus, treated as a negative consumption equal (as far as absolute values are concerned) to the price obtained for the old house and positive consumption equal to the price paid for the new one. Effectively, utility of net sum of negative and positive consumption is taken. In this way, utility is measured only for the part of the new house value that is in excess over the old house value.

4. Stochastic nature of goals and their interdependencies

To show the way in which the goals may be taken into account in the value function of the household, it is not necessary to identify the distributions precisely. Knowledge of some of their general properties may be, however, useful. Of course, to construct a fully functional model, the detailed probabilities will be needed. The next important question is about the dependence between the goals. Constructing a model with a multivariate distribution of times and magnitudes for all goals would be a very tough task. And such approach would make the model hardly applicable in practice. In the proposition put forward here the relationships between some chosen goals are simplified to deterministic influence.

4.1. Goal dependency map

First, a map of relationships between the goals was created. It is a proposition only, but based on life experience, logical reasoning and general common knowledge of the nature of the goals. It is, yet, a simplification and it should be treated as such. The simplification is justified by the fact that the plan is revised every year (comp. section 2) and the goals may be shifted in time, added, removed or modified in respect of magnitude by the household. Thus, the attempt to include all possible stochastic and deterministic relationships might give only a spurious impression of precision. What is more important here is constructing a model of general influence structure. As a result, a kind of causal network of influences, mainly of deterministic character, was obtained. It is to a vast extent of (a kind of) hierarchical structure, because the direction of influences in this model is such that some goals rather exert influence, whereas other ones are rather influenced. The map of relationships between goals is constructed on the basis of the assumption that the main event in life of a household is birth of the first child. Thus, it may be said that it is the most "influential" one in the aforementioned hierarchy of influences.

A general map of relationships between goals for a stylized typical household may look for example like the one in the Figure 1. Arrows indicate influence directions.

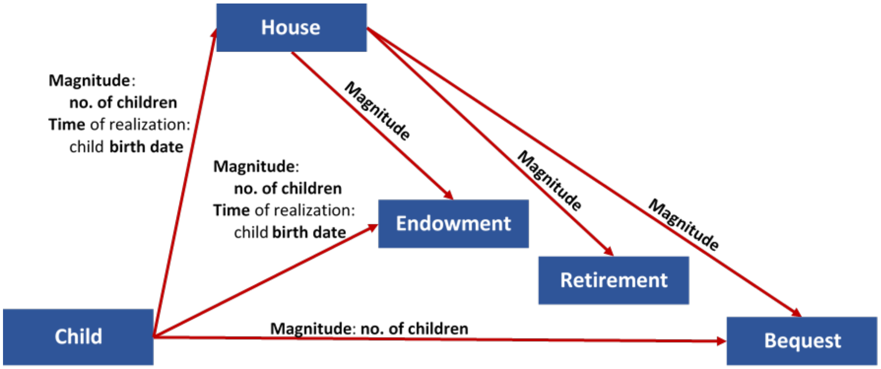


Figure 1. Stylized map of relationships between goals

In Figure 1 the following influences are marked:

- Time of house purchase is influenced by the time of a child's birth, and magnitude of this goal is influenced by the number of children (in further discussion the second relation is neglected).
- Time of endowment may be influenced by the time when children appear (particularly if it is the first flat for a child, study for children or some kind of dowry). Magnitude of endowment depends on the number of children if the addressee of the endowment are children of the main household members.
- Magnitude of bequest is not directly controlled by the household, but it depends on the bequest motive parameter, which is at the discretion of household members. The bequest motive may, in turn, depend on the number of children. Also, the value of the house contributes to bequest, since the house is a part of wealth that may be bequeathed. Thus, the goal "house" and the goal "child" influence the magnitude of bequest.

- Retirement income may be (partially) obtained from inverted mortgage secured by the house the person lives in. Thus, the size of retirement is influenced by the value of the house.

The model of the goal relationship map may be used to construct a “child-driven” model of the household life-long financial plan. Certainly, if the household does not plan children, all dependencies between the time of a child's birth and other goals disappear. For example, Type III goal (House) is then not influenced by any goal in this model, but may still exert influence on other goals.

If the household plans a child (children), stochastic time of a child's birth (conditional on the planned time) is added to the main sources of uncertainty from the basic version of the model (lifetimes of the two main household members). For the sake of simplicity it may be assumed that the dates of buying a house and making an endowment are just shifted by a constant number of years in relation to the date of the first child's birth (let the shifts be denoted by Δ_h and Δ_e, respectively).

4.2. Time and magnitude distributions

It is proposed to distinguish four distributions:

- for the time of realization of the type I goal (Child) – conditional distribution of a child's birth, under the condition of the planned time,
- for magnitude of the type III goal (House) – distribution of real estate price at the moment of purchase,
- for magnitude of the type IV goal (Endowment) – none or distribution of (commodity/real estate) prices if some commodity or real estate is to be endowed,
- for magnitude of the type II goal (Retirement) – distribution obtained from a survival model (it refers only to the cases when a household member dies before retirement, in all other cases the retirement goal is set on the basis of the cost of purchasing a life annuity, being a scalar derived from expected life time),
- for time of realization of the type V goal (Bequest) – distribution obtained from a survival model (distribution of the maximum of lifetimes of the two main household members).

The summary of time and magnitude distributions is given in the Table 2.

Table 2. Distributions of goal realization time and magnitude

	Child	House	Endowment	Retirement	Bequest
Magnitude distribution	None	**Distribution of real estate prices**	**Distribution of prices** or none (alternatively)	**Distribution obtained from a survival model**	None
Time-of-realization distribution	**Conditional distribution of time of realization** **(conditional on planned time)**	None	None	None	**Distribution obtained from a survival model**
Dependence between goals	**Stochastic:** **each next occurrence depends on the previous ones**	Deterministic: Time of realization: - planned, corrected by time of a child's birth; Magnitude: - planned, corrected by the no. of children	Deterministic: Time of realization: - planned, corrected by time of a child's birth; Magnitude: - planned, corrected by the no. of children	None	None – bequest motive taken into account in the household preferences

It is possible to characterize in some more details the distributions listed in the Table 2.

- **Conditional distribution of a child's birth**

Albeit a parametric model is not yet known, it is possible to formulate some postulates about the distribution.

For a univariate case (e.g. conditional distribution of the time when the first child is born, conditional on the planned time) the distribution may have the following properties:

- It is asymmetric. If the mother's age at the planned time of birth is young, then it is right-skewed. If the mother is of a more mature age at the planned time, the distribution is skewed to the left. For the planned age somewhere in the middle of these two extremes, the distribution may be more like a symmetric one.
- Modal value of the distribution should be close to the planned date.
- It is truncated. Its domain is bounded from the left by the moment of plan preparation and from the right because of biological limitations (with some reservations – e.g. adoption).

The shape suggests that Gumbel or some alpha-stable distributions might be used to model it. Figure 1 shows suggested shapes for two planned times (mother, respectively, younger and older on the assumed date).

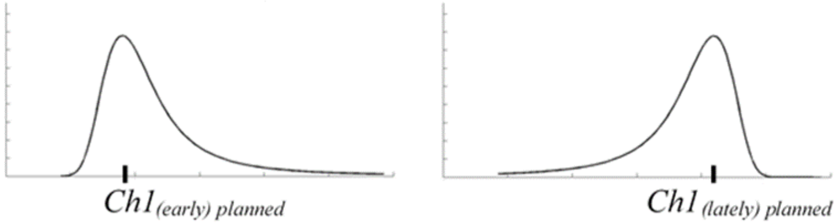
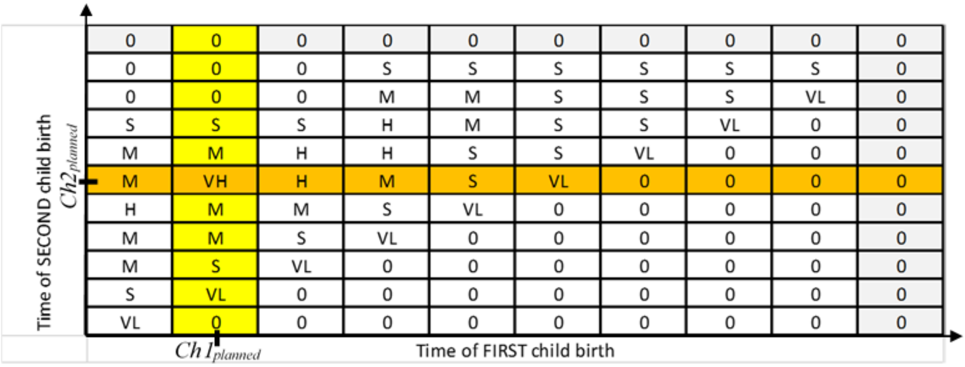


Figure 2. Stylized examples of conditional first child's birth distributions, under the condition that the panned time is $Ch1_{(early)\ planned}$ and $Ch1_{(lately)\ planned}$, respectively.

Certainly, life is more complex and the time of the first child's birth is not a sufficient piece of information indeed. More comprehensive, but also much more difficult, would be a model of joint distribution of times of birth for a number of children, including possibilities of multiple pregnancy. It might be also a kind of a stochastic process model of subsequent births, but such that also took into account the information about desired/planned times.

A very rough model of joint conditional-times-of-birth distribution for two children, given in a discrete version and only with qualitative description of probabilities, is proposed in Figure 3.This is just an illustration of how the joint distribution – if in a discrete version – might look like.


