STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Volume 19, Number 2, June 2018

EDITORIAL OFFICE **ISSN 1234-7655**

Scientific Secretary
Marek Cierpiał-Wolan, Poland, e-mail: m.wolan@stat.gov.pl
Secretary
Patryk Barszcz, Poland, e-mail: P.Barszcz@stat.gov.pl, phone number 00 48 22 — 608 33 66
Technical Assistant
Rajmund Litkowiec, Poland

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

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 177−180

FROM THE EDITOR

This issue of the *Statistics in Transition new series* appears at a very special moment in its history, for two seemingly unrelated reasons which coincide in time: the 25th anniversary of the journal, and 100th anniversary of the Central Statistical Office, renamed recently on Statistics Poland, which sponsors the publication of this international journal of the Polish Statistical Association. As it was announced in the previous issues, there will be a special topical stream within the upcoming 2nd Congress of Polish Statistics (July10-12, 2018), envisaged as a way to celebrate this outstanding moment, and devoted to discussing the role of such a type of scientific (statistical) journals in promoting statistics as a discipline and as an instrument of creation and sustain community of specialists, and other stakeholders. Without falling into the tone of celebration, it seems worthwhile mentioning here the systematic progression of our journal in terms of its growing visibility in numerous international indexation bases, and of scores (impact factor) obtained from some of the most prestigious ones (for instance, above three times higher Scopus' CiteScore metrics for 2017 compared to 2016). We are totally aware of the fact that the primary source of such recognition was an increasing quality of the journal's articles. I would like to take this opportunity to express my deep appreciation to authors and peer-reviewers for their contributions to our joint efforts towards an excellence. Thanks to them the next quarter of a century of the *Statistics in Transition new series* looks optimistic and worth to support it.

The structure of this issue follows its regular thematic frame, i.e., it contains four sections, starting with *Sampling methods and estimation* and closing up with *Research Communicates and Letters*.

Carl-Erik Särndal's, Imbi Traat's and **Kaur Lumiste's** paper ***Interaction between data collection and estimation phases in surveys with nonresponse*** discusses approaches to deal with problems encountered in inference in surveys with nonresponse. The traditional focus on the estimation phase resulted in excelling some methods to reduce the nonresponse bias (propensity weighting and calibrated weighting) while the data collection phase has come into focus only recently. The authors take an integrated view where data collection and estimation are considered together. For a chosen auxiliary vector, they define the concepts incidence and inverse incidence and show their properties and relationship, showing that incidences are used in balancing the response in data collection; and that the inverse incidences are important for weighting adjustment in the estimation.

The paper by **Ceylan Talu Yozgatligil** and **H. Öztaş Ayhan, *Univariate sample size determination by alternative components: issues on design efficiency for complex samples*** is focused on the sample size determination taking into account some desired objectives: the level of confidence of estimates and the desired precision of the survey results, and the cost of enumeration.

Recently, some international organizations have been using univariate sample size determination approaches for their multivariate sample designs. These approaches also included some design efficiency and error statistics for the determination of the univariate sample sizes. They should be used for determining the survey quality measures after the data collection, not before. The additional components of the classical sample size measure will create selection and representation bias of survey estimates, which is discussed in this article.

Mir Subzar, Showkat Maqbool, Tariq Ahmad Raja, Surya Kant Pal and **Prayas Sharma** propose new estimators in the paper on ***Efficient estimators of population mean using auxiliary information under simple random sampling.*** An improved family of estimators for estimation of population mean is developed using the auxiliary information of median, quartile deviation, Gini's mean difference, Downton's Method, Probability Weighted Moments and their linear combinations with correlation coefficient and coefficient of variation. Their performance is analysed by mean square error and bias and compared with the existing estimators in the literature. By this comparison the authors go to conclusion that their proposed family of estimators is more efficient than the estimators offered in the literature. The theoretical results are supported by the empirical study.

O. Olawale Awe's and **A. Adedayo Adepoju's** article ***Modified recursive Bayesian algorithm for estimating time-varying parameters in dynamic linear models*** starts with observation that Estimation in Dynamic Linear Models (DLMs) with Fixed Parameters (FPs) has been faced with considerable limitations due to its inability to capture the dynamics of most time-varying phenomena in econometric studies. Since an attempt to overcome this limitation resulted in the use of Recursive Bayesian Algorithms (RBAs) - which also suffers from increased computational problems in estimating the Evolution Variance (EV) of the Time-Varying Parameters (TVPs) - the authors developed an alternative procedure. They propose a modified RBA for estimating TVPs in DLMs with reduced computational challenges.

In the next paper, ***Generalized exponential type estimator of population mean in the presence of non-response*****, Siraj Muneer, Javid Shabbir** and **Alamgir Khalila** propose a class of generalized exponential type estimators to estimate the finite population mean using two auxiliary variables under non-response in simple random sampling. The proposed estimator under non-response in different situations has been studied and gives minimum mean square error as compared to all other considered estimators. Usual exponential ratio type estimator, exponential product type estimator and many more estimators are also identified from the proposed estimator. They use three real data sets to obtain the efficiencies of estimators.

The second section, *Research articles,* starts with **Mauro Mussini's** paper ***On measuring polarization for ordinal data: an approach based on the decomposition of the Leti index.*** The measurement of polarization for ordinal data - which occurs in the distribution of an ordinal variable – involves the decomposition of the Leti heterogeneity index. The ratio of the between-group component of the index to the within-group component is used to measure the degree of polarization for an ordinal variable. This polarization measure does not

require imposing cardinality on ordered categories to quantify the degree of polarization in the distribution of an ordinal variable. Author addresses the practical issue of identifying groups by using classification trees for ordinal variables. This tree-based approach uncovers the most homogeneous groups from observed data, discovering the patterns of polarization in a data driven way. An application to Italian survey data on self-reported health status is shown.

Ujjwal Das and **Nader Ebrahimi** in the paper entitled ***New method for covariate selection in Cox model*** undertake the problem of selection right predictors starting with discussion of the criterion of penalized regression, known as “least absolute shrinkage and selection operator" (LASSO). The LASSO regression involves a penalizing parameter (commonly denoted by l) which controls the extent of penalty and hence plays a crucial role in identifying the right covariates. The author's propose an information theory-based method to determine the value of l in association with the Cox proportional hazards model. Furthermore, an efficient algorithm is discussed in the same context. They demonstrate the usefulness of the proposed method through an extensive simulation study. Finally, the performance of it is compared with existing methods and the algorithm is illustrated using a real data set.

In **Wioletta Grzenda's** and **Ewa Frątczak's** article ***Cohort patterns of fertility in Poland based on staging process – generations 1930-1980*** addressed is the problem of unprecedented changes in the fertility. Currently, the total fertility rate level is very low, about 1.3 children per woman, which is below the replacement level. Many studies have described changes in fertility based on the cross-sectional approach. In the authors' view, the changes of cohort fertility have been described not quite sufficiently. Therefore, they attempt to fill in this gap by the assessment of stochastic fertility tables, calculated for five-year generations of women born in the period 1930-1980. The main goal of this study is to analyse changes in the cohort patterns of female fertility in Poland. During the transformation period in Poland the model of nuclear family changed from two-child model into one-child model, with a high percentage of childless families in the general structure. More recent analysis of 15 Central and East European (CEE) countries, including Poland, confirms such tendencies and shows that despite the growth in fertility rates in the late 2000s, the fertility still remains at a low level.

The section *Other articles* contains a paper based on presentation at the 2018 Multivariate Statistical Analysis Conference in Łódź by **Daniel Kosiorowski, Dominik Mielczarek, Jerzy P. Rydlewski and Małgorzata Snarska,** ***Generalized exponential smoothing in prediction of hierarchical time series.*** The authors stars with presentation of a grouped functional time series forecasting approach being a combination of individual forecasts obtained using the generalized least squares method. They modify the Shang-Hyndman methodology using a generalized exponential smoothing technique for the most disaggregates functional time series in order to obtain a more robust predictor. They discuss some properties of their proposals based on the results obtained via simulation studies and analysis of real data related to the prediction of demand for electricity in Australia (in 2016).

The final section, *Research Communicates and Letters,* contains two articles. In the first, by **Czesław Stępniak,** entitled ***On a surprising result of two-candidate election forecast based on the first leadership time***, author presents "a simple but provocative note". He considers an election with two candidates, and under assumption that candidate A was the leader until counting n votes, and he asks the question "How to use this information in predicting the final results of the election?" According to the common belief the final number of votes for the leader should be a strictly increasing function of n. Assuming the votes are counted in random order, it is possible to derive the Maximum Likelihood predictor of the final number of votes for the future winner and loser based on the first leadership time. It appears that this time has little effect on the predicting. The first leadership time is informative for the final results of the election only in the trivial case.

In the next paper, ***The wellbeing effect of community development. Some measurement and modeling issues,*** **Włodzimierz Okrasa** and **Dominik Rozkrut** discuss the interconnected methodological tasks, measurement and modeling, in the context of exploration of the cross-level interaction between the local community development and individual wellbeing. The preliminary results illustrate usefulness of an analytical framework aimed to assess an impact of the local development on individual wellbeing through multilevel modeling, accounting for spatial effects. To this aim, a dual measurement system is employed with data from two independent sources: (i) the Local Data Bank (LDB) for calculating a multidimensional index of local deprivation (MILD) and (ii) the Time Use Survey data to construct the U-index (‘time of unpleasant state’), considered as a measure of individual wellbeing. Since one of the implications of the main hypothesis on the interaction between community development and individual wellbeing is the importance of 'place' and 'space', a special emphasize has been put on spatial effects, i.e. geographic clusters and spatial associations (autocorrelation, dependence). The evidence that place and space matter for this relationship provides support for validity of both multilevel and spatial approaches (ideally, combined) to this type of problems.

Włodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 181

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.,
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, June 2018
Vol. 19, No. 2, pp. 183–200, DOI 10.21307/stattrans-2018-011

INTERACTION BETWEEN DATA COLLECTION AND ESTIMATION PHASES IN SURVEYS WITH NONRESPONSE

Carl-Erik Särndal[1], **Imbi Traat**[2], **Kaur Lumiste**[3]

ABSTRACT

Inference in surveys with nonresponse has been studied extensively in the literature with a focus on the estimation phase. Propensity weighting and calibrated weighting are among the adjustment methods used to reduce the nonresponse bias. The data collection phase has come into focus more recently; the literature on adaptive survey design emphasizes representativeness and degree of balance as desirable properties of the response obtained from a probability sample. We take an integrated view where data collection and estimation are considered together. For a chosen auxiliary vector, we define the concepts *incidence* and *inverse incidence* and show their properties and relationship. As we show, incidences are used in balancing the response in data collection; the inverse incidences are important for weighting adjustment in the estimation.

Key words: adaptive survey design, auxiliary vector, incidence, inverse incidence, nonresponse adjustment, response imbalance.

1. Introduction

Weighting techniques are important in producing statistics from sample surveys. Units under-represented in the sample ought to be given a higher weight in the estimation, those over-represented should get a lower weight. This intuitive understanding was probably practiced well before theoretical advancement in the 1930's made it formal: Unbiased estimation in stratified sampling calls for weighting units by the inverse of the stratum sampling rate; the rates may differ considerably between strata. Later and more generally, the Horvitz-Thompson estimator principle established that if the sampling design gives inclusion probability π_k to unit k, then the weights $1/\pi_k$ will grant *design unbiased estimation* of a population total. That holds in the absence of nonresponse. This principle has had a great impact on survey methodology for at least 60 years, and continues to be a backbone for methodology, particularly in national statistical institutes, despite heavy unit nonresponse affecting many surveys today, especially those of individuals and households (Bethlehem et al., 2011).

[1] Statistics Sweden. Sweden. E-mail: carl.sarndal@telia.com.
[2] Institute of Mathematics and Statistics, University of Tartu, Estonia. E-mail: imbi.traat@ut.ee.
[3] Questro Analytics Ltd., Tartu, Estonia. E-mail: kaur.lumiste@eesti.ee.

When we come to surveys with nonresponse, specifically to NMAR (not missing at random) nonresponse, weighting techniques continue to be attractive and important, but are less successful in that estimates are no longer unbiased. An inspection of the realized set of respondents may reveal that certain types of sample units are markedly under- or over-represented. Weighting is used to compensate for this, then called "weighting adjustment". Intuitively, this can reduce bias, perhaps considerably, compared with a passive attitude of a flat weighting, as when we simply use the respondent mean multiplied by the population size. But weighting adjustment will not fully eliminate the bias.

A comprehensive review of nonresponse weighting adjustment was presented by Brick (2013). He identifies three major themes in nonresponse research: (a) Study of the response mechanism; (b) Data collection methods to reduce damage by nonresponse, (c) Adjustment of the survey weights to adjust for survey nonresponse. We are concerned in this article with (b) and (c), and more particularly with the interaction between them. As Brick (2013, p. 347) also notes, a deeper understanding of nonresponse in surveys is prevented by the complexity of the survey process; many unknown factors contribute to it.

With the considerable attention paid recently to responsive (or adaptive) survey design, the practice of weighting comes into a new light. Such designs can bring a more appropriate final set of respondents, compared with a stationary design where the data collection obeys a fixed unchanging protocol from beginning to end. A better balanced response is, potentially, a better starting point for the weighting adjustment in the estimation phase. A review of the literature of adaptive and responsive survey designs is found in Tourangeau et al. (2017). They also suggest directions for further improvement of such designs, and for data collection management more generally.

An adaptive data collection does not follow a stationary protocol. Interventions may take place during the data collection period. Representativeness and low imbalance are general objectives for the ultimate set of respondents. The *R-indicator* of Schouten et al. (2009) is a measure of the former concept. In a similar vein, Särndal (2011), Särndal and Lundquist (2014) used the *Imbalance* statistic to monitor the data collection. Representativeness and balance are related. Both are measured with respect to an auxiliary vector composed of auxiliary variable values known at least for the sample units, possibly for all population units.

Response propensity is another important concept for the data collection. It is a conditional response probability, given the auxiliary vector (Schouten et al., 2011). It is thus a theoretical quantity, defined either at the population level or at the sample level. It can be estimated from a response set. In adaptive design, the response propensity of the sample units is evolving during the data collection period, in tune with of the recruitment protocol changes (Olson and Groves 2012, Schouten et al. 2011).

Until recently, the data collection phase and the estimation phase have been seen largely as separate fields of research. Estimation under nonresponse has a long history and a large literature, namely, on how to apply statistical estimation theory to get the best possible – least biased – estimates, with the "frozen" set of respondents that the data collection happened to give, let alone how "good" or "representative" that set of respondents may be.

With the recent attention paid to adaptive design for the data collection, the need has arisen to know more about how a representative or well-balanced response may help the search for less biased estimates. Designs that optimize collection and adjustment simultaneously need to be developed (Kaminska 2013, p. 356).

We discuss terminology and concepts important for the two phases, the data collection and the estimation. We focus on the interrelation of the two phases and explore the connections that exist, via a multivariate auxiliary vector, between a realized set of respondents and the full (unrealized) probability sample. We see the response not as fixed and frozen but as dynamic, subject to change through the adaptive data collection. Important concepts introduced and studied in Sections 2 and 3 of the paper are *incidence* (of different types of sample units) and *inverse incidence*. The former is used for balancing the response during the data collection period, see Section 4, the latter for weighting responding units at the estimation stage, see Section 5. The two concepts do not necessarily assume a probabilistic response mechanism. A concluding discussion is the topic in Section 6.

2. Response Set and Sample Set: One Reflected in the Other

Suppose the survey data collection has resulted in a non-empty response set *r*, out of a probability sample *s* drawn from the population $U = \{1, \cdots, k, \cdots, N\}$; $r \subset s \subset U$. The response *r* is the set of units *k* having delivered the value y_k of the study variable *y*. The survey may have several study variables; the discussion and the formulas will necessarily focus on one. The sample *s* is drawn from *U* so that unit *k* has the known inclusion probability $\pi_k > 0$ and the sampling weight $d_k = 1/\pi_k$. The mechanism that generates *r* from *s* is unknown. The (sample-weighted) survey response rate is $P = \sum_{k \in r} d_k / \sum_{k \in s} d_k$, where $0 < P < 1$ is assumed.

2.1. The Auxiliary Vector

In the nonresponse context, three types of variables play a role: The study variable (continuous or categorical) *y* has values y_k observed for $k \in r$ only, and used to estimate the population total $Y = \sum_{k \in U} y_k$. The response indicator I has value $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s - r$.

The auxiliary vector x with value x_k is available at least for $k \in s$, possibly for $k \in U$. The $J \geq 1$ variables in the vector x can be continuous or categorical. They are recorded from registers or available as paradata from the data collection process. An early use of the latter information is in Politz and Simmons (1949), a more recent one in Beaumont (2005).

Since x_k is known for $k \in s$ we can note, in an ongoing data collection, which values x_k of the sample units are over-represented (have high incidence) in the realized response *r*, and which are under-represented (have low incidence). At the end of data collection, we can analyse the final response outcome with respect the specified vector x.

In an important special case, all auxiliary variables are categorical. We denote the number of distinct values x_k by *M*, a number possibly different from the vector

dimension *J*. More particularly, x can be a group vector, that is, of the form $\mathrm{x}_k = (0, \ldots, 1, \ldots, 0)'$ with a single entry "1" to indicate the group membership of *k*. Then $J = M$. For other kinds of x-vector, $J < M$, where *M* may be considerably greater than *J*.

To illustrate, if x represents a crossing of 2 sexes, 3 exhaustive education categories and 4 exhaustive age categories, then x is a group vector with dimension $J = 2 \times 3 \times 4 = 24$ and $J = M = 24$. If the same three variables are used to define instead the auxiliary vector x with sex and education crossed, while the categorical age is coded as one of (1,0,0), (0,1,0), (0,0,1) and (0,0,0), then the dimension is only $J = 2 \times 3 + 3 = 9$, but *M* is unchanged at 24.

We assume that all x-vectors used here have the following feature: There exists constant vector μ (not depending on *k*) such that

$$\mu' \mathrm{x}_k = 1 \text{ for all } k. \tag{1}$$

Most vectors of interest satisfy this requirement. When x is a group vector, the vector μ with all elements equal to "1" satisfies (1). In the example above, where x has sex and education crossed, and age contributing three more positions, the vector $\mu = (1,1,1,1,1,1,0,0,0)'$ satisfies (1). The reason for the requirement is convenience in many derivations.

2.2. The Response Described by the Incidence of the Sampled Units

To say that the response *r* is a subset of the sample *s*, and to say, inversely, that the set *s* contains *r*, are weak and uninformative descriptions of the relationship between *r* and *s*. Their relationship is made more explicit through the intermediary of chosen vector x and its values x_k known for $k \in s$. No assumptions about the probabilistic nature of the response mechanism are needed in this description.

Given *r* and an x-vector, we ask: What values f_k, attached to the sample units $k \in s$, will give agreement with the observed response mean $\bar{\mathrm{x}}_r = \sum_{k \in r} d_k \mathrm{x}_k / \sum_{k \in r} d_k$? We seek f_k for $k \in s$ to satisfy

$$\sum_{k \in s} d_k f_k \mathrm{x}_k / \sum_{k \in s} d_k = \bar{\mathrm{x}}_r. \tag{2}$$

Further specification is needed to get a unique solution. One is obtained by letting f_k be linear in the x-vector: $f_k = \mathrm{A}' \mathrm{x}_k$ for some *J*-vector A. Inserting into (2), and solving, we get $\mathrm{A}' = \bar{\mathrm{x}}'_r \Sigma_s^{-1}$, where the $J \times J$ matrix

$$\Sigma_s = \sum_{k \in s} d_k \mathrm{x}_k \mathrm{x}'_k / \sum_{k \in s} d_k \tag{3}$$

is assumed non-singular. Therefore,

$$f_k = \bar{\mathrm{x}}'_r \Sigma_s^{-1} \mathrm{x}_k \,, \quad k \in s. \tag{4}$$

We call f_k the incidence (factor) of unit k. The mean incidence over s, as a consequence of (1), is $\bar{f}_s = \sum_{k \in s} d_k f_k / \sum_{k \in s} d_k = 1$. The variance over s,

$\sum_{k \in s} d_k (f_k - \bar{f}_s)^2 / \sum_{k \in s} d_k$, is minimal under the constraint in (2). The proof is in the Appendix.

Units with the same value of x_k share the same incidence f_k. In the simple example where gender is the only x-variable, we have $J = M = 2$, $\mathrm{x}_k = (1,0)'$ for all men, $\mathrm{x}_k = (0,1)'$ for all women. Then (4) says that all sampled men have the

incidence $f_k = P_{\text{men}}/P$, all sampled women have $f_k = P_{\text{women}}/P$, where P_{men} and P_{women} are the gender response rates and P the overall rate. This crude kind of response analysis describes how the response for men differs from that of women.

For x-vectors typically used in practice, the number M of distinct values can be large. The response rate within groups of units with the same x_k-value is replaced by the wider concept generalized response rate, $P_k = P \times f_k$, which can also be seen as an estimated response propensity for unit k characterized by x_k. The mean of P_k over s is $P\bar{f}_s = P$, the overall response rate.

2.3. The Sample Described by the Inverse Incidence of the Responding Units

After a completed data collection, the composition of the response *r* can no longer be changed or influenced. We can describe the relationship between *r* and *s* by the *inverse incidence.* The direction here is to make the smaller set *r* conform to the larger set *s*, by weighting the units in *r*.