Time of SECOND child birth
$Ch2_{planned}$
$Ch1_{planned}$
Time of FIRST child birth

0	0	0	0	0	0	0	0	0	0
0	0	0	S	S	S	S	S	S	0
0	0	0	M	M	S	S	S	VL	0
S	S	S	H	M	S	S	VL	0	0
M	M	H	H	S	S	VL	0	0	0
M	VH	H	M	S	VL	0	0	0	0
H	M	M	S	VL	0	0	0	0	0
M	M	S	VL	0	0	0	0	0	0
M	S	VL	0	0	0	0	0	0	0
S	VL	0	0	0	0	0	0	0	0
VL	0	0	0	0	0	0	0	0	0

Figure 3. An attempt to construct a rough model of bivariate distribution of a child's birth, conditional on planned/desired times of birth. Probabilities given only in an ordinal scale. VH denotes “very high”, H – “high”, M – “medium”, S – “small”, VL – “very low”, 0 – zero.

- **Real estate price distribution**

Modelling of real estate price distributions is less problematic. There is a rich literature (e.g., Willcocks, 2009; Ghysels et al., 2012; Ohnishi et al., 2011) and data sets of real estate price indices are available, though the indices suffer from many limitations.

One of the biggest problems in statistical analysis of real estate prices is incoherence of data. Besides inhomogeneity, real estate market is characterized by high transaction costs, low liquidity, substantial cost of carry, lack of short sales, etc. (Ghysels et al., 2012).

To model real estate distributions one may use real-estate-suited models that try to take at least the most important idiosyncrasies of real estate market into account or, for the sake of simplicity, borrow solutions from some financial-price models. Then, the most common approach would be assuming log-normal distribution of prices.

Ohnishi et al. (2011) demonstrated that house prices in Tokyo show fat-tailed distributions (tails closer to power-laws than tails of lognormal distribution). But they also observed that size-adjusted prices, defined in their research as simple functions of house sizes and natural logarithms of house prices, are normally distributed. The last holds for almost all periods but speculative bubble when a fat right tail is observed (which refers both to crude prices and the size-adjusted constructs).

This gives a good ground to assume that modelling price distribution for the needs of the model will not face any conceptual difficulties, though technical problems may arise from the reasons discussed above. Let us assume for now that the prices are log-normally distributed. The further question is how to take the spectrum of possible prices into account in the model that is based on a discrete number of scenarios. A simple solution to this issue is proposed in section 5 (subsection 5.3)

- **Time of house purchase**

As it has already been mentioned in subsection 4.1, time of goal realization is the planned one, corrected by the actual time of a child's birth. At this stage of the model development, only the time of the first child's birth is used.

- **Distribution of household end (for time of bequest goal realization)**

It is not a distribution of maximum of household member lifetimes. The same maximum may be obtained in many ways, generating along the line different trajectories of household financial surplus. Instead, a bivariate survival process is considered, with regard to the financial processes this survival model underlies. Each pair of dates of death $\left(D1, D2\right)$ constitutes a different survival scenario, with a corresponding financial scenario (reflected in this model by a cumulated surplus trajectory).

Then, unconditional bi-dimensional survival time distribution is used here. A selected subset of possible $(D1, D2)$ pairs, together with their probabilities, is used as the main grid of scenarios in each of the value functions presented in section 5. Upon them other scenarios, like the time a child's birth, may be built.

- **Distribution of survival scenario (for retirement goal size, but also used for determining probabilities of scenarios under which the plan is optimized)**

As it has been just mentioned, the unconditional bivariate distribution of survival is used. It may be obtained from any survival model. For the needs of calculations performed by Feldman, Pietrzyk and Rokita (2014c) and Pietrzyk and Rokita (2014) a combination of two independent univariate survival processes was used following Gompertz (1825) law. This is, of course, a simplification that is far from reality, since any dependences between survival processes within a couple are neglected. Instead, one might use bivariate survival models (Brockett, 1984; Carriere, 2000; Gutiérrez et al., 2008; Georges et al., 2001). The choice of a survival model, from which unconditional probabilities of scenarios (as used in section 5) are derived, does not change the general concept.

The distributions listed above are then used in the value function of the household. On the basis of the distributions, probabilities of some chosen scenarios are determined. The value functions are function of expected discounted utilities of consumption and bequest, calculated for the scenarios.

5. Multiple goals in household financial plan

This section puts forward propositions of the value functions of the household with different types of financial goals taken into account. First, the basic, with retirement goal (type II) and bequest motive is recalled. Then, the goal function is augmented to include the goal type I (Child). In order to show how the value function might be further extended, a function with goal type III (House) is also proposed. In all cases the goal function is a sum of expected discounted utilities of consumption and bequest. The difference consists in the scenarios for which the utilities are calculated, and also in the functions calculating consumption and bequest (that take arguments connected with goals). The analytical form of utility function is the same and it may be any function fulfilling conditions of a utility function.

5.1. Retirement and bequest only

In the basic version of the model (Feldman, Pietrzyk and Rokita, 2014c; Pietrzyk and Rokita, 2014) the only stochastic factors are lifetimes of the two main household members (decision makers). The uncertainty about length of life is expressed in the value function by means of the so-called *range of concern*.

Since the model is defined in a discrete time, the range of concern may be treated as a sub-matrix of the survival scenario matrix. Namely – such sub-matrix of scenarios which includes only the dates of death that are between γ^*years before and δ^*years after the expected lifetime of a given person. Formally, the range of concern covers such date-of-death pairs that fulfil the following condition:

$$(D1_G, D2_G) = [E(D1)-\gamma^*, E(D1)+\delta^*]\times[E(D2)-\gamma^*, E(D2)+\delta^*] \qquad (1)$$

The parameters γ^* and δ^* are at the same time risk aversion parameters in respect of length-of-life risk. The parameter γ^* corresponds to premature-death risk aversion and δ^* to longevity risk aversion. The higher risk aversion the broader the range of concern.

The idea of the range of concern does not consist only in simplification and a convenient way of expressing risk aversion, but it also has a deeper sense. There are two more advantages of such solution. The first is avoiding too demanding plans that would require draconian saving and investing measures in order to be protected against some very unlikely, though theoretically probable, scenarios. The second is cutting off those scenarios that, in mathematical sense, are probable but should be removed from the scope of considered ones because of psychological reasons. For example, it is hard to treat a young widow as still the same household in the scenario in which her or his spouse dies very young (after some time this person would rather launch a new household with a new financial plan, and plausibly even with a new life partner).

The value function is calculated in the following way: for each survival scenario belonging to the range of concern a sum of discounted utilities of consumption is taken throughout the whole consumption path (from the starting moment $t_0 = 0$ until the maximum of the two dates of death for this particular scenario). Utility of bequest is for each scenario calculated only once, namely for the end of the scenario. The sums of discounted utilities of each scenario are weighted with respective unconditional probability. At this stage, the only type of scenario is a survival scenario.

The value function in this basic version is given with the equation (2):

$$V(c_0, v) =$$

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

(2)

where:

c_0 – consumption rate at the moment 0,

v_0 – proportion of Person 1 investment in total one-period contribution of the household ($v \equiv v_1, v_1 = 1 - v_2$),

$u(.)$ – utility function (the same in all segments of the formula),

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

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

$\gamma(t)$ – premature death risk aversion measure (depends on γ^*),

$\delta(t)$ – longevity risk aversion measure (depends on δ^*),

$p_{D_1^* D_2^*}$ – (unconditional) probability of such scenario that $\left(D1 = D_1^*, D2 = D_2^*\right)$,

α – consumption preference,

β – bequest preference,

r_C – discount rate of consumption,

r_B – discount rate of bequest,

$max\left\{D_1^*, D_2^*\right\}$ – time of household end under the scenario of $\left(D1 = D_1^*, D2 = D_2^*\right)$,

$C\left(t; D_1^*, D_2^*\right)$ – consumption at the moment *t* in the $\left(D_1^*, D_2^*\right)$ scenario,

$B\left(t; D_1^*, D_2^*\right)$ – cumulated investments and surplus of both household members at the moment *t* in the $\left(D_1^*, D_2^*\right)$ scenario; for $t = max\left\{D_1^*, D_2^*\right\}$ this is just amount of available bequest.

5.2. Augmenting the model by stochastic childbirth time

Let us assume that a pair is planning two children. They think of some time as the best for the first and the second child's birth, but certainly the true time of birth does not depend only on their decision. It is a random variable, conditional on the planned time.

The model with the Type I goals (2 children) is constructed using the same concept of a discrete grid of scenarios as in the previous subsection. The difference is that there is a number of childbirth scenarios put on each survival scenario. The range of possible the first child's births is from the start of the plan (

$t_0 = 0$) until the date of death of the woman in a given survival scenario, the second child's birth dates are between the first child's birth until the scenario end for the woman. Any number of children may be added to the model in this way, but here it is limited to only two for simplicity. The probabilities attached to each childbirth scenario are taken from the distribution of conditional childbirth times discussed in section 4, subsection 4.2.

The value function formula used in this variant of the model is as presented in the equation (3):

$$V(c_0, v) = \sum_{D_2^* = E(D2) - \gamma^*}^{E(D2) + \delta^*} \sum_{D_1^* = E(D1) - \gamma^*}^{E(D1) + \delta^*} p_{D_1^* D_2^*} \left[\sum_{Ch1 = t_0}^{W(D_1^*, D_2^*)} \sum_{Ch2 = Ch1}^{W(D_1^*, D_2^*)} p_{Ch1Ch2} \left[\begin{array}{l} \alpha \left(\sum_{t=0}^{max\{D_1^*, D_2^*\}} \frac{1}{(1+r_c)^t} u\Big(C\big(t; D_1^*, D_2^*, Ch1, Ch2\big)\Big)\big(\gamma(t) + \delta(t)\big) \right) + \\ \beta \frac{1}{(1+r_B)^{max\{D_1^*, D_2^*\}}} u\Big(B\big(max\{D_1^*, D_2^*\}; D_1^*, D_2^*, Ch1, Ch2\big)\Big) \end{array} \right] \right] \to \max \tag{3}$$

where:

$W\left(D_1^*, D_2^*\right)$ – time of death of the woman,

$Ch1$ – time of the first child's birth,

$Ch2$ – time of the second child's birth,

p_{Ch1Ch2} – probability of a scenario of child 1 and child 2 time births.