We ask: What numbers g_k applied to the responding units will reproduce the auxiliary sample mean $\bar{\mathrm{x}}_s = \sum_{k\in s} d_k \mathrm{x}_k / \sum_{k\in s} d_k$? It is futile to ask that question for y_k, because it is missing for $k \in s - r$. This is the inverse of the question in the preceding section. We seek g_k for $k \in r$ to satisfy

$$\sum_{k\in r} d_k g_k \mathrm{x}_k / \sum_{k\in r} d_k = \bar{\mathrm{x}}_s. \tag{5}$$

There is no unique solution. One solution is obtained by forming g_k as a linear combination of the x-variables: For some *J*-vector B, set $g_k = \mathrm{B}'\mathrm{x}_k$. Inserting into (5), solving for B, and assuming that

$$\Sigma_r = \sum_{k\in r} d_k \mathrm{x}_k \mathrm{x}_k' / \sum_{k\in r} d_k \tag{6}$$

is non-singular, we get

$$g_k = \bar{\mathrm{x}}_s' \Sigma_r^{-1} \mathrm{x}_k,\ k \in r. \tag{7}$$

We call g_k the *inverse incidence* (*factor*), or weight, of unit $k \in r$. The mean over *r* is $\bar{g}_r = \sum_{k\in r} d_k g_k / \sum_{k\in r} d_k = 1$, using (1). The variance over *r*, $\sum_{k\in r} d_k (g_k - \bar{g}_r)^2 / \sum_{k\in r} d_k$, is minimal under the constraint in (5). The proof is analogous to the corresponding one for f_k, which is given in the Appendix. Note that g_k is computable for all $k \in s$, because x_k is available for $k \in s$.

3. Properties of Incidence and Inverse Incidence

3.1. The Moments and the Interrelation

The equation (2) makes a sample *s* conform to a realized response *r* through the *incidence factor f* with values $f_k = \bar{\mathrm{x}}_r' \Sigma_s^{-1} \mathrm{x}_k$ given in (4) for $k \in s$. The equation (5) makes an "upweighted" response *r* conform with a given sample *s* through the *inverse incidence factor* (or weight factor) *g* with values $g_k = \bar{\mathrm{x}}_s' \Sigma_r^{-1} \mathrm{x}_k$ given in (7) for $k \in s$. The values $f_k \times \mathrm{g}_k$ for $k \in s$ define the *product factor*.

Example. Let x be a group vector of dimension *J*, $\mathrm{x}_k = (0, \ldots, 1, \ldots, 0)'$, coding the same number of different groups of sample units. Suppose that *s* is a self-weighting fixed size *n* sample. Then $d_k = N/n$ for all *k*, and $m\bar{x}_r =$

$(m_1, \dots, m_j, \dots, m_J)'$, where m_j is the number of responding units in group *j*. Alternatively expressed, m_j is the size of the *j*:th response group r_j, and $m = \sum_{j=1}^{J} m_j$ is the size of *r*. From (4) and (7) we obtain $f_k = P_j/P$, $g_k = P/P_j$ for all units *k* in the same sample group s_j, where $P = m/n$, $P_j = m_j/n_j$ is the group j response rate and n_j is the size of s_j, $j = 1, \dots, J$. Hence, when x is a group vector, g_k is the inverse of f_k in an exact numerical sense: $f_k g_k = 1$ for every *k*. □

In practice, the incidences f_k for $k \in s$ are used at the data collection phase, as tools for an adaptive data collection to create a well-balanced final response. This is reviewed in Section 4. The inverse incidences g_k are used in the estimation phase for weighting adjustment. This is the topic of Section 5. Here we present general properties of f_k and g_k.

We derive mean and variance of f_k, g_k and of their product $f_k \times g_k$, over the response and over the full sample. For the f factor, these moments are defined as

$$\bar{f}_r = \text{mean}_r(f) = \sum_{k\in r} d_k f_k \,/ \sum_{k\in r} d_k,\ \bar{f}_s = \text{mean}_s(f) = \sum_{k\in s} d_k f_k \,/ \sum_{k\in s} d_k, \qquad (8)$$

$$\text{var}_r(f) = \sum_{k\in r} d_k (f_k - \bar{f}_r)^2 \,/ \sum_{k\in r} d_k,\ \text{var}_s(f) = \sum_{k\in s} d_k (f_k - \bar{f}_s)^2 \,/ \sum_{k\in s} d_k. \qquad (9)$$

For the corresponding moments of the g factor, replace f by g. For the product factor, replace f by $f \times g$ and f_k by $f_k \times g_k$ in (8) and (9).
The moments of the three factors are shown in Table 1 for an arbitrary vector x. Some of the table entries involve quadratic forms in the vector difference $\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s$:

$$Q_s = (\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s)' \Sigma_s^{-1} (\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s); \qquad Q_r = (\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s)' \Sigma_r^{-1} (\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s), \qquad (10)$$

where the $J \times J$ weighting matrices Σ_s and Σ_r (non-singular) are given by (3) and (6). Four of the variances have less transparent expressions and are shown only as concepts.

Table 1. Mean and variance of *f*, g and $f \times g$. The quantities Q_r and Q_s are given in (10).

Factor	mean in *s*	mean in *r*	variance in *s*	variance in *r*
f	1	$1 + Q_s$	Q_s	$\text{var}_r(f)$
g	$1 + Q_r$	1	$\text{var}_s(g)$	Q_r
$f \times g$	1	1	$\text{var}_s(f \times g)$	$\text{var}_r(f \times g)$

The properties in Table 1, used in later sections, follow from the definitions in (8) and (9) by standard matrix and vector manipulations, using also $\bar{\mathrm{x}}_s' \Sigma_s^{-1} \mathrm{x}_k = \bar{\mathrm{x}}_r' \Sigma_r^{-1} \mathrm{x}_k = 1$ for all k, and $\bar{\mathrm{x}}_r' \Sigma_s^{-1} \bar{x}_s = \bar{\mathrm{x}}_s' \Sigma_r^{-1} \bar{x}_r = 1$; these follow from (1).

By Table 1, $\bar{f}_r = 1 + Q_s \geq 1 = \bar{f}_s$. Equality holds only for $Q_s = 0$, implying $\bar{\mathrm{x}}_r = \bar{\mathrm{x}}_s$. In general $f_k \times g_k \neq 1$ for any particular unit k , but Table 1 shows that the mean of the products $f_k \times g_k$ is 1, over s as well as over r. This interesting property says that one factor is the inverse of the other, in a generalized sense. In

the group vector case the inverse relationship holds in an exact numerical sense, $f_k g_k = 1$ for every k.

The covariances are

$$\mathrm{cov}_s(f,g) = \mathrm{mean}_s(f \times g) - \bar{f}_s \bar{g}_s = 1 - 1 \times (1 + Q_r) = -Q_r < 0, \quad (11)$$

$$\mathrm{cov}_r(f,g) = \mathrm{mean}_r(f \times g) - \bar{f}_r \bar{g}_r = 1 - (1 + Q_s) \times 1 = -Q_s < 0 \quad (12)$$

Hence, f_k and g_k are negatively correlated, over *s* as over *r*. More specifically, the coefficient of correlation over *s* is usually large negative, not far from -1. This claim is justified by an approximation shown in the Appendix, whereby

$$\mathrm{corr}_s(f,g) \approx -1/(1 + Q_s). \quad (13)$$

The right-hand side is greater than -1, but not far from -1, because compared with 1, Q_s is small positive. The approximation in (13) may not be highly accurate for all outcomes *r*, given *s*, but a large negative correlation is indicated.

The covariances with the auxiliary vector are

$$\mathrm{cov}_s(f,\mathrm{x}) = \textstyle\sum_{k\in s} d_k (f_k - 1)(\mathrm{x}_k - \bar{\mathrm{x}}_s) / \sum_{k\in s} d_k = (\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s), \quad (14)$$

$$\mathrm{cov}_r(g,\mathrm{x}) = \textstyle\sum_{k\in r} d_k (g_k - 1)(\mathrm{x}_k - \bar{\mathrm{x}}_r) / \sum_{k\in r} d_k = -(\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s). \quad (15)$$

It is interesting to note that $\mathrm{cov}_s(f,\mathrm{x}) = -\mathrm{cov}_r(g,\mathrm{x})$.

The fit of a linear regression with intercept of g_k on f_k over $k \in s$ gives the slope coefficient $b = \mathrm{cov}_s(f,g)/\mathrm{var}_s(f) = -Q_r/Q_s$ and the intercept $a = \bar{g}_s - b\bar{f}_s = 1 + Q_r + Q_r/Q_s$. The predicted g_k-value from this linear fit is $\hat{g}_k = a + bf_k$, so for every $k \in s$ we have the equation

$$(\hat{g}_k - 1)/Q_r + (f_k - 1)/Q_s = 1. \quad (16)$$

3.2. Empirical Illustration of the Relationship

Figure 1 illustrates the relationship between the f- and g-factors in a specific experiment. From a data set collected in an Estonian household survey a simple random sample s of 700 households (HH) was drawn and then kept fixed. A number of characteristics of each household and head of household (HD) were recorded. Response probabilities ϕ_k (where *k* designates a household) were then computed for $k \in s$ by the model

logit(ϕ) = 5 – 4 × *HD sex* + 2 × *HD employment status* – 0.0004 × *HH income.*

Here, *HD sex* (1 for woman, 0 for man) and *HD employment status* (1 for employed, 0 for unemployed) are dichotomous; *HH income* is continuous. The model deliberately assigns lower response probability to high income households where the head is unemployed female. One single response set *r*, with response rate $P = 60\%$, was realized by giving household *k* the response probability ϕ_k. Given that set *r*, computations were carried out with the vector

x = (HD education, HD sex, HH size, HH children, HD employment status, HH expenditure).

Here, *HD education*, with 3 exhaustive categories, was coded as (1,0,0), (0,1,0) and (0,0,1). The variables *HH size* and *HH children* (the number of children in household) are discrete univariate; *HH expenditure* is continuous. The dichotomous *HD sex* and *HD employment status* are as explained earlier. This x is not a group vector, so the inverse relationship $g_k = 1/f_k$ will not hold with exactness for all *k*, but it does so to the degree of approximation that Figure 1 illustrates. The dimension of x is 8: The first variable occupies 3 positions, the other 5 variables one position each. The response set *r* has considerable imbalance; *IMB* = 0.055, computed by (17) below.

The f- and g-factors were computed on the realized *r* and *s*. The 700 points (f_k, g_k) for $k \in s$ are plotted (as hollow small circles) in Figure 1. The figure illustrates that f_k can be negative for a small number of units $k \in s$. In the figure, none of the points with $f_k < 0$ belong to *r.* Consequently, the linear approximation of g_k through f_k works quite well in the response set *r.* The solid line is the linear regression line $g = a + bf$, with $a = 1 + Q_r + Q_r/Q_s$= 3.145 and $b = -Q_r/Q_s$=−1.863. The dashed curve is $g = 1/f$. We verified empirically, for the group vector x = (*HD education* × *HD employment status*), with 3×2 = 6 groups, that $g_k = 1/f_k$ holds exactly for all *k*, as it should.

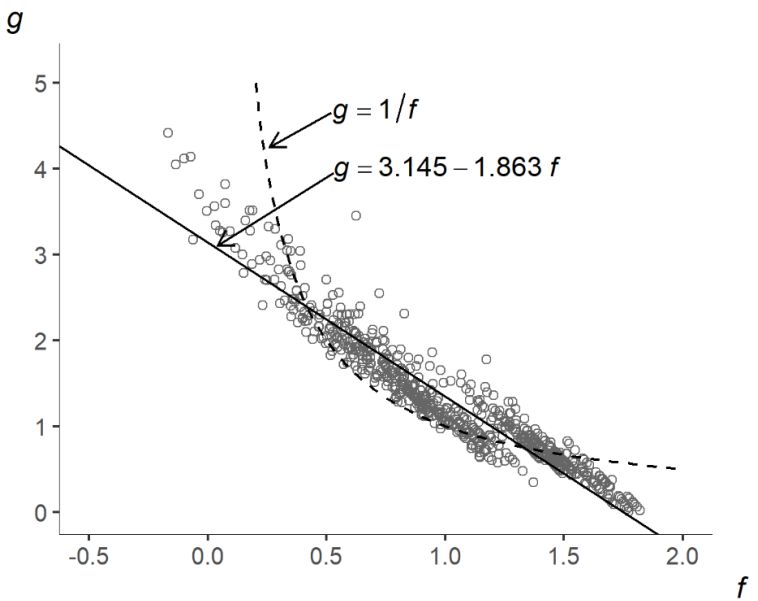


Figure 1. Relationship between f- and g-factors for a sample of size 700. Each circle represents a sample element.

4. Achieving Low Imbalance in the Data Collection

The incidences f_k are important for the data collection. They are used for creating a well balanced response set. The response *r* is called perfectly balanced with respect to the vector x if $\bar{x}_r = \bar{x}_s$ (Särndal, 2011). It follows from (2)

that the equality in means is achieved if $f_k = 1$ for all *k*. The equality $\bar{\mathrm{x}}_r = \bar{\mathrm{x}}_s$ also holds if $g_k = 1$ for all *k*, as seen from (5). To get a perfectly balanced response *r* is a distant possibility in a survey data collection, especially for a long x vector. We can strive to come close. But ordinarily, a perfect balance is not achieved. Since $\bar{\mathrm{x}}_r - \bar{\mathrm{x}}_s$ is a vector, a scalar measure of the difference is created, called the *imbalance* of the response *r* with respect to the vector x for the given sample *s*,

$$\mathrm{IMB}(r, \mathrm{x}|s) = P^2 Q_s, \tag{17}$$

where P is the response rate and Q_s is given in (10) (Särndal, 2011; Lundquist and Särndal, 2013; Särndal and Lundquist, 2014). Although $\mathrm{IMB}(r, \mathrm{x}|s)$ is more descriptive, we shall use for simplicity the notation IMB. For any *r*, *s* and vector x, $0 \leq \mathrm{IMB} \leq P(1-P) \leq 0.25$. For most survey data, IMB does not come close to the upper bound $P(1-P)$; typical values are in the range 0.03 to 0.06.

A measure related to IMB is the R-indicator, with R for "representativeness" (Bethlehem et al., 2011). It is different in its background, which is estimation of response probabilities assumed to exist for all population units.

The incidences f_k, computable for all $k \in s$, are tools for an adaptive data collection aiming at an ultimate response set *r* with low imbalance. A property making this possible is that the variance (computed over *s*) of the (estimated) propensities $P_k = Pf_k$ is equal to the imbalance, $\mathrm{IMB} = P^2 Q_s$ (see Table 1). The P_k can be computed continuously during an ongoing data collection period. Therefore, an avenue to low imbalance in the final response *r* is to manage the data collection to achieve in the end a low variance of P_k, and therefore low IMB. There may be several ways to accomplish this. One is the threshold method proposed in Särndal and Lundquist (2014), which we now describe.

The data collection, which may last several days or weeks, is seen as a dynamic process where inspections and change of protocol may take place, at specified points. For example, one may decide, at a certain point, to focus the continued data collection on specific types of units, say those that are so far underrepresented.

In the threshold method, the propensities $P_k = Pf_k$ are computed for $k \in s$ at several points, say four to six, in the data collection period, and with a "monitoring vector" x designated for this purpose.

At the first inspection point, units with propensity greater than a fixed threshold, say 0.60, are set aside and not further contacted during the period. Contact attempts continue with the remaining non-responding sample units; as a result more units join the response set. At the second inspection point, P_k is computed again for all $k \in s$, and some more units, those with the new propensity P_k greater than 0.60, join those already set aside. This pattern is repeated at each remaining inspection point; at each of these some more units are set aside. Non-responding units remaining at the last inspection point are pursued until the very end of the data collection period. By the mechanics of this procedure, the variability of the propensities - and therefore the imbalance IMB - is more and more reduced. In the end, the imbalance IMB can be quite low. Alternative adaptive designs can be constructed with a similar objective.

5. The Estimation Stage

After a completed data collection, it remains to produce estimates of important finite population parameters, such as the population total $Y = \sum_{k\in U} y_k$, using the values y_k available for $k \in r$. The estimates are design biased, more or less.

If individual response probabilities ϕ_k were known, then $\hat{Y}_{2\text{ph}} = \sum_{k\in r} d_k \phi_k^{-1} y_k$ would be unbiased for the total $Y = \sum_{k\in U} y_k$. This claim derives from design-based theory for two-phase selection: First a probability sample *s* from *U*, then a response *r* from the given *s*. Since ϕ_k is unknown, $\hat{Y}_{2\text{ph}}$ should be adjusted. Brick (2013) reviews three types of weighing adjustment procedures in surveys with nonresponse. In the first of these, the unknown individual response probabilities ϕ_k in $\hat{Y}_{2\text{ph}}$ are replaced by estimates $\hat{\phi}_k$. This results in

$$\hat{Y}_{\text{ADJ}} = \sum_{k\in r} d_k \hat{\phi}_k^{-1} y_k \ , \tag{18}$$

also referred to as "quasi-randomization" estimators. Access to suitable auxiliary variables and the choice of the model for the response mechanism play an important role in (18).

Brick's (2013) second type is the weighting class estimator. It is a special case of (18), where $\hat{\phi}_k^{-1}$ is equal to the inverse of a group response rate. That is, if the sample *s* is divided into *J* mutually exclusive and exhaustive subgroups s_j with r_j as the responding subset of s_j, $j = 1, \ldots, J$, then $\hat{\phi}_k^{-1} = \sum_{k\in s_j} d_k / \sum_{k\in r_j} d_k$, common to all units *k* in a group.

The third weighting adjustment estimator in Brick's (2013) review is the calibration estimator. It differs in its construction from (18) but is still unmistakably design-based in its orientation. All three weighting adjustment procedures are imperfect under nonresponse because they fail to meet the design-based criterion of unbiased estimation.

Here, we distinguish three arguments for constructing an estimator for $Y = \sum_{k\in U} y_k$. They are: Weighting by inverse incidence (Section 5.1), calibration estimation (Section 5.2) and estimation by explicit modelling/prediction (Section 5.3).

5.1. Weighting by Inverse Incidence

Weighting by inverse incidence does not require any response model. It reflects the intuitive idea that units in *r* with low incidence get relatively higher weight, and vice versa.

The incidence factor f_k is given in (4), the inverse incidence factor g_k in (7). Now put

$$P_k = P f_k;\ \nu_k = P^{-1} g_k, \tag{19}$$

where P is the overall response rate. P_k and ν_k are each other's inverse, in that the mean of their product, $P_k \nu_k = f_k g_k$, is equal to one, over *r* and over *s* (see Table 1). The inverse incidence weighting estimator of $Y = \sum_{k\in U} y_k$ is then given by

$$\hat{Y}_{\text{WEI}} = \sum_{k\in r} d_k \nu_k y_k. \tag{20}$$

This is weighting adjustment as in (18), if we let $\hat{\phi}_k^{-1} = \nu_k$. Moreover, P_k is reminiscent of a second phase inclusion probability for unit *k*, that is, in "drawing" the response set *r* from *s*. The sample mean of P_k is $\bar{P}_s = \sum_{k\in s} d_k P_k / \sum_{k\in s} d_k = P\bar{f}_s = P$, the overall response rate.

The weighting in (20) is motivated purely by inverse incidence, based on a given x-vector, with no particular variable *y* in mind. The same weights are applied to all variables *y*, whatever their special characteristics. This is appealing in surveys where many *y*-variables require estimation, none of them deemed to be truly more important or different in nature. Implicit in the inverse incidence weighting is a relationship between the 0/1 indicator of the response and the auxiliary vector x that determines the incidence f_k.

5.2. Calibration Estimation

A well-known weighting adjustment estimator is the calibration estimator. Weighting is based on x with implicit *y*-to-x relationship. Still, all *y*-variables are typically given the same weighting. For comparability reasons, we consider calibration up to *s*. Weight factors u_k are calibrated "from *r* up to *s*", to satisfy the calibration equation

$$\sum_{k\in r} d_k u_k \mathrm{x}_k = \sum_{k\in s} d_k \mathrm{x}_k. \tag{21}$$

The resulting calibration estimator is then

$$\hat{Y}_{\mathrm{CAL}} = \sum_{k\in r} d_k u_k y_k. \tag{22}$$

If we choose u_k to be linear in x_k, $u_k = \lambda' \mathrm{x}_k$, it follows from the derivation in Section 2.3 that $u_k = P^{-1} g_k$, where g_k is the inverse incidence given in (7). Then, (22) is the linear calibration estimator, $\hat{Y}_{\mathrm{CALlin}}$, which we can express in several ways:

$$\hat{Y}_{\mathrm{CALlin}} = \sum_{k\in r} d_k \nu_k y_k = P^{-1} \sum_{k\in r} d_k g_k y_k = \sum_{k\in s} d_k \hat{y}_k = \hat{N}\, \bar{\mathrm{x}}_s' \mathrm{b}_r, \tag{23}$$

where $\hat{y}_k = \mathrm{x}_k' \mathrm{b}_r$ and b_r is the regression coefficient vector in a linear regression fit of y on x over *r*,

$$\mathrm{b}_r = (\sum_{k\in r} d_k \mathrm{x}_k \mathrm{x}_k')^{-1} \sum_{k\in r} d_k\, \mathrm{x}_k y_k. \tag{24}$$

Hence, the inverse incidence weighting estimator $\hat{Y}_{\mathrm{WEI}}$ in (20) has a double identity: It is at the same time a (linear) calibration estimator.

The purely mechanical aspect of the calibration approach is to deliver weights to satisfy (21) – which has an unbiased Horvitz-Thompson estimator on the right hand side – and to apply these weights in the estimation. But the purpose is also to explain the *y*-variable through the auxiliary vector x. The calibration approach is thus double-natured: The weighting aspect is combined with implicit relationship *y*-to-x. This can be seen when we examine the deviation of $\hat{Y}_{\mathrm{CALlin}}$ from the unbiased estimator requiring full response, $\hat{Y}_{\mathrm{FUL}} = \sum_{k\in s} d_k y_k$. This deviation can be written as

$$\hat{Y}_{\mathrm{CALlin}} - \hat{Y}_{\mathrm{FUL}} = -\sum_{k\in s} d_k e_k \tag{25}$$

with the residual $e_k = y_k - \mathrm{x}_k'\mathrm{b}_r$, where b_r is the regression vector given in (24). If the model fits well in the response set, the residuals are small, and $\hat{Y}_{\mathrm{CALlin}}$ based on the response is close to the unbiased $\hat{Y}_{\mathrm{FUL}}$.