In addition to the new set of scenarios, a modification is also needed in the functions calculating consumption and cumulated wealth that may be bequeathed. Both these functions take two more arguments now, namely – dates of children's births.

5.3. Type III goal in the model

As it has been assumed in section 4, time of realization of goal type III is deterministically dependent on the child's birth. Thus, the only new set of scenarios that would have to be taken into account is the price of the intended purchase. It is natural to treat the price as a continuous random variable, which leads to an infinite number of scenarios. Fortunately, the direction of price influence on the general financial situation of the household is known and pretty obvious. At the time when the household intends to buy a residential real estate, it is the better the lower the market price, and the higher the price – the worse. And inversely, at the moment when the house becomes a part of wealth to be bequeathed, the higher market price the better, and the lower price the worse.

Let aversion to real-estate price risk in respect of the type III goal be expressed in terms of a tolerance level. The tolerance level, by analogy to *VaR* or *CFaR* tolerance levels, is here understood as a significance indicating the accepted probability of adverse scenarios. Here, it is some small pre-defined probability that a scenario of real estate prices will fall out of the range of scenarios for which the financial plan gives protection. Put differently, the household wants the plan to be optimized and meet budget and other constraints for at least all other scenarios.

For a tolerance level q, two quantile-based scenarios are considered. The first, for the moment of purchase. It is a right-tail quantile of real estate price distribution, corresponding to probability$1-q$. The distribution used here is unconditional (conditional on the state form the moment $t_0 = 0$) price distribution for the moment of purchase. The second, for the moment of bequest, is a conditional left-tail quantile corresponding to probability$q$, conditional on the upper quantile from the moment of purchase.

Let us assume that no new (larger) house nor any house expanding is planned when the second and next children are born.

The value function taking into account type III goal is given by the eq. (4):

$$V\left(c_0,v\right)=$$
$$\sum_{D_2^*=E(D2)-\gamma^*}^{E(D2)+\delta^*}\ \sum_{D_1^*=E(D1)-\gamma^*}^{E(D1)+\delta^*} p_{D_1^*D_2^*}\left[\sum_{Ch1=t_0}^{W\left(D_1^*;D_2^*\right)}\ \sum_{Ch2=Ch1}^{W\left(D_1^*;D_2^*\right)} p_{Ch1Ch2}\left[\begin{array}{l}\alpha\left(\displaystyle\sum_{t=0}^{max\left\{D_1^*,D_2^*\right\}}\frac{1}{\left(1+r_c\right)^t}u\left(C\left(t;D_1^*,D_2^*,Ch1,Ch2,S_{1-q}\right)\right)\left(\gamma\left(t\right)+\delta\left(t\right)\right)\right)+\\ \beta\dfrac{1}{\left(1+r_B\right)^{max\left\{D_1^*,D_2^*\right\}}}u\left(B\left(max\left\{D_1^*,D_2^*\right\};D_1^*,D_2^*,Ch1,Ch2,S_{1-q},S_q^*\right)\right)\end{array}\right]\right]\to\max \tag{4}$$

where:

$S_{1-q} = F_S^{-1}\left(1-q;T_h\right)$,

$S_q^* = F_{S|S(T_h)=S_{1-q}}^{-1}\left(q;T_B\right)$,

$T_h = Ch1+\Delta_h$.

$T_B = \max\{D_1^*, D_2^*\}$.

The $S_{1-q} = F_S^{-1}\left(1-q;T_h\right)$ denotes unconditional quantile of real estate price distribution at the moment T_h for probability $1-q$.

$S_q^* = F_{S|S(T_h)=S_{1-q}}^{-1}\left(q;T_B\right)$ denotes quantile corresponding to probability q of the conditional real estate price distribution at the moment T_B, conditional on the price at the moment $T_h$being equal to S_{1-q}.

Moreover, consumption and bequest functions take on new arguments.

Consumption function depends now on the price of the real estate at the moment of purchase (T_h). The wealth to be bequeathed depends also on the price of purchase at the moment T_h, because it influences consumption and thus also cumulated surplus. But in addition to that, bequest depends also on the price at the moment T_B, since it is a part of the residual wealth.

In practice, the risk aversion of the household in this respect may be calibrated on the basis of the maximum price of a real estate (given standard, technical state and location) the decision makers would accept to pay in the future and the minimum price they would accept when selling it then. Of course, the prices would need to be brought to present values as of the time when the plan is prepared. Otherwise, household members would not have a feeling of weather they are high or low in real sense. Then, probabilities might be calculated for the decision maker automatically in the planning system to translate the input into terms used in the model.

6. Summary

The proposed concept of the household financial plan model takes into account a source of uncertainty that is hardly ever addressed in the personal finance literature. Other stochastic factors, like survival of the household members, returns on investments, interest rates of credits, labour income, health condition, or events of insurance type (both life and non-life) are covered by many models, though usually not all of them together at the same time. Here, in turn, the time of goal realization and the magnitude of some chosen goals is discussed and an attempt to identify and review their main statistic properties is made. Also, relationships between goals, both in the sense of statistical dependence and causal links, are generally discussed. As a result “child-birth-time driven” model of household financial planning is obtained.

The model at its present version of development is a proposition of theoretic concept, still based on a number of unverified assumptions and dependent on parameters that have not yet been estimated. And in the case of conditional distribution of a child's birth under the condition of the planned time, no type of distributions (nor even a family) have been specified yet. Also the choice of links between financial goals is made arbitrary, though on the ground of some life experience, and logical reasoning. Empirical research may change somewhat the structure of the goal dependency map.

Further research, besides solving technical problem of this version of the model, will concentrate on augmenting it by other stochastic elements. Particularly, the risk factors connected with financing the goals need to be modelled.

Also, an important issue that has not yet been taken into account in this model is a trade-off between life insurance and capital-based protection against unexpected events in the area of survival process (Ibbotson et al. 2005, Huang et al., 2008).

Another question that may be taken up in further research is including different possible retirement goal realizations. In the accumulation phase different pension plans may be compared and selected (Blake et al., 2001). In the distribution phase different mixes of life annuity and asset fund may be used, plus some additional solutions (Blake et al., 2003; comp. also: Huang and Milevsky 2011; Milevsky and Huang 2011; Gong, Webb 2008). For a two-person household also a joint annuity variant, of the kind discussed by Brown and Poterba (2000), is worth considering.

Acknowledgement

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

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.

BLAKE, D., CAIRNS A., DOWD, K., (2001). Pensionmetrics: Stochastic Pension Plan Design During the Accumulation Phase. Insurance: Mathematics and Economics, Vol. 29, Issue 2, pp. 187−215.

BLAKE, D., CAIRNS A., DOWD, K., (2003). Pensionmetrics 2: Stochastic Pension Plan Design During the Distribution Phase. Insurance: Mathematics and Economics, Vol. 33, issue 1, pp. 29−47.

BODIE, Z., TREUSSARD, J., WILLEN, P., (2008). The Theory of Optimal Life-Cycle Saving and Investing. In: Z. Bodie, D. McLeavy, L.B. Siegel, eds., The Future of Life-Cycle Saving and Investing. The research Foundation of CFA.

BROCKETT, P. L., (1984). General bivariate Makeham laws. Scandinavian Actuarial Journal, Vol. 1984, Issue 3, pp. 150−156.

BROWN, J. R., POTERBA J. M., (2000). Joint Life Annuities And Annuity Demand By Married Couples, Journal of Risk and Insurance, 2000, 67(4), pp. 527−553.

BRUHN K., STEFFENSEN M., (2010). Household consumption, investment and life insurance. Insurance: Mathematics and Economics, 48, pp. 315−325.

CARRIERE, J. F., (2000). Bivariate Survival Models for Coupled Lives. Scandinavian Actuarial Journal, 2000:1, pp. 17−32.

COX, J.C., HUANG, CH., (1989). Optimal consumption and portfolio policies when asset prices follow a diffusion process. Journal of Economic Theory, Vol. 49, Issue 1, pp. 33−83.

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: M. Spiliopoulou, L. Schmidt-Thieme, R. Janning, eds., Data Analysis, Machine Learning and Knowledge Discovery. Studies in Classification, Data Analysis and Knowledge Organization. Berlin-Heidelberg: Springer, pp. 255−264.

FELDMAN, L., PIETRZYK, R., ROKITA, P., (2014b). Multiobjective optimization of financing household goals with multiple investment programs. Statistics in Transition, New Series, Spring, Vol. 15, No. 2, pp. 243−268.

FELDMAN, L., PIETRZYK, R., ROKITA, P., (2014c). Cumulated Surplus Approach and a New Proposal of Life-Length Risk Aversion Interpretation in Retirement Planning for a Household with Two Decision Makers (November 6, 2014). Available at SSRN: <http://ssrn.com/abstract=2473156>.