Calibration estimators have been extensively studied for the last 20 years. One direction is to use information both in the sample and population levels. Another direction is to use non-linear forms of calibration. Some references are Deville (1998), Deville and Särndal (1992), Folsom and Singh (2000), Estevao and Särndal (2000), Montanari and Ranalli (2003, 2005, 2012), Särndal and Lundström (2005), Chang and Kott (2008), Kott and Chang (2010), Kott and Liao (2012).

5.3. Estimation by Explicit Modelling/Prediction

The modelling/prediction approach is based on replacing missing *y*-values by the best possible substitutes that statistical theory can offer. This argument is, on surface at least, very different from both incidence weighting and calibration weighting. Its importance is illustrated by Little's (2013) discussion of Brick (2013).

This approach focuses directly on one *y*-variable at a time. From an explicitly formulated (linear or non/linear) model for the *y*-to-x relationship, and a fit of that model based on (y_k, x_k) for $k \in r$, predicted values are obtained for the non-observed y_k, using the values x_k known for $k \in s - r$. Observed y_k together with predictions $\hat{y}_k$ are used to build the estimator of the population *y*-total,

$$\hat{Y}_{\mathrm{PRED}} = \sum_{k\in r} d_k y_k + \sum_{k\in s-r} d_k \hat{y}_k. \qquad (28)$$

Examination of the design-based behaviour of $\hat{Y}_{\mathrm{PRED}}$ has shown that strong regression relationship holds good prospects for a considerable reduction of the (design-based) nonresponse bias. Early references are Bethlehem (1988) and Cassel et al. (1983).

A variety of models and methods can be entertained to get the predicted values $\hat{y}_k$. A simple application is by ordinary linear regression fit of *y* on x, resulting in the regression vector b_r in (24) and predicted values $\hat{y}_k = \mathrm{x}_k'\mathrm{b}_r$ for $k \in s$. Note that $\sum_{k\in r} d_k(y_k - \hat{y}_k) = 0$ because of (1). Then

$$\hat{Y}_{\mathrm{PREDlin}} = \sum_{k\in r} d_k y_k + \sum_{k\in s-r} d_k \hat{y}_k = \sum_{k\in s} d_k \hat{y}_k = \hat{Y}_{\mathrm{CALlin}} = \hat{Y}_{\mathrm{WEI}}. \qquad (29)$$

It can also be seen as a result of the linear generalized regression (GREG) construction;

$$\hat{Y}_{\mathrm{GREG}} = \sum_{k\in s} d_k \hat{y}_k + \sum_{k\in r} d_k(y_k - \hat{y}_k) = \sum_{k\in s} d_k \hat{y}_k. \qquad (30)$$

Hence the inverse incidence weighting estimator $\hat{Y}_{\mathrm{WEI}}$ in (20) has multiple identities: It is at the same time (a) a calibration estimator, (b) a prediction estimator, and (c) a GREG estimator. It is important to note that this equivalence happens under the linear formulation, and under the x-vector condition in (1).

We can link the bias to the tendency of nonresponse to misrepresent the regression relationship: Denote by $\mathrm{b}_s = (\sum_{k\in s} d_k \mathrm{x}_k \mathrm{x}_k')^{-1} \sum_{k\in s} d_k \mathrm{x}_k y_k$ the regression coefficient vector in the linear fit of *y* on x over *s*. Then, by (1), $\hat{N}\,\bar{\mathrm{x}}_s'\mathrm{b}_s = \hat{N}\bar{y}_s = \sum_{k\in s} d_k\, y_k = \hat{Y}_{\mathrm{FUL}}$ and the deviation from the unbiased estimation can be written as

$$\hat{Y}_{\mathrm{PREDlin}} - \hat{Y}_{\mathrm{FUL}} = \hat{Y}_{\mathrm{CALlin}} - \hat{Y}_{\mathrm{FUL}} = \hat{N}\,\bar{\mathrm{x}}_s'(\mathrm{b}_r - \mathrm{b}_s), \qquad (31)$$

where b_r is given in (24). As is well known from regression theory, the selection effect is likely to distort an estimated regression relationship, that is, to make the regression vectors b_r and b_s differ considerably, and thus $\hat{Y}_{\mathrm{PREDlin}}$ to differ from the unbiased $\hat{Y}_{\mathrm{FUL}}$. Särndal et al. (2016) evaluate the deviation $\Delta = (\hat{Y}_{\mathrm{CALlin}} - \hat{Y}_{\mathrm{FUL}})/\hat{N} = \bar{\mathrm{x}}_s'(\mathrm{b}_r - \mathrm{b}_s)$ under certain assumptions, and find potential for improved accuracy under adaptive design. Expressions for the design-based bias have been derived for some types of regression-based estimators (Fuller et al. 1994, Särndal and Lundström 2005, Brick and Jones 2008).

In the model-based version of the modelling/prediction approach, the sampling design and the sampling weights d_k may not enter at all. A comprehensive coverage is found in books such as Valliant et al. (2000) and Chambers et al. (2012). Other recent contributions are Breidt and Opsomer (2000), Breidt et al. (2005), Little (1986).

6. Conclusion

We have examined a survey setting where nonresponse is occurring in a probability sample from the finite population. We emphasized an integrated view, in which the data collection and the estimation stage can benefit from each other, and support each other, in making inference about the population.

We have assumed that an appropriate auxiliary vector was formulated, from the available supply of auxiliary variables, categorical or continuous. We discussed the auxiliary vector's important role in forming a bridge between a realized set of respondents and the full probability sample. To that end, we formulated the concepts of *incidence* and *inverse incidence* of the sample units. A realized response set can be described by the (computable) incidences of the sample units; vice versa, the drawn sample can be described by the (also computable) inverse incidences of the responding units.

As we showed, the incidences are used in an adaptive data collection to realize a final response set with low imbalance. The inverse incidences are used at the estimation stage, for building a weighted estimator. It is one that does not use any assumptions about a probabilistic response mechanism. We pointed out that it coincides, in the special case of a "linear formulation", with estimators derived by other approaches: Calibration, modelling/prediction and GREG. These approaches have branched out in their own directions and have generated a stream of literature that we do not review here.

To a considerable degree, this article has dealt with concepts and principles. This has left unanswered a number of other important aspects. Among these is the question whether a reduced imbalance in the ultimate response set will lead to reduced bias in the estimates, over and beyond what (weighting) adjustment alone can accomplish at the estimation stage. There is some positive evidence in this direction in the recent literature. A relationship between auxiliary vector x and survey variable *y* is implicitly assumed; one can say that balancing the survey response gives some added protection against large nonresponse bias. Recent articles in this direction are Schouten et al. (2016) and Särndal et al. (2016). Also, Tourangeau et al. (2017) confirm that a bias reduction, although perhaps marginal, can be realized by balancing, and these authors claim that further

improvement may be possible, through alternative and better adaptive designs. These and other recent contributions to the literature underline the need for an integrated view, one where data collection and estimation are considered together; in this article, we have also taken a step in that direction.

Acknowledgements

This work was partly supported by the Institutional Research Funding IUT34-5 of Estonia.

REFERENCES

BEAUMONT, J. F., (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. Survey Methodology, 31, pp. 227–231.

BETHLEHEM, J. G., (1988). Reduction of nonresponse bias through regression estimation. Journal of Official Statistics, 4, pp. 251–260.

BETHLEHEM, J. G., COBBEN, F., SCHOUTEN, B., (2011). Handbook on Nonresponse in Household Surveys. New York: Wiley.

BREIDT, F. J., CLAESKENS, G., OPSOMER, J. D., (2005). Model-assisted estimation for complex surveys using penalised splines. Biometrika, 92 (4), pp. 831–846.

BREIDT, F. J., OPSOMER, J. D., (2000). Local polynomial regression estimators in survey sampling. Annals of Statistics, 28 (4), pp. 1026–1053.

BRICK, J. M., (2013). Unit nonresponse and weighting adjustments: A critical review. Journal of Official Statistics, 29 (3), pp. 329–353.

BRICK, J. M., JONES, M. E., (2008). Propensity to respond and nonresponse bias. Metron, 66 (1), pp. 51–73.

CASSEL, C., SÄRNDAL, C. E., WRETMAN, J., (1983). Some uses of strategical models in connection with the nonresponse problem. In: Incomplete Data in Sample Surveys, ed. By W. G. Madow and I.Olkin, Vol. 3, New York: Academic Press.

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

CHANG, T., KOTT, P. E., (2008). Using calibration weighting to adjust for nonresponse under a plausible model. Biometrika, 95 (3), pp. 555–571.

DEVILLE, J. C., (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. Paper presented at Congrès de l'ACFAS, Sherbrooke, Québec.

DEVILLE, J. C., SÄRNDAL, C. E., (1992). Calibration estimation in survey sampling. Journal of the American Statistical Association, 87 (418), pp. 375–382.

ESTEVAO, V., SÄRNDAL, C. E., (2000). A functional form approach to calibration. Journal of Official Statistics, 16, pp. 379–399.

FOLSOM, R. E., SINGH, A. C., (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse and poststratification. Proceedings, Section of Survey Research Methods, American Statistical Association, Washington DC, pp. 598–603.

FULLER, W. A., LOUGHIN, M. M., BAKER, H. D., (1994). Regression weighting in the presence of nonresponse with application to the 1987-1988 nationwide food consumption survey. Survey Methodology, 20, pp. 75–85.

KAMINSKA, O., (2013). Unit nonresponse and weighting adjustments: a critical review: discussion. Journal of Official Statistics, 29, pp. 355–358.

KOTT, P. S., CHANG, T., (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. Journal of the American Statistical Association, 105 (491), pp. 1265–1276.

KOTT, P. S., LIAO, P., (2012). Providing double protection for unit nonresponse with a nonlinear calibration weighting. Survey Research Methods, 6 (2), pp. 105–111.

LITTLE, R. J. A., (2013). Discussion. Journal of Official Statistics, 29, pp. 363–366.

LITTLE, R. J. A., (1986). Survey nonresponse adjustments for estimates of means. International Statistical Review, 54 (2), pp. 139–157.

LUNDQUIST, P., SÄRNDAL, C.-E., (2013). Aspects of responsive design. With applications to the Swedish Living Conditions Survey. Journal of Official Statistics, 29, pp. 557–582.

MONTANARI, G. E., RANALLI, M. G., (2003). Nonparametric methods in survey sampling. In M. Vinci, P. Monari,, S. Mignani, A. Montanari (eds.), New Developments in Classification and Data Analysis. Berlin: Springer.

MONTANARI, G. E., RANALLI, M. G., (2005). Nonparametric model calibration estimation in survey sampling. Journal of the American Statistical Association, 100, pp. 1429–1442.

MONTANARI, G. F., RANALLI, M. G., (2012). Calibration inspired by semiparametric regression as a treatment for nonresponse. Journal of Official Statistics, 28, pp. 239–277.

OLSON, K., GROVES, R. M., (2012). An Examination of within-person variation in response propensity over the data collection field period. Journal of Official Statistics, 28, pp. 29–51.

POLITZ, A., SIMMONS, W., (1949). An attempt to get “Not at Homes” into the sample without callbacks. Journal of the American Statistical Association, 44 (245), pp. 9–31.

SÄRNDAL, C.-E., (2011). The 2010 Morris Hansen lecture: Dealing with survey nonresponse in data collection, in estimation. Journal of Official Statistics, 27, pp. 1–21.

SÄRNDAL, C.-E., LUMISTE, K., TRAAT, I., (2016). Reducing the response imbalance: Is the accuracy of the survey estimates improved? Survey Methodology, 42 (2), pp. 219–238.

SÄRNDAL, C.-E., LUNDSTRÖM, S., (2005). Estimation in Surveys with Nonresponse. New York: John Wiley& Sons, Inc.

SÄRNDAL, C.-E., LUNDQUIST, P., (2014). Accuracy in estimation with nonresponse: a function of degree of imbalance and degree of explanation. Journal of Survey Statistics and Methodology, 2, pp. 361–387.

SCHOUTEN, B., COBBEN, F., BETHLEHEM, J., (2009). Indicators for the representativeness of survey response. Survey Methodology, 35, pp. 101–113.

SCHOUTEN, B., COBBEN, F., LUNDQUIST, P., WAGNER, J., (2016). Does more balanced survey response imply less non-response bias? Journal of the Royal Statistical Society, Series A, 179 (3), pp. 727–748.

SCHOUTEN, B., SHLOMO, N., SKINNER, C., (2011). Indicators for monitoring and improving representativeness of response. Journal of Official Statistics, 27, pp. 1–24.

TOURANGEAU, R., BRICK, J. M., LOHR, S., LI, J., (2017). Adaptive and responsive survey designs: a review and assessment. Journal of the Royal Statistical Society, Series A, 180 (1), pp. 201–223.

VALLIANT, R., DORFMAN, A. H., ROYALL, R. M., (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: John Wiley & Sons Inc.

APPENDIX

Proof that the incidence factors f_k in (4) have minimal variance subject to (2):

Using the Lagrange multiplier method, we seek the minimum of

$$\sum_{k\in s} d_k(f_k - \bar{f}_s)^2 - 2\lambda'(\sum_{k\in s} d_k f_k \mathrm{x}_k - (\sum_{k\in s} d_k)\bar{\mathrm{x}}_r). \qquad (32)$$

Setting the derivative with respect to f_k equal to zero gives

$$2d_k(f_k - \bar{f}_s) - 2d_k\lambda'\mathrm{x}_k = 0\ ; \quad f_k - \bar{f}_s = \lambda'\mathrm{x}_k. \qquad (33)$$

Determine λ from the condition in (2): $\lambda' = \bar{\mathrm{x}}_r'\Sigma_s^{-1} - \bar{f}_s\bar{\mathrm{x}}_s'\Sigma_s^{-1}$. Post-multiply by x_k and note that $\bar{\mathrm{x}}_s'\Sigma_s^{-1}\mathrm{x}_k = 1$ by (1). This gives $\lambda'\mathrm{x}_k = \bar{\mathrm{x}}_r'\Sigma_s^{-1}\mathrm{x}_k - \bar{f}_s$ and $f_k = \bar{f}_s + \lambda'\mathrm{x}_k = \bar{\mathrm{x}}_r'\Sigma_s^{-1}\mathrm{x}_k$, as given in (4).

Derivation of the approximation in (13):

By definition, $\mathrm{corr}_s(f,g) = \mathrm{cov}_s(f,g)/(\mathrm{var}_s(f)\mathrm{var}_s(g))^{1/2}$. First use $\mathrm{var}_s(g)/\bar{g}_s^2 \approx \mathrm{var}_r(g)/\bar{g}_r^2$, assuming that the coefficient of variation of g (standard deviation divided by mean) is roughly the same over *r* as over *s*. Then by Table 1, $\mathrm{var}_s(g) \approx Q_r(1+Q_r)^2$, and $\mathrm{var}_s(f) = Q_s$. Both Q_r and Q_s are small compared to 1 and not greatly different, so $(1+Q_r)/(1+Q_s) = 1+\delta$ for some small δ. Then

$$\mathrm{corr}_s(f,g) = \frac{-Q_r}{(Q_sQ_r)^{1/2}(1+Q_r)} = -\frac{1}{1+Q_s}h(\delta), \qquad (34)$$

where $h(\delta) = (1+(1+Q_s^{-1})\delta)^{1/2}/(1+\delta)$. Now, for small δ, $h(\delta) \approx 1$. The formula in (13) follows.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 201–218, DOI 10.21307/stattrans-2018-012

UNIVARIATE SAMPLE SIZE DETERMINATION BY ALTERNATIVE COMPONENTS: ISSUES ON DESIGN EFFICIENCY FOR COMPLEX SAMPLES

Ceylan Talu Yozgatligil[1], **H. Öztaş Ayhan**[2]

ABSTRACT

Sample size determination for any sample survey can be based on the desired objectives of the survey as well as the level of confidence of the desired estimates for some survey variables, the desired precision of the survey results and the size of the population. In addition to these, the cost of enumeration can also be considered as an important criterion for sample size determination. Recently, some international organizations have been using univariate sample size determination approaches for their multivariate sample designs. These approaches also included some design efficiency and error statistics for the determination of the univariate sample sizes. These should be used for determining the survey quality measures after the data collection, not before. The additional components of the classical sample size measure will create selection and representation bias of survey estimates, which is discussed in this article.

Key words: univariate sample size, representation bias, sample allocation, error statistics, design efficiency measures.

1. Introduction

Sample size determination for univariate cases has been commonly used for many years. Surveys which are based on large population sizes require other sample size determination methodologies than the univariate cases, because they are based on criteria for multivariate observations. Therefore, univariate sample size determination methodologies cannot satisfy the multivariable criteria. Recently, some national and international survey organizations have been using some modified univariate sample size determination formulas which have low efficiency. As a result, this can lead to under- or overrepresentation of the population by the selected sample. These modified formulas contain unnecessary components, such as *design effect* and *response* or *nonresponse rates*, etc. This article highlights the components which create selection and representation bias of survey estimates. Therefore, the methodological problem is not the concern of

[1] Department of Statistics, Middle East Technical University, 06800 Ankara, Turkey. E-mail: ceylan@metu.edu.tr
[2] Department of Statistics, Middle East Technical University, 06800 Ankara, Turkey. E-mail: oayhan@metu.edu.tr

this article. The main aim is to emphasize the correct usage of the sample size determination formulas for the multivariable case. Some researchers may want to follow the methodology (or formula) used by the respectful organizations, which may not fully (or correctly) represent the target population.

2. Classical Univariate Sample Size Measures

Sample size determination is an important aspect of the representativeness of the survey results. There are many approaches which can be taken. Generally, all surveys utilize too many variables. Some of these variables may be more important than others for decision-makers. The researcher generally wishes to satisfy the representation of several survey variables, which are important.

There are many studies on determination of sample size in different disciplines. Some of them propose a new methodology and some others gather the existing ones and compare their performances. Dell et al. (2002) discussed simple methods of estimating the number of animals needed for various types of variables and experiments. They showed that it is crucial to choose the power, the significance level, and the size of the effect to be detected, and to estimate the population variability of the variable being studied, and using a complicated design and statistical analysis usually results in the highest power to detect any difference. Shore (2008) addressed sample size determination relating to hypothesis testing, parameter estimation, relational modelling and optimal sampling. Sathien et al. (2010) gave a few suggestions regarding the methods to determine an optimum sample size in descriptive and analytical studies. Marshall et al. (2013) described basic requirements for sample size determination and the sample size determination methods to estimate a normal distribution mean, standard deviation, quantile, binomial proportion and Poisson occurrence rate. In his book, Ryan (2013) discussed many sample size calculation techniques with applications using software. Siddiqui (2013) presented the guidelines described in the literature as to determine the appropriate sample size for the various statistical techniques. Safo et al. (2015) compared the performance of the existing sample size method and the sample size method developed by the authors for lasso logistic regression. Placzek and Friede (2017) proposed methods for planning and analysing a multiple nested subgroups design and described sample size determination prior to the trial and sample size recalculation via a blinded review in an internal pilot study.

For the representation of survey results, a very important single survey variable (*univariate case*) can be chosen and the sample size only for this variable can be evaluated. Alternatively, two variables (*bivariate case*) may affect one another and the sample size determination may be based on the presence of these. Finally, several variables (*multivariate case*) may become very important to determine the minimum sample size, by utilizing multivariate information. One of the common and most practical solution to these problems is to select several independent variables (*univariate case*) and compute sample sizes for each of these separately and choose the largest computed sample size to satisfy all variables.

The use of several survey variables one at a time has some practical conveniences. On the other hand, the type of measurement scale of the survey

variable also leads to the use of different test statistics as input information for the sample size determination model. Here, the case of the proportion and another case of the sample mean for the determination of sample size will be illustrated.

2.1. Determination of Sample Size for Proportion

For the test of the following hypothesis; $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, the sample proportion (*p*) of success is distributed asymptotically as $N\{\theta, \theta(1-\theta)/n\}$ with the requirement that $\Pr\left[\,|p-\theta| \le d \mid \theta\right] \ge 1-\alpha.$ This leads to the sample size estimation as: $n \ge \theta\,(1-\theta)\,\chi^2_{(1),\alpha}\,/\,d^2$, where α is the level of significance and *d* is the level of tolerance of the estimate. For the unknown population proportion, we take θ = 0.5 and the sample size estimate will be: $n \ge 0.25\,\chi^2_{(1),\alpha}\,/\,d^2$. This naturally represents the worst case, which creates the maximum variance, to be on the safe side as the sample designer. If we have prior information about the population proportion, then consequently we can have relatively smaller sample size estimation.

Hence, the sample size can be determined as: $n \ge 0.25\,\chi^2_{(1),\alpha}\,/\,d^2$. The overall sampling fraction for this design will be, *f* = *n*/*N* = 1/F. Here, (*N*) is the total number of Housing Units (HUs) in the population, and (*n*) is the total number of Housing Units (HUs) in the selected sample. For self-weighting sample designs, the sampling fraction for any domain will be the same as any other domain in the design. Furthermore, this will also be equal to one another within any prefecture as well as the total population.