GEORGES, P., LAMY, A.-G., NICOLAS, E., QUIBEL, G., RONCALLI, T, (2001). Multivariate Survival Modelling: A Unified Approach with Copulas (May 28, 2001). Available at SSRN: <http://ssrn.com/abstract=1032559> [Version: 28 May 2001. Accessed: 04 Dec. 2014].

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.

GHYSELS, E., PLAZZI, A., TOROUS, W. N., VALKANOV, R. I., (2012). Forecasting Real Estate Prices. In: G. Elliott, A. Timmermann, eds., Handbook of Economic Forecasting. Vol. II, Elsevier.

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.

GUTÍERREZ, R., GUTÍERREZ-SANCHEZ, R., NAFIDI, A., (2008). A bivariate stochastic Gompertz diffusion model: statistical aspects and application to the joint modeling of the Gross Domestic Product and CO_2 emissions in Spain. Environmetrix Vol. 19, Issue 6, pp. 643−658.

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., WANG, J., (2008). Portfolio Choice and Life Insurance: The CRRA Case. Journal of Risk & Insurance, Vol. 75, Issue 4, pp. 847−872.

IBBOTSON, R. G., CHEN, P., MILEVSKY, M. A., ZHU, X., (2005). Human Capital, Asset Allocation, and Life Insurance. Yale ICF Working Paper No. 05-11. Available at SSRN: <http://ssrn.com/abstract=723167>.

KONICZ, A. K., PISINGER, D., RASMUSSEN, K. M., STEFFENSEN, M., (2014). A Combined Stochastic Programming and Optimal Control Approach to Personal Finance and Pensions. Available at SSRN: <http://ssrn.com/abstract=2432869> [Version: 30 April 2014. Accessed: 24 Nov. 2014].

KOTLIKOFF, L. J., SPIVAK, A., (1981). The Family as an Incomplete Annuities Market. Journal of Political Economy, Vol. 89, No. 2, (April 1981), pp. 372−391.

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 Planet 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.

OHNISHI, T., MIZUNO, T., SHIMZU, CH., WATANABE, T., (2011). The Evolution of House Price Distribution. RIETI Discussion Paper Series 11-E-019. Available at <http://www.rieti.go.jp/jp/publications/dp/11e019.pdf> [Accessed: 04 Dec. 2014].

PIETRZYK, R. A., ROKITA, P. A., (2014). Facilitating Household Financial Plan Optimization by Adjusting Time Range of Analysis to Life-Length Risk Aversion (October 22, 2014).
Available at SSRN: <http://ssrn.com/abstract=2513393>.

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.

RUSZCZYŃSKI, A., SHAPIRO, A., (2003). Stochastic Programming Models. In: Ruszczyński, A., Shapiro, A., eds, Handbooks in Operations Research and Management Science, 10: Stochastic Programming, pp. 1−64.

VICKSON, R. G., ZIEMBA, W. T., eds, (2006). Stochastic Optimization Models in Finance. World Scientific.

WILLCOCKS, G., (2009). UK Housing Market: Time Series Processes with Independent and Identically Distributed Residuals. Journal of Real Estate Finance and Economics, Vol. 39, No. 4, 2009, pp. 403−414.

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

ZALEGA, T., (2007). Gospodarstwa domowe jako podmiot konsumpcji (Households as consuming actors) (in Polish). Materials and Studies, Faculty of Management, University of Warsaw.

ZIEMBA, W. T., (2003). The Stochastic Programming Approach to Asset, Liability, and Wealth Management. The Research Foundation of AIMR™.

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 137–152

ROBUST REGRESSION IN MONTHLY BUSINESS SURVEY

Grażyna Dehnel[1]

ABSTRACT

There are many sample surveys of populations that contain outliers (extreme values). This is especially true in business, agricultural, household and medicine surveys. Outliers can have a large distorting influence on classical statistical methods that are optimal under the assumption of normality or linearity. As a result, the presence of extreme observations may adversely affect estimation, especially when it is carried out at a low level of aggregation. To deal with this problem, several alternative techniques of estimation, less sensitive to outliers, have been proposed in the statistical literature. In this paper we attempt to apply and assess some robust regression methods (*LTS, M-estimation, S-estimation, MM-estimation*) in the business survey conducted within the framework of official statistics.

Key words*:* robust regression, outlier detection, business statistics.

1. Introduction

One of the main problems involved in estimating population parameters is distributions of enterprises in terms of the variable of interest and auxiliary variables, which are characterised by a high variation, strong asymmetry and kurtosis. This is due to, inter alia, non-response survey errors, a large proportion of zero values for survey variables and extreme values. In this article we focus on the third issue – extreme values. Although some observations are extreme, they need not necessarily be incorrect but may be part of the survey population. The statistical literature refers to these observations as outliers.

Undoubtedly, in many business surveys conducted within the framework of official statistics sample sizes are large enough to compensate for the presence of outliers, which have a relatively small impact on estimates. However, at low levels of unit aggregation the impact of outliers might be significant. The appearance of outlying observations is particularly noticeable in estimates for short-term statistics, where surveys are repeated monthly or quarterly.

[1] Poznan University of Economics, Department of Statistics. E-mail: g.dehnel@ue.poznan.pl.

The aim of outlier treatment is to make estimates as close to the parameters of the population as possible. This is not a simple task in the presence of outliers, since estimators do not retain their properties, such as resistance to bias or efficiency. This means that outlier treatment should provide some kind of a trade-off between variance and bias. And in the case of a population known to comprise outliers (such as populations of enterprises) a robust analysis should be considered in addition to the classical approach. In the statistical literature several robust methods have been proposed. The aim of the present study was to compare the usefulness of four robust regression methods: *LTS, M-estimation, S-estimation, MM-estimation* against LS regression estimation based on data derived from a business survey. An empirical example is conducted in SAS.

2. Robust regression methods

The main objective of robust regression methods is to provide stable results when fundamental assumptions of the least squares regression are not fulfilled due to the presence of outliers. In order to achieve this stability, robust regression limits the influence of outliers by reducing the weights of outliers, changing the values of outliers or using robust estimation techniques (Chen, 2007). Many methods have been developed for these problems, but those most commonly used today are Huber *M-estimation, Least Trimmed Squares (LTS) estimation, S-estimation* and *MM-estimation.*

M-estimation

The most widely used general method of robust regression is *M-estimation,* introduced by Huber (1964), which is nearly as efficient as LS (Huber, 1964).

Instead of minimizing the sum of squares of the residuals, a Huber-type M-estimator $\hat{\theta}_M$ of θ minimizes the sum of less rapidly increasing functions of the residuals:

$$\hat{\theta}_M = \arg\min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{r_i}{s}(\theta)\right) \tag{1}$$

where $r_i = y_i - X\theta$,

s - scale parameter,

ρ(·) is a loss function, which is even, non-decreasing for positive values and less increasing than the square function.

To guarantee scale equivariance (i.e. independence with respect to the measurement units of the dependent variable), residuals are standardized by a measure of dispersion *s* (Verardi, Croux, 2009).

The estimator is not robust with respect to leverage points, but it is useful in analyzing data for which it can be assumed that the contamination is mainly in the *y*-direction.

Assuming *s* to be known, the M-estimate is found by solving:

$$\sum_{i=1}^{n} \Psi\left(\frac{y_i - \sum_{k=1}^{p} x_{ik}\theta_k}{s} \right) x_i = 0 \tag{2}$$

where Ψ is the first derivative of ρ.

The choice of the Ψ function is based on the preference of how much weight to assign to outliers and this leads to different variants of M-estimators. A monotone Ψ function does not assign weight to large outliers as big as least squares do. A redescending Ψ function increases the weight assigned to an outlier until a specified distance (e.g. 3σ) and then decreases the weight to 0 as the outlying distance gets larger (Alma, 2011).

The choice of the Ψ function is not critical to gaining a good robust estimate, and many choices will give similar results that offer great improvements, in terms of efficiency and bias, over classical estimates in the presence of outliers (Huber, 1981). *M-estimation* has a breakdown point of 1/n.

Least Trimmed Squares (LTS) Estimation

The *least trimmed squares (LTS) estimate* proposed by Rousseeuw (1984) is given by

$$\hat{\theta}_{LTS} = \arg\min_{\theta} \sum_{i=1}^{h} Q_{LTS}(\theta) \tag{3}$$

where $Q_{LTS}(\theta) = \sum_{i=1}^{h} r_{(i)}^2$,

$r_{(1)}^2 \le r_{(2)}^2 \le \ldots \le r_{(n)}^2$ - are the ordered squared residuals

h – is defined in the range $\frac{n}{2} + 1 \le h \le \frac{3n + p + 1}{4}$ or $h = \frac{n + p + 1}{2}$,

p – number of parameters.

LTS is calculated by minimizing the *h* ordered squares residuals. The largest squared residuals are excluded, which allows those outlier data points to be removed completely.

Depending on the value of *h* and the outlier data configuration, LTS can be very efficient. In fact, if the exact numbers of outliers are trimmed, this method is computationally equivalent to LS (Alma, 2011). However, if there are more extreme values than are trimmed, this method is not as efficient. In turn, if there is more trimming than there are outlying data points, then some good data will be excluded from the estimation. LTS is considered to be a high breakdown method with a breakdown point of 50% (Rousseeuw, Leroy, 1987; Rousseeuw, Driessen, 1998).

S-estimation

S-estimation proposed by Rousseeuw and Yohai (1984) minimizes the dispersion of the residuals. However, it uses a robust measure for the variance. It is defined as

$\hat{\theta}_s = \arg\min_{\theta} \hat{\sigma}(r(\theta))$ where $\hat{\sigma}(r)$ is an M-estimator of scale, found as the solution of

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho\left(\frac{Y_i - x_i'\theta}{\hat{\sigma}}\right) = K \tag{4}$$

where $K = const = E[\rho]$.

The final scale estimate, $\hat{\sigma}$, is the standard deviation of the residuals from the fit that minimized the dispersion of the residuals.

Rousseeuw and Yohai (1984) suggest a redescending influence function as:

$$\rho(x) = \begin{cases} \dfrac{x^2}{2} - \dfrac{x^4}{2c^2} + \dfrac{x^6}{6c^4} & \text{for } |x| \le c \\ \\ \dfrac{c^2}{6} & \text{for } |x| > c \end{cases} \tag{5}$$

The parameter *c* is the tuning constant. Trade-offs in breakdown and efficiency are possible based on choices for tuning constant *c* and *K* (Alma, 2011). The usual choice is *c*=1.548 and K=0.1995 for 50% breakdown and about 28% asymptotic efficiency (Rousseeuw, Leroy, 1987). *S*-estimation is a high breakdown value method.

MM-estimation

MM-estimation is a combination of high breakdown value estimation and efficient estimation, which was introduced by Yohai (1987).

The procedure consists of three steps (Alma 2011):

1. Calculation of an *S*-estimate with the influence function
2.

$$\rho(r) = \begin{cases} 3\left(\dfrac{r}{c}\right)^2 - 3\left(\dfrac{r}{c}\right)^4 + \left(\dfrac{r}{c}\right)^6 & \text{for } |r| \le k \\ \\ 1 & \text{for } |r| > k \end{cases} \tag{6}$$

The value of the tuning constant, c, is selected as 1.548.

3. Calculation of the MM parameters that provide the minimum value of $\sum_{i=1}^{n} \rho\left(\frac{Y_i - x_i'\theta_{MM}}{\hat{\sigma}}\right)$ where $\rho(r)$ is the influence function used in the first stage with the tuning constant 4.687 and 0 σˆ is the estimate of scale from the first step (standard deviation of the residuals).
4. Calculation of the MM estimate of scale as the solution to

$$\frac{1}{n-p}\sum_{i=1}^{n} \rho\left(\frac{Y_i - x_i'\theta_{MM}}{\hat{\sigma}}\right) = 0{,}5 \tag{7}$$

MM-estimation is a special type of M-estimation. It is the estimation with a high breakdown point (50%) and high efficiency (70%) under normal error (Stromberg, 1993).

3. Data source

Information for the study came from the DG1 survey conducted by the Statistical Office in Poznan and from tax register. The DG1 survey is a business activity report submitted by large, medium-sized and small enterprises. It is the basic source of short-term information about economic activity of businesses, such as *revenue from sales (of products and services), number of employees, gross wages, volume of wholesale trade and retail sales, excise tax, specific subsidies*. The sample frame includes 98,000 units, of which 19,000 are medium-sized and large enterprises (with over 49 employees), 80,000 are small enterprises (from 10 to 49 employees). Tax register served as the auxiliary data source.

4. Description of the study

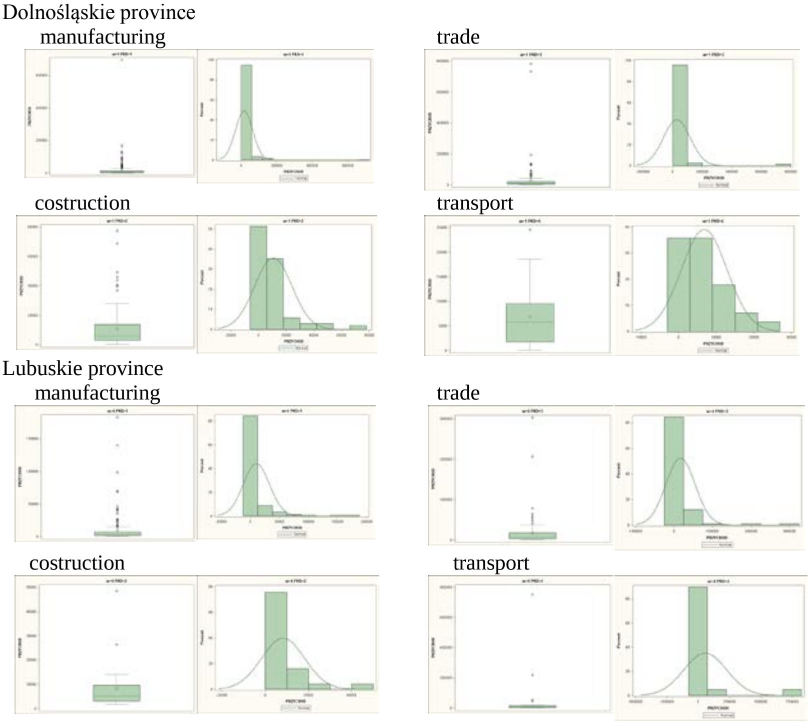
The study was limited to small enterprises (from 10 to 49 employed persons) that were active in December 2011. We took into consideration two models with a different number of auxiliary variables. There was one auxiliary variable (cost) in the first model and there were three auxiliary variables (income, cost and *revenue)* in the second model. *Revenue from sales of products (goods and services)* was the target variable. The general population included all small and medium-sized enterprises that participated in the DG1 survey. This choice enabled access to detailed information about the target and auxiliary variables. The level of aggregation adopted for the study was a combination of the territorial division by province and the biggest four sections of economic activity (NACE Rev.2) - *manufacturing, construction, trade* and *transport*. The population was thus broken down into 64 domains (16 provinces x 4 NACE sections). Owing to the large volume of study results, the following presentation is limited to two provinces - Dolnośląskie, Lubuskie and the 4 NACE sections. The selection of provinces was made on the basis of information about the goodness of fit of the model. The main objective of choice was to include domains with a high variation of the coefficient of determination (from 0.041 to 0.999), see Table 1.

Table 1. Coefficient of determination for the regression models of *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces

AUXILIARY VARIABLES	*cost*	*income, cost, revenue*	*cost*	*income, cost, revenue*
NACE / Provinces	*Dolnośląskie*		*Lubuskie*	
Manufacturing	0.975	0.996	0.985	0.991
Construction	0.742	0.815	0.980	0.995
Trade	0.989	0.991	0.999	0.999
Transport	0.498	0.510	0.041	0.050

Source: Own calculations based on DG1 survey and tax register from December 2011.

The analysis started with the assessment of the distributions of the variable of interest coming from DG1 survey. A look at the distributions of *revenue* shows that a relatively large percentage of economic entities display zero values (or close to zero) in this respect. Moreover, there are very long right-hand tails in the histograms, as expected, see Fig. 1. This is the justification for the use of statistical techniques that are able to cope with or to detect outlying observations.



Dolnośląskie province

manufacturing

trade

costruction

transport

Lubuskie province

manufacturing

trade

costruction

transport

Figure 1. Distribution of enterprises by revenue for *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces

Source: Own calculations based on DG1 survey and tax register from December 2011.

Based on the distributions, the analysis of outliers was divided into two stages: the first one involved detecting outlying observations, the second focused on ways of handling them to reduce their effect on survey estimates by applying robust regression methods.

To identify outlying observations *RSTUDENT* was applied. This is one of the most widely used measures to identify outliers. The value of the RSTUDENT for each observation is the difference between the observed y_i and the predicted value of $\hat{y}$ excluding this observation from the regression and can be calculated using the following formula:

$$r_i * = \frac{e_i}{\sqrt{MSE_i} \cdot \sqrt{1-h_i}} \tag{8}$$

where: $r_i *$ - RSTUDENT,

$e_i = y_i - \hat{y}$ - the i-th residual,

MSE$_{\mathrm{i}}$ - the error variance estimated without the i-th observation

$h_i = \mathbf{x}_{\mathrm{i}} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{x}_i$ − the i-th diagonal of the *hat matrix* (projection matrix).

If $|RSTUDENT| \geq 2$ then the observation is identified as an outlier, see Table 2.

Table 2. Number of enterprises and percentage of outliers for *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces

NACE / Provinces	Number of observations	Percentage of outliers	Number of observations	Percentage of outliers
	Dolnośląskie		*Lubuskie*	
manufacturing	772	1,6	368	3.3
construction	207	4,3	56	3.6
trade	315	1,3	149	4.7
transport	80	3,8	46	4.3

Source: Own calculations based on DG1 survey and tax register from December 2011.

It is important not only to identify outliers but also to classify them into types. This can be achieved by calling on the graphical tool − scatter plot of the RSTUDENT versus Leverage (the leverage statistic measures how far the observation is from the centroid of the *x*-space), see Fig. 2, 3. In order to better illustrate the relationship between the study variable *revenue* and auxiliary variable *cost*, scatter plots (with 95% confidence intervals) are presented, see Fig. 2, 3. We can distinguish tree types of extreme values (Rousseeuw, Leroy, 1987): *outliers in the y-direction* (in Fig. 2,3 denoted as *outliers), outliers in the* *x-direction* (in Fig. 2, 3 denoted as *Leverage)*, and *good leverage points*. Graphing relationships among variables reveal exceptions to general rules. The picture in the Figure 2 shows *outliers in the y-direction, in the x-direction and good leverage points.*

The presence of o*utliers in the y-direction* affects the estimated intercept of LS. *Good leverage points* that are outlying in the space of explanatory variables and are located close to the regression line do not affect the LS estimation. Their presence has an impact on statistical inference by deflating the estimated standard errors. The presence of *outliers in the x-direction* that are both outlying in the space of explanatory variables and located far from the regression line influences the LS estimation of both the intercept and the slope (Verardi, Croux, 2009).