2.2. Determination of Sample Size for Frequency Type Variable

The frequentist case of sample size determination is concerned with the normal distribution with known variance. When the random variable *X* is distributed as $N\left(\mu, \sigma^2\right)$, the mean μ may be estimated with absolute error (*d)* and probability 1 – *α* by the sample mean (*m*) if

$$\Pr\left[\,|m-\mu| \le d \mid \mu\right] \ge 1-\alpha.$$

Since $\sqrt{n}(m-\mu)/\sigma \sim N(0,1)$, it follows that the above inequality is satisfied when the sample size (*n*) satisfies $n \ge \sigma^2\, Z^2_{\alpha/2}\,/\,d^2$. Here, tolerance level refers to $d = Z_{\alpha/2}\left(\hat{\sigma}/\sqrt{n}\right)$. We can also easily create an application for this case just like the previous one. If we take the same element variance value and the same tolerance level for this case, then the estimated sample size will be the same as before. Hence, the sample size is determined by

$$n \ge \sigma^2\, Z^2_{\alpha/2}\,/\,d^2$$

formula, which is affected by the level of confidence, level of tolerance (desired error variance) and the element variance. Due to changes in these parameters will consequently result in differing outcomes.

3. Sample Design for the Survey

The sample design for the survey will be based on the latest available information on the population. The population will be stratified into domains (prefectures) and self–weighting samples will be selected for each domain.

3.1. Sampling Frame and Stratification

The latest population figures are based on the population projections for the survey time. The aggregated data from the urban–rural information for the available districts will be aggregated into several prefectures within a nested structure in defined geographic areas. Dividing the *total urban and rural population* $\left(M_h\right)$ for each domain (prefecture) by their Population Census *average household size* $\left(\overline{H}_h\right)$ of each prefecture, we can compute the *number of urban and rural Housing Units* $\left(N_h = M_h / \overline{H}_h\right)$ for the survey date. This calculation is based on the assumption that: the *average household size does not change significantly over the years*. This assumption is verified and used in many countries of the world.

In summary, Desu and Raghavarao (1990) and Adcock (1997) proposed the following measures for frequentist methods.

$n_0 \geq \sigma^2 Z_{\alpha/2}^2 / d^2$ where *d* is the *absolute error*; $d = Z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}$.

Alternatively, for studies aiming at the hypothesis testing

$$n^* = \sigma^2 \left(Z_{\alpha/2} + Z_\beta\right)^2 / d^2 .$$

For the studies with binary response, i.e. binomial distribution,

$$n_0 \geq \theta(1-\theta)\, \chi^2_{(1),\alpha} / d^2 .$$

The ultimate sample size is adjusted for the known population size as:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} .$$

3.2. Measures of Design Efficiency

The following measures of the design efficiency are commonly used for many surveys, after the data collection. There are several measures of design efficiency in survey research. Basically, it is the ratio of sampling variances, which is based on two different sample designs. The comparison of the two variances has to be based on the same sample sizes for both designs.

i. Design efficiency

Design efficiency is the ratio of two sampling variances for given sample designs (*Di; i=1,2*).

$$DesEff = \text{var}(\bar{y}_{D1}) / \text{var}(\bar{y}_{D2})$$

where *D2* is not based on Simple Random Sample (*SRS*) design.

ii. Design effect

A design effect (*deff*) measures the relative increase or decrease in the variance of an estimator due to departures from simple random sampling. Kish (1965) presented *deff* as a convenient way of gauging the effect of clustering on an estimator of a mean (Henry and Valliant, 2015). Later work by Rao and Scott (1984) and others found that more complicated versions of *deff*'s were useful to adjust inferential statistics calculated from complex survey data (Sirken, 2002).

A specialized version of *deff* was proposed in Kish (1965), who addressed only the effect of using weights that are not all equal. Kish derived the "design effect due to weighting" for a case in which weights vary for reasons other than statistical efficiency (Henry and Valliant, 2015). There are also sample designs and estimators where having varying weights can be quite efficient.

Design effect is the ratio of two sampling variances for given sample designs.

$$deff = \text{var}\,(\bar{y}_{D1}) / \text{var}\,(\bar{y}_{SRS})$$

where *Design 2* is based on *SRS* only (Kish, 1965 & 1982). The original definition of the design effect is based on the sampling variance of a given complex design, which is compared with the SRS sampling variance of the same sample size. Theoretically, SRS has to be taken as an independent sample from the same population rather than adjusting the complex sample design boundaries as if it was selected as a SRS.

The efficient sample size calculations assume simple random sampling. If the sample design deviates from SRS, the efficient sample size will also vary. *deff* is a measure for the relative efficiency of an estimator under a studied sampling design. It is the direct way of measuring the effect of design on sampling variability. The planned sample size computation for the univariate case naturally corresponds to the "gross sample size". After the data collection "net sample size" will be achieved. The difference can be reflected through the computation of the nonresponse amount. On the other hand, the *deff* computation will be based on the sampling variance of the existing data, which is collected from the net sample size. Naturally, this will not include the planned inclusion probabilities and the clustering for the complex sample design in particular.

iii. Design factor

Design factor is the ratio of two standard errors for given sample designs (Kish, 1965).

$$deft = se\,(\bar{y}_{D1}) / se\,(\bar{y}_{SRS})$$

where *Design 2* is based on *SRS* only. Here, $deft = \sqrt{deff}$ and $deft^2 = deff$.

deft is a measure of efficiency of a given sample design compared to a direct simple random sampling of individuals, defined as the ratio between the standard error using the given sample design and the standard error that would result if a simple random sample was used. A *deft* value of 1.0 indicates that the sample design is as efficient as the simple random sample.

3.3. Computed Error Statistics for the Analysis of Design Efficiency

The sample design efficiency for a given design will be compared with some error statistics, in order to show the data quality measures. These measures are based on the error statistics which are based on the complex multivariate designs when compared with the base design, which is SRS with replacement. The basic error statistics which are obtained for this comparison will be: *standard error, design factor, design effect, rate of homogeneity, cluster size*, etc. Some examples of these statistics are given in Table 1 below, which is based on the "2013 Turkey: Population and Health Survey" (HÜNEE, 2014). It is based on the complex sample survey design, which has 14,496 target sample households. The total sample household population was 41,476 persons. The household population consists of 78% urban and 22% rural domains. The aim of the presentation of these figures is merely to highlight the importance and usage of these error statistics. Here, the interpretation of the survey results is not intended to be the main purpose of this study.

Table 1. Sampling Related Error Statistics for Selected Survey Variables Turkey 2013

Survey Variables	Ratio mean $r = y/x$	Standard error *se(r)*	Design factor *deft*	Relative error *se(r)/r*
Never married women	0.275	0.006	1.277	0.021
Currently married women	0.683	0.006	1.276	0.009
Number of live births	1.667	0.020	1.133	0.012
Number of living children	2.919	0.050	1.252	0.017
Wants no children	0.474	0.007	1.202	0.015
Ideal number of children	2.721	0.019	1.507	0.007
Total fertility rate (*3 years*)	2.258	0.069	1.360	0.031
Infant mortality rate (*5 years*)	13.282	2.345	1.111	0.177

Source: HÜNEE (2014).

The purpose of computing these statistics is to compare the efficiency of the latest design used. On the other hand, some survey institutions are mistakenly proposing to include these error statistics into their selection procedures. They are utilizing a univariate sample size formula, which is combined with some of these error statistics as well as response or nonresponse rate components, in order to pre-adjust the sample size. This article has shown that the use of additional unrelated components will create the selection and representation bias for the estimation of selected population parameters.

4. Some Modified Sample Size Estimators

For large scale surveys, an ideal way of obtaining the required sample size should be based on multivariate sample size determination. Some international organizations are insisting on using univariate sample size determination methods with some modifications to the formulae in place. For this case, their argument is based on using the univariate sample size methodology for several design variables separately. Then, they intend to compensate for the missing components by adding some error statistics (*deff, nonresponse*, etc.) in advance which are based on complex sample designs. They also argue that adding these statistics to their modified sample size formulae will solve their methodological bias.

These error statistics are theoretically used for measuring the design efficiencies of their complex sample designs, when compared with the unrestricted design (i.e. SRS-WR). However, they are not proposed to be used prior to sample selection as an additional design component. Another important point is when these additional components are used within the desired sample selection formulae, they will naturally effect the overall sample selection probabilities in an undesired way, which will create sample selection bias. Consequently, it is not advised to use the modified univariate sample size determination formulas of this type. We would like to justify our argument by giving two different modified formulas in the following subsections.

Survey sampling statisticians are responsible for designing sample surveys and determining the ideal and unbiased sample results for their surveys. When they are comparing their survey results with several internationally organized surveys, where their sample selection was biased due to the use of undesired sample size formulation, which created biased results. Consequently, these methodological problems naturally concern survey sampling statisticians overall. In addition, naturally these issues have to be brought to the attention of survey methodology community.

4.1. Demographic and Health Surveys (DHS)

The DHS (2012) has used the following formula for calculating the final sample size in terms of the number of households while taking design effect and non-response into account in advance, and is given by:

$$n_{DHS} = deft^2 \frac{(1/P - 1)}{\alpha^2} \Big/ (R_i \; R_h \; d).$$

The formula in terms of our notation is given by

$$n_{DHS} = deft^2 \frac{(1/p - 1)}{d^2} \Big/ (R_i \; R_h \; e)$$

where

n is the sample size in households;

*deft** is the design effect (a default value of 1.5 is used for *deft* if not specified);

p is the estimated proportion;

d is the desired relative standard error, $\sqrt{p(1-p)/n}\,/\,p$;

R_i is the individual response rate;

R_h is the *household gross response rate*; and

e is the number of eligible individuals per household.

(*): The symbol *deft* actually denotes the design factor, not the design effect. (Default value of *deft=1.5* is recommended in DHS manual as a special case. Naturally, this corresponds to *deff=2.25*). In practice, this can be an acceptable threshold value for complex clustered sample designs.

The *household gross response rate* is the number of households interviewed over the number selected. DHS reports the net household response rate, which is the number of households interviewed over the number valid households found in the field (i.e. excluding vacant and destroyed dwellings). The practical aspects of R_h and R_i rates are discussed in Ayhan (1981) for the Turkish Fertility survey data. Ayhan (1981) has used the WFS (1975) recommendations that the first visit to the household (or individual) plus the number of re-calls constitute total calls. For a household survey, 1 + 3 = 4 total calls, and for individual survey, 1 + 2 = 3 total calls are proposed as threshold values.

For a required precision with a relative standard error α, the net sample size (number of completed interviews) needed for a simple random sampling (SRS) is given by:

$$n_{SRS} = \frac{(1/p - 1)}{\alpha^2}.$$

Since a simple random sampling is not feasible for DHS, the sample size for a complex survey with clustering such as DHS can be calculated by inflating the above calculated sample size by using a design effect (*deff=deft²)*.

A simple random sample would be a random selection of individuals or households directly from the target population. This is not feasible for DHS surveys because a list of all eligible individuals or households is not available.

4.2. Multiple Indicator Cluster Surveys (MICS)

Another survey which is based on the complex sample design is the MICS (2006). Methodological manuals of the United Nations Children's Fund (UNICEF), Statistics and Monitoring Division, propose using the modified univariate sample selection formulae for their multivariable surveys.