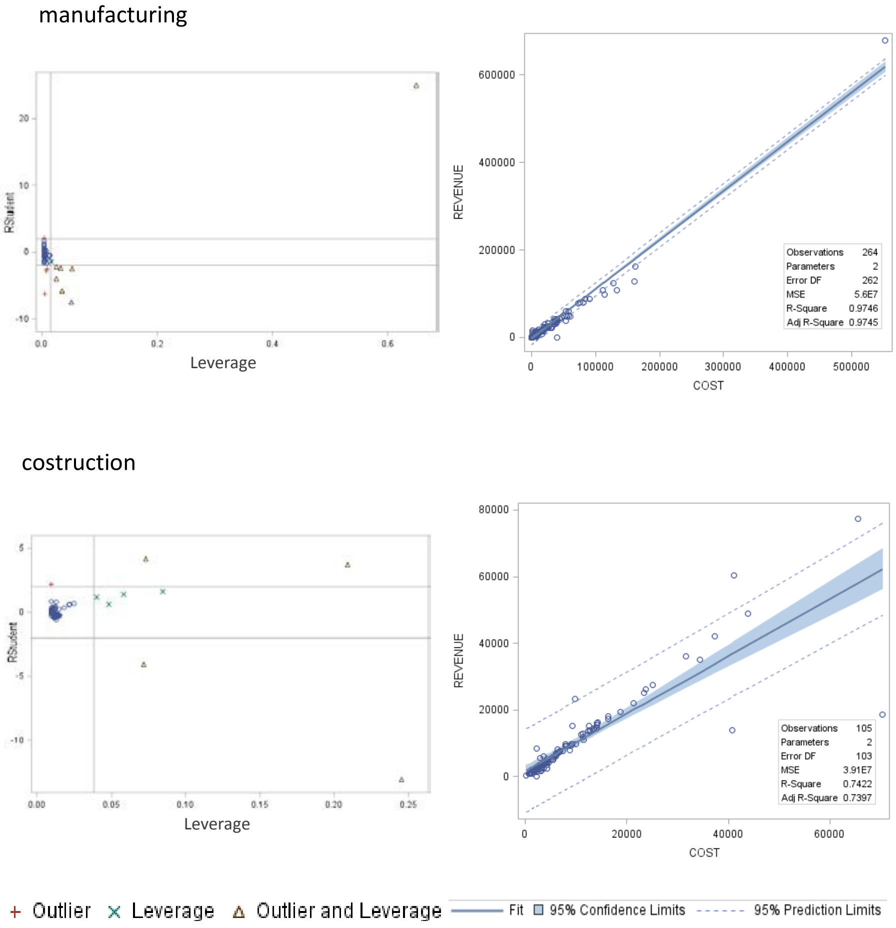


Figure 2. Outlier and Leverage diagnostic for *manufacturing and construction in* Dolnośląskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

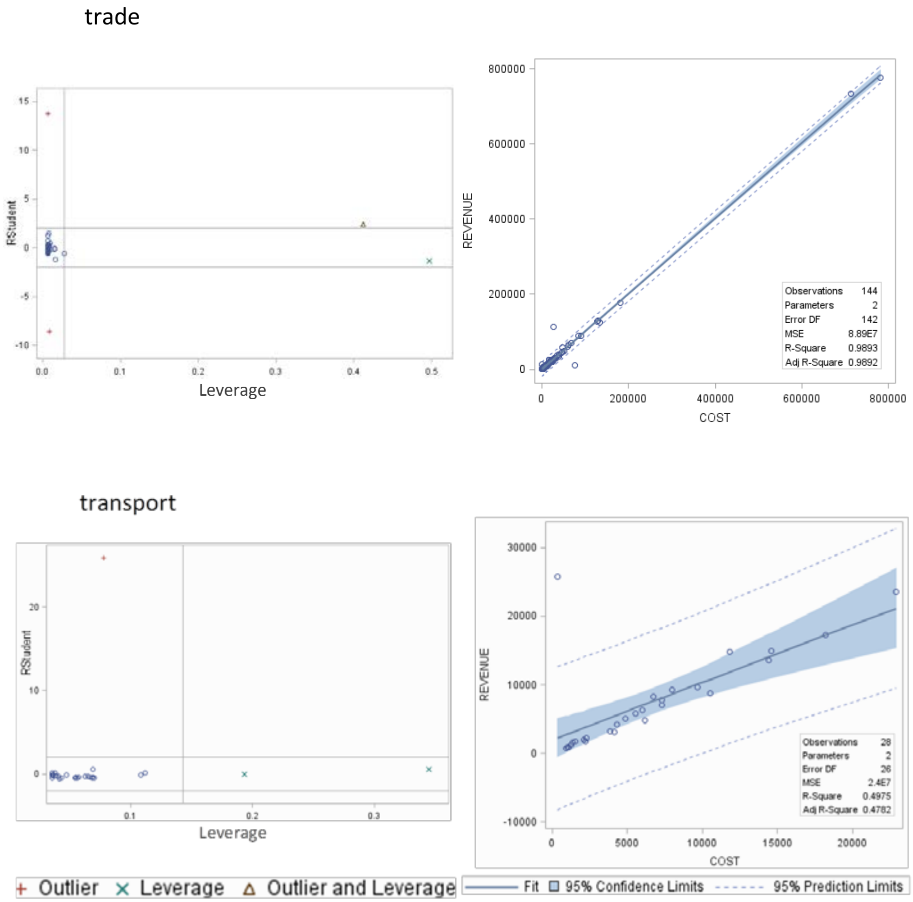
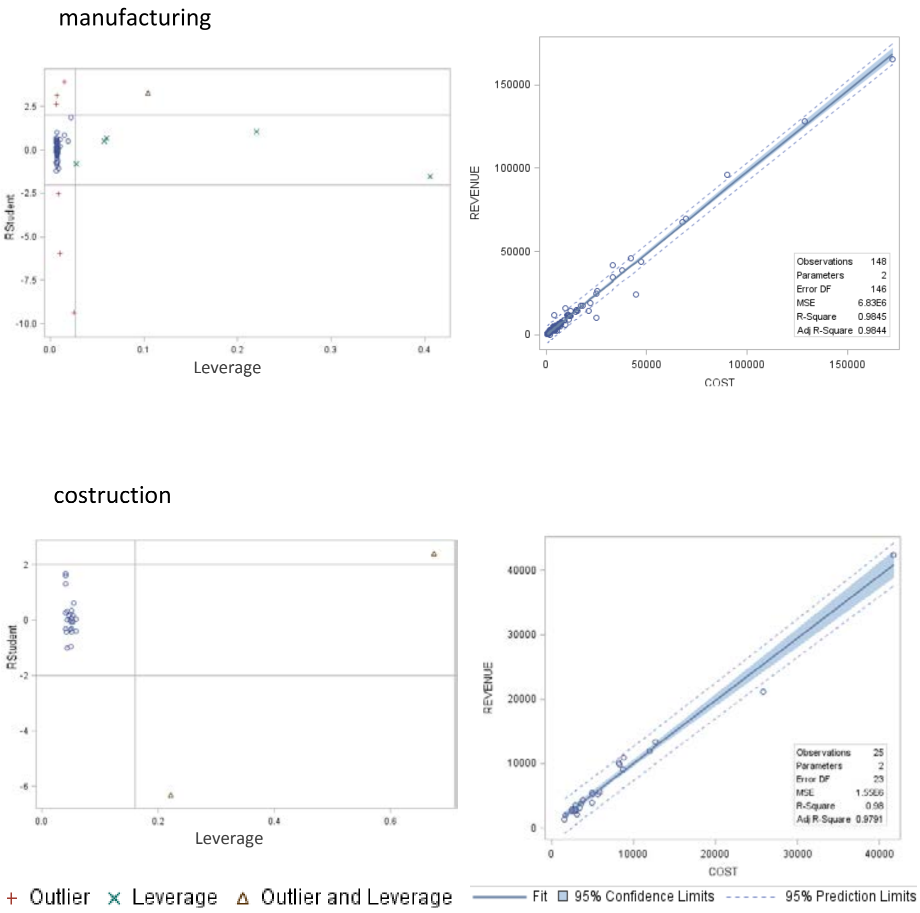


Figure 2a. Outlier and Leverage diagnostic for *trade and transport in* Dolnośląskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.



manufacturing

costruction

Figure 3. Outlier and Leverage diagnostic for *manufacturing and construction in* Lubuskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

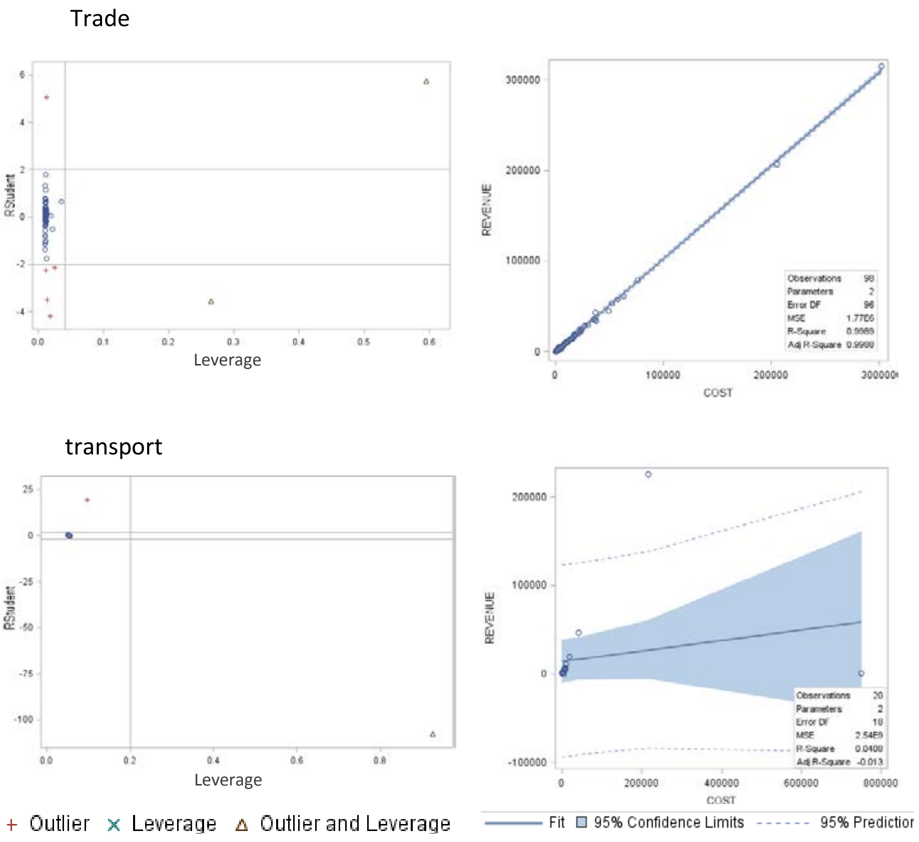


Figure 3. Outlier and Leverage diagnostic for *trade and transport in* Lubuskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

Outliers violate the assumption of normally distributed residuals in the least squares regression. It means that the estimation of model parameters using classical LS is not credible. Therefore, four robust regression methods were applied. The objective of this study was to compare *M-estimation*, *LTS*, *S-estimation* and *MM-estimation* against the LS regression estimation method in terms of the goodness of fit of the model that is represented by the coefficient of determination. The robust version of the coefficient of determination is defined as:

$$R^2 = \frac{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right) - \sum \rho\left(\frac{y_i - x_i^T \hat{\theta}}{\hat{s}}\right)}{\sum \rho\left(\frac{y_i - \hat{\mu}}{\hat{s}}\right)} \tag{9}$$

where ρ is the objective function for the robust estimate, $\hat{\mu}$ is the robust location estimator, and $\hat{s}$ is the robust scale estimator in the full model.

The results are presented in Table 3. The variation between the robust regression methods reflects their inherent sensitivity to the presence of outliers, see Fig. 2, 3. Moreover, their performance (and the resulting estimates) also depends on the type of outliers and their distance from the bulk of the data. The use of *M-estimation* improves the goodness of fit of the model only if the contamination is mainly in the *y*-direction. This limitation also applies to *MM-estimation*, which is determined by *M-estimation*.

Table 3. Coefficient of determination regression methods for *manufacturing, construction, trade and transport in* Dolnośląskie and Lubuskie provinces

auxiliary variables: cost

Regression methods	**LS**	**M**	**LTS**	**S**	**MM**
NACE / PROVINCE	**Dolnośląskie**				
manufacturing	0.975	0.648	0.980	0.978	0.760
construction	0.742	0.740	0.979	0.984	0.699
trade	0.989	0.703	0.994	0.994	0.772
transport	0.498	0.717	0.979	0.977	0.707
Lubuskie					
manufacturing	0.985	0.673	0.980	0.975	0.662
construction	0.980	0.718	0.979	0.970	0.773
trade	0.999	0.704	0.997	0.996	0.775
transport	0.041	0.016	0.934	0.851	0.791

auxiliary variables: income. cost. *revenue*

Regression methods	**LS**	**M**	**LTS**	**S**	**MM**
NACE / PROVINCE	**Dolnośląskie**				
manufacturing	0.996	0.677	0.998	0.998	0.768
construction	0.815	0.695	0.999	0.997	0.766
trade	0.991	0.703	0.999	0.998	0.770
transport	0.510	0.800	0.999	0.998	0.810
Lubuskie					
manufacturing	0.991	0.680	0.997	0.992	0.778
construction	0.995	0.854	0.986	0.981	0.823
trade	0.999	0.744	0.999	0.999	0.780
transport	0.050	0.070	0.966	0.862	0.845

Source: Own calculations based on DG1 survey and tax register from December 2011.

A closer look at section *transport* in Lubuskie province reveals a low value of the coefficient of determination for LS. The characteristics of the regression models for this domain are presented in Table 3. As can be seen, the use of *M-estimation*, in the presence of leverage points, had an adverse effect on the goodness of fit of the model. The value of the coefficient of determination decreased from 0.041 to 0.016. Both in the case of LS and *M-estimation*, the slope estimate equals 0.06 and -0.004 respectively, which means a lack of correlation between variables, see. Fig. 4. *Cost* is an insignificant auxiliary variable (its *p-value* equals 0.4), see Table 4. A considerable improvement in the goodness of fit can be observed for *S-estimation*, which is characterized by a high breakdown point. The *S-estimation* method performs better than *M-estimation* and *MM-estimation* because the data contains high leverage points.

The best goodness of fit was obtained for LTS (R^2 = 0.934), where the model accounts for 0.82% observations – observations with the largest squared residuals are excluded ($h = 36$, n=46).

As we can notice, the percentage of outliers for this domain equals 4.3%, see Table 2.

This means that the number of removed observations is higher than the number of outliers. None of the outlying observations affects the parameter of estimation. Unfortunately, some good data was excluded from the estimation, as should be expected.

Table 4. Characteristics of the regression models for *revenue* based on cost, transport, Lubuskie province

METHOD	Parameter	Estimate	Standard error	95% Confidence interval		p-value
LS	**intercept**	14256.241	11874.01	7.20	28504.80	0.2455
	COST	0.059	0.07	0.00	0.12	0.3933
M	**intercept**	3621.988	822.48	2009.95	5234.03	<.0001
	COST	-0.004	0.01	-0.01	0.01	0.4183
LTS	**intercept**	-590.616	207.13	-996.58	-184.65	0.0044
	COST	1.044	0.01	1.04	1.05	<.0001
S	**intercept**	-571.268	292.29	-1144.14	1.61	0.0506
	COST	1.045	0.01	1.03	1.06	<.0001
MM	**intercept**	-960.423	487.64	-1916.18	-4.67	0.0489
	COST	1.049	0.01	1.03	1.07	<.0001

Source: Own calculations based on DG1 survey and tax register from December 2011.

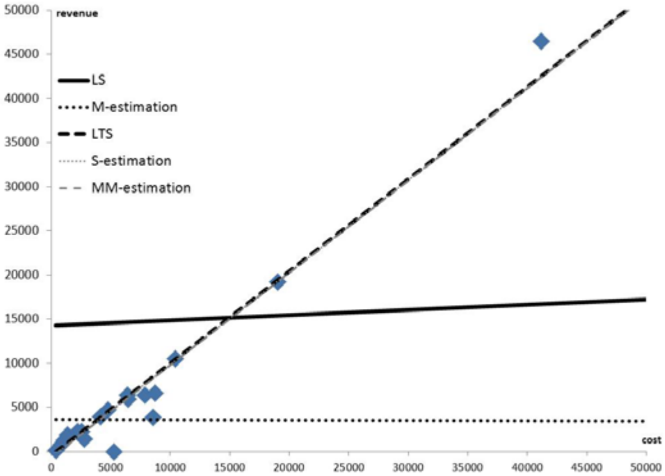


Figure 4. Regression lines for the data using LS, M-estimation, LTS, S-estimation and MM-estimation Methods, TRANSPORT, Lubuskie province

Source: Own calculations based on DG1 survey and tax register from December 2011.

Figure 4 presents information contained in Tables 3 and 4. The scatterplot displays a roughly linear relationship between two variables *revenue* and *cost*. The LS and *M-estimation* are immediately affected by the leverage points, so the estimated slope is close to zero – in this case *cost* turns out to be insignificant. After applying LTS, *S-estimation* and *MM-estimation* auxiliary variable *cost* became significant, see Table 4. Regression lines for *LTS, S-estimation* and *MM-estimation* coincide.

7. Conclusion

Conclusions about the assessment of robust regression methods drawn on the basis of the study overlap with those published in the literature of the subject (concerning the properties of the estimators in question):

- in general, the use of the M-estimator in the presence of outliers tends to improve the efficiency and reduce the bias compared to the classical methods of estimation,
- the M-estimator is not robust with respect to high leverage points, so it should be used in situations where high leverage points do not occur,
- the LTS method can be very efficient, but only under specific circumstances – when the number of trimmed observations is equal to the number of outliers. If there are more outliers than trimmed observations, the efficiency of LTS

method is low. In turn, if there is more trimming than there are outlying data points, then some good data will be excluded from the estimation,

- the MM-estimator is affected by the M-estimation, so it should be used with caution given the presence of high leverage points,
- robust regression methods can considerably improve estimation precision, but they should not be automatically applied instead of the classical methods.

REFERENCES

ALMA, Ö., G., (2011). Comparison of Robust Regression Methods in Linear Regression, [in:] Int. J. Contemp. Math. Sciences, Vol. 6, no. 9, pp. 409−421.

CHEN, C., (2007). Robust Regression and Outlier Detection with the ROBUSTREG Procedure, SUGI,
http://www2.sas.com/proceedings/sugi27/pp.265-27.pdf.

COX, B. G., BINDER, A., CHINNAPPA, N. B., CHRISTIANSON, A., COLLEDGE, M. J., KOTT, P. S., (1995). Business Survey Methods, John Wiley and Sons.

GROSS, W. F., BODE, G., TAYLOR, J. M., LLOYD–SMITH, C. W., (1986). Some finite population estimators which reduce the contribution of outliers, [in:] Proceedings of the Pacific Statistical Conference, 20-24 May 1985, Auckland, New Zealand.

HUBER, P. H., (1964). Robust estimation of a location parameter, The Annals of Mathematical Statistics, 35, pp.7−101.

HUBER, P. H., (1981). *Robust Statistics*, New York: John Wiley and Sons.

ROUSSEEUW, P. J., (1984). Least Median of Squares Regression, [in:] *Journal of the American Statistical Association*, 79, pp. 871−880.

ROUSSEEUW, P. J., YOHAI, V., (1984). Robust regression by means of S-estimators, [in:] W. H. J. Franke and D. Martin (Editors.), Robust and Nonlinear Time Series Analysis, Springer-Verlag, New-York, pp. 256−272.

ROUSSEEUW, P. J., LEROY, A. M., (1987). Robust Regression and Outlier Detection. Wiley-Interscience, New York.

ROUSSEEUW, P. J., DRIESSEN, K., (1998). Computing LTS regression for large data sets, Technical Report, University of Antwerp.

STROMBERG, A. J., (1993). Computation of high breakdown nonlinear regression parameters, [in:] Journal of the American Statistical Association, 88 (421).

VERARDI, V., CROUX, C., (2009). Robust regression in Stata, [in:] The Stata Journal, 9, Number 3, pp. 439−453.

YOHAI,V.J., (1987). High breakdown-point and high efficiency robust estimates for regression, The Annals of Statistics, 15, pp. 642−656.

STATISTICS IN TRANSITION new series, Spring 2015
Vol. 16, No. 1, pp. 153−156

ABOUT THE AUTHORS

Das Manjula works as an Associate Professor in the department of Mathematics at Faculty of Engineering & Technology (Institute of Technical Education & Research) under Siksha 'O' Anusandhan University, Odisha, India. She received her PhD in 2007 from Utkal University in Statistics in the area of Statistical Analysis of Women's Issues in Odisha - health, labour and education, under the supervision of Prof. AKPC Swain. Her research interest is statistical analysis in socio-economic issues and estimators of population mean in sampling.

Dehnel Grażyna is an Assistant Professor at the Department of Statistics, Poznań University of Economics. Her main research domain is short-term and structural business statistics, small area estimation. She is also interested in outlier robust regression applied on business data, data matching and business demography.

Elkasabi Mahmoud A. is a Sampling Expert at ICF International. He has a PhD from the University of Michigan in survey methodology. His specialties include population-based data collection, coverage and nonresponse adjustment, sample weighting, data calibration, dual frame estimation and methodological challenges in mixed-mode surveys. He has extensive experience in designing dual frame surveys, as well as other kinds of surveys for the internationally acclaimed Demographic and Health Surveys.

Górecki Tomasz received the M.Sc. degree in mathematics from the Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznań, Poland, in 2001. There, in 2005, he received a PhD degree. Currently he is an Assistant Professor at this University. His research interests include machine learning, times series classification and data mining.

Heeringa Steven G. is a Senior Research Scientist at the University of Michigan Institute for Social Research (ISR). He is a member of the Faculty of the University of Michigan Program in Survey Methods and the Joint Program in Survey Methodology. He is a Fellow of the American

Statistical Association and elected member of the International Statistical Institute. He is the author of many publications on statistical design and sampling and is the lead author of Applied Survey Data Analysis (Chapman & Hall, 2010). Steve has over 38 years of statistical sampling experience in the development of the SRC National Sample design, as well as research designs for ISR's major longitudinal and cross-sectional survey programs.

Krzyśko Mirosław is a Professor Emeritus at the Department of Probability and Mathematical Statistics in Adam Mickiewicz University, Poznań, Poland. His research interests are multivariate statistical analysis, analysis of multivariate functional data, statistical inference and data analysis in particular. Professor Krzyśko has published more than 150 research papers in international/national journals and conferences. He has also published five books/monographs. Professor Krzyśko is an active member of many scientific professional bodies.

Lepkowski James M. is a Research Professor in Survey Methodology at the Institute for Social Research and a Professor in Biostatistics at the University of Michigan. He directs the Michigan Summer Institute in Survey Research Techniques and is the Associate Director of the Michigan Graduate Program in Survey Methodology. His research interests include survey sample design and estimation and methods for compensating for missing data.

Lone Hilal Ahmad's area of research is sampling techniques. He received MPhil and PhD from School of Studies in Statistics, Vikram University Ujjain in 2012 and 2015 respectively. The title of his PhD thesis is **"**Contribution of Auxiliary information in Improved Estimation of Population Parameters in Sample Surveys**"**. He has published 10 research papers in reputed international journals and have 5 accepted papers in different reputed journals across the world. Presently he is a referee of three international journals, namely Communication for Statistical Application and Methods (CSAM), Model Assisted Statistics and Applications (MASA), Journal of Statistics Application and Probability Letters (JSAPL), and one national journal, namely Journal of Reliability and Statistical Studies. He has completed his postgraduate degree in statistics at the Department Of Statistics, University of Kashmir.

Longford Nicholas T. is the director of SNTL Statistics Research and Consulting (since 2004) and an academic visitor in the Department of Economics and Business, University Pompeu Fabra, in Barcelona, Spain.
He is the author of six monographs, the latest of them entitled 'Statistical Studies of Income, Poverty and Inequality in Europe: Computing and Graphics in R Using EU-SILC' (Taylor and Francis, 2015).
His professional interests include dealing with missing data, causal inference in observational studies, decision theory and statistical computing (in R) in general. His previous appointments include Department of Medical Statistics, De Montfort University, Leicester, England (1995-2004) and Educational Testing Service, Princeton, NJ, U.S.A (1986-1995).

Pietrzyk Radosław is an Assistant Professor at Wroclaw University of Economics, Department of Financial Investments and Risk Management. His main areas of interest are risk management, portfolio management and portfolio performance evaluation. Currently, his interest is focused on life-cycle-spanning integrated model of household risk. Within the research he develops the building blocks of the model that are connected with household financial goals and their financing, investment portfolio of the household, its performance evaluation and risk control in the financial plan.

Rokita Paweł is an Assistant Professor at Wroclaw University of Economics, Department of Financial Investments and Risk Management. For several years his research has been focused on the area of market risk measurement and dependence measurement, including dependence between extreme values. Currently his scientific interests encompass stability of financial system, systemic approach to financial markets and personal finance. In the field of personal finance he is a member of a research team developing a life-cycle-spanning integrated model of household risk. Within the project he is responsible for risk identification and measurement with regard to various types of risk faced by households, as well as integrating results of the research in this field with a model of household life-long financial plan.

Singh Abhishek is working as an Assistant Professor (Agricultural Statistics) in the Department of Farm Engineering, Institute of Agricultural Sciences, BHU, India. He received his PhD degree from Banaras Hindu University in 2012. He is a lifetime member of Indian Society of

Agricultural Statistics. His main research areas are time series forecasting, modelling and design of experiments.

Swain A. K. P. C. is a former Professor of Statistics, Utkal University, Vani Vihar, Bhubaneswar-751004, Odisha, India, having educated at Ravenshaw College, Cuttack, Utkal University and London School of Economics and Political Science, England, UK, and Indian Agricultural Statistics Research Institute, New Delhi. He was a recipient of Commonwealth Scholarship in the UK, awarded by the Commonwealth Scholarship Commission in the UK. He has visited Liverpool university, UK on Younger scientists' exchange programme and also visited Collegio de Postgraduados, Chapingo, Mexico under cultural exchange programme. He has published 40 research papers in Indian and international journals and successfully supervised eleven research scholars towards PhD degree in statistics, Utkal University. His research interests include sampling theory, inference and applied statistics.

Tailor Rajesh was born in 1976 in Ujjain (M.P.), India. He obtained his PhD degree from Vikram University, Ujjain under the supervision of renowned statistician Prof. H.P. Singh. He served National Council of Educational Research and Training, New Delhi, India as a lecturer and joined Vikram University as a Reader in 2007. In the field of sampling theory he has more than 60 research papers published in national and international peer-reviewed reputed journals.

Wołyński Waldemar is an Assistant Professor at the Faculty of Mathematics and Computer Science Adam Mickiewicz University in Poznań, Poland. His major research interests focus on various aspects of multivariate statistical analysis. He has been an author or co-author over 40 research papers.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, *Key words* (in bold italics) should be followed by three to four key words or brief phrases, preferably other than used in the title of the paper**.**
- ***Sectioning***. The paper should be divided into sections, and into subsections and smaller divisions as needed. Section titles should be in bold and left-justified, and numbered with **1.**, **2.**, **3.**, etc.
- ***Figures and tables***. In general, use only tables or figures (charts, graphs) that are essential. Tables and figures should be included within the body of the paper, not at the end. Among other things, this style dictates that the title for a table is placed above the table, while the title for a figure is placed below the graph or chart. If you do use tables, charts or graphs, choose a format that is economical in space. If needed, modify charts and graphs so that they use colours and patterns that are contrasting or distinct enough to be discernible in shades of grey when printed without colour.
- ***References.*** Each listed reference item should be cited in the text, and each text citation should be listed in the References**.** Referencing should be formatted after the Harvard Chicago System – see http://www.libweb.anglia.ac.uk/referencing/harvard.htm. When creating the list of bibliographic items, list all items in alphabetical order. References in the text should be cited with authors' name and the year of publication. If part of a reference is cited, indicate this after the reference, e.g. (Novak, 2003, p.125).