The sample size calculating formula for MICS is given by

$$n_{MICS} = \frac{4r(1-r)deff(1.1)}{(0.12r)^2\, p.n_h}.$$

The formula in terms of our notation is given by

$$n_{MICS} = \frac{\chi^2_{(1),\alpha}\, p(1-p)deff(r)}{(0.12p)^2\, k.n_h}$$

where

n_{MICS} is the required sample size, expressed as number of households

$\chi^2_{(1),\alpha}$ is a factor to achieve the $(1-\alpha)$ per cent level of confidence

p is the predicted or anticipated prevalence (coverage rate) for the indicator being estimated

r is the factor necessary to raise the sample size for nonresponse. (*for example, for 10% nonresponse rate r should be 1.1*.)

deff is the design factor

0.12p is the margin of error to be tolerated at the 95 per cent level of confidence, defined as 12 per cent of p (*12 per cent thus represents the relative sampling error of p*)

k is the proportion of the total population upon which the indicator p is based, and

n_h is the average household size.

For the Multiple Indicator Cluster Surveys (MICS), *UNICEF* proposes r = 1.1 as an early compensation for the nonresponse amount. This will correspond to 10% increase in the sample size before the data collection, which intends to compensate the same amount of loss in the collected sample following the data collection. This approach cannot be accepted due to several bias producing aspects. Firstly, nonresponse rate is a part of survey error, which should not be included as the sample selection component. Secondly, 10% nonresponse rate can be a lower bound threshold value for this error statistics. For many surveys, the nonresponse rates are higher than this in the literature. Recently, there has been even a tendency of increase in nonresponse rates for the sample surveys of some developed countries.

For the MICS methodology, relative sampling error (value of *0.12p*) has been used for margin of error in the previous formulae because it scales the margin of error to result in comparable accuracy regardless of whether a high coverage indicator or low coverage indicator is chosen as the key one for sample size determination.

Recently, UNICEF, Statistics and Monitoring Section has decided not to clarify the sample selection formulae by removing the related methodology from their website. Instead, they provided a “sample size determination” template electronically. This template is naturally based on the previously discussed methodology for a univariate sample size determination algorithm, for a multivariate complex sample design.

5. Design Efficiency of Alternative Sample Sizes

This section clearly shows the partition of the components which are based on the modified sample size formulas of the two international institutions. The bias which will be created for the estimation by using the unrelated sample size formulae is given for the examined two large scale surveys.

The formula proposed by DHS is:

$$n_{DHS} = \frac{deff \dfrac{\left(1/p - 1\right)}{d^2}}{R_i R_h e}$$

where n_{DHS} is the sample size in HH's offered by DHS.

The formula used in this study is

$$n_{DHS} = \frac{\left[(1-p)/p\right]}{d^2} \frac{deff}{R_i . R_h . e} .$$

The formula proposed by MICS is

$$n_{MICS} = \frac{\chi^2_{(1),\alpha}\, p(1-p) deff(r)}{(0.12p)^2 k.n_h}$$

where n_{MICS} is the sample size in HH's offered by MICS.

The formula used in this study is

$$n_{MICS} = \frac{\chi^2_{(1),\alpha}\left[p(1-p)\right]}{d^2} \frac{deff(r)}{k.n_h} .$$

The relationship between the classical sample size formulization and DHS's sample size formulization can be given as

$$n_{DHS} = n_C \frac{1}{\chi^2_{(1),\alpha} p^2} \frac{deff}{R_i . R_h . e} ,$$

where n_C is the classical sample size formula for the binary response.

The relationship between the classical sample size formulization and MICS's sample size formulization can be given as

$$n_{MICS} = n_C \frac{deff.(r)}{k.n_h} .$$

6. Issues on Selection Bias Representation

A comparison of the outcomes for the classical sample size determination methods and modified sample size determination methods provides information on the population representation and related biases. If we compare the results, in terms of overall sampling fractions, the following comparison can be used.

Overall sampling fraction of the *classical estimate*:

$$f_{SRS} = \frac{n_{SRS}}{N} = \frac{1}{F} = \frac{\chi^2_{(1),\alpha}\left[p(1-p)\right]}{d^2} \Big/ N .$$

Overall sampling fraction of the *modified estimate of DHS*:

$$f_{DHS} = \frac{n_{DHS}}{N} = n_C \frac{1}{\chi^2_{(1),\alpha} p^2} \frac{deff}{R_i.R_h.e} \Big/ N = f_{SRS} \frac{1}{\chi^2_{(1),\alpha} p^2} \frac{deff}{R_i.R_h.e}.$$

Overall sampling fraction of the *modified estimate of MICS*:

$$f_{MICS} = \frac{n_{MICS}}{N} = n_C \frac{deff.(r)}{k.n_h} \Big/ N = f_{SRS} \frac{deff.(r)}{k.n_h}.$$

Selection bias of the estimates for DHS:

$$B(f_{DHS}) = f_{DHS} - f_{SRS} = f_{SRS} \left(\frac{1}{\chi^2_{(1),\alpha} p^2} \frac{deff}{R_i.R_h.e} - 1 \right),$$

where bias will be *0* if and only if $\frac{1}{\chi^2_{(1),\alpha} p^2} \frac{deff}{R_i.R_h.e} = 1.$

Selection bias of the estimates for MICS:

$$B(f_{MICS}) = f_{MICS} - f_{SRS} = f_{SRS} \left(\frac{deff.(r)}{k.n_h} - 1 \right),$$

where bias will be *0* if and only if $\frac{deff.(r)}{k.n_h} = 1.$

Selection bias formulas show that the sample size calculation used by the surveys affect the overall sample selection probabilities. Accordingly, it is not recommended to use the modified univariate sample size determination formulas of this type for the complex sample designs.

Sample size determination for a two stage cluster sampling is proposed by Desu and Raghavarao (1990), and Aliaga and Ren (2006). Hansen *et al.* (1953) has evaluated the general cost function model for the complex sample designs. For the multivariable designs, there is no established standard computation formula for sample sizes. Depending on the type of design, the related parameters constitute variables for complex sample designs.

7. Weighting Adjustment Procedures

After the determination of the univariate sample size, the actual SRS sample is selected from the population by using a randomization process. For the purpose of the allocation of complex sample survey designs, the selected sample is then reallocated to the proposed new sample design. The proposed design can be allocated to complex designs, which may be based on *equal allocation*, *proportional allocation, probability proportional to size (PPS) allocation, weighted PPS allocation, optimum allocation,* and *clustering*. A comparison of the sample allocation methods is summarized by Ayhan and Islam (2005). Under this

approach, the following adjustment methodologies can be used after the data have been collected from a complex sampling plan.

7.1. Weighting Independent Stages

Data weighting methods have been covered by Kish (1992), Kalton and Flores-Cervantes (2003), and Ayhan (2003), and Alkaya et al. (2017) in detail. Several alternative approaches, such as cell weighting, ranking, linear weighting, GREG weighting and several others can be proposed (Vaillant et al., 2013; Brick, 2013).

7.1.1. Adjustments for Design Weights

For complex or stratified sample designs, design weights have to be used for the adjustment of the sample selection probabilities if the sample design is not self-weighting. For self-weighting sample designs selection probabilities of each domain will be the same as the overall, that is $f = \frac{n}{N} = \frac{n_i}{N_i} = f_i \quad \forall i$. Then, the design weights are given by

$$w_i = \frac{1}{p_i}\left[\frac{n}{\sum\left(1/p_i\right)}\right], \; i = 1,2,\cdots,n$$

where p_i is the selection probability of the domain and $\sum w_i = n$.

7.1.2. Adjustments for Non-Response Weight

Non-response weights should be used as an error correction component after the data collection, not before. Non-response adjustment weights are used to compensate for the losses of non-response amounts when the overall non-response rate is greater than 10 per cent and the domain non-response rates are more than 5 per cent for any domain (WFS, 1977). Sample design outcomes other than the above restrictions do not require any weighting adjustments for the sample outcomes. Hence, the non-response weights are for each domain given by

$$w_j = \overline{R}/R_j \; , \; j = 1,2,\cdots,$$

here $\overline{R} = \frac{\sum n_j}{\sum\left(n_j/R_j\right)}$, where R_j is the non-response rate for the domain (or strata), $\overline{R}$ is the average non-response rate overall domains, and n_j is the domain size. These rates (R_j and $\overline{R}$) are recommended by WFS (1977) and has been used in 42 WFS country surveys, including Turkish Fertility Survey 1980 (Ayhan, 1981).

7.1.3. Adjustments for Post-Stratification Weights

Computation of the post-stratification weights is required for each domain in order to avoid bias due to cross-tabulation of the data. Kalton and Flores-Cervantes (2003) have proposed an alternative combined adjustment methodology for sample surveys.

This procedure adjusts the sample weights so that the sample totals conform to the population totals on a cell-by-cell basis. The weights for each respondent (typically, the inverse of the probability of the case) in a weighting cell (or post-stratum) is multiplied by an adjustment factor (Tourangeau et al., 2013). Then, the weight formula is given as

$$w_{2ij} = \frac{N_j}{\sum_{n_j} w_{1ij}} w_{1ij},$$

in which w_{2ij} is the adjusted or post-stratified weight, w_{1ij} is the unadjusted weight, and the adjustment factor is the ratio between the population total for cell *j* (N_j) and the sum of the unadjusted weights for the respondents in that cell.

Rather than using independent weighting and adjustment procedures for each stage of the weighting, alternative approaches can also be used. This is based on combined weighting methods, which take into account the conditional probability approach for the previous stages. As an alternative to the weighting independent stages, the combined weighting methods can be proposed to avoid bias for the sample estimates.

7.2. Combined Weighting Methods

Ayhan (2003) and Alkaya et al. (2017) have proposed the following combined weighting procedure for sample surveys. These weighting procedures are used in a sequential manner for each weighting component. The weights are proposed as products for each weighting stage in a combined way. Sample design may be self-weighting or non-self-weighting. Design weights have to be introduced for non-self-weighting designs in the following way.

The probability of selection of the overall sample is obtained simply by the sampling fraction of the selected sample $f = x/X = 1/F$ for the total sample. On the other hand, after using some method of stratification, the sampling fraction of any strata is $f_i = x_i / X_i = 1/F_i$.

7.2.1. Design Weights

Design weights (Ayhan, 1991; Verma, 1991) for non-self-weighting sample designs can be computed for each domain *i* with the same probability of selection p_i (Ayhan, 2003).

For *combined ratio mean* $\theta = Y/X = \sum_{i}^{H} Y_i \Big/ \sum_{i}^{H} X_i$ where *i = 1 ,2*,.., *H*, here *H* represents the number of domains, estimated by $\hat{\theta} = y/x = \sum_{i}^{H} y_i \Big/ \sum_{i}^{H} x_i$. On the other hand, for a *separate ratio mean* $\theta_w = \sum_{i}^{H} W_i\ \theta_i$, estimated by

$$\hat{\theta}_w = \sum_{i}^{H} W_i\ \hat{\theta}_i = \sum_{i}^{H} W_i\ \left[y_i / x_i \right] ,$$

$$W_i = \left[\sum_{i=1}^{H} x_i \Big/ \sum_{i=1}^{H} \left\{ x_i / \left[(X/x)\, p_i \right] \right\} \right] \Big/ \left[(X/x)\, p_i \right] = P_0 / P_i$$

where $\sum_{i=1}^{H} \left(W_i\ x_i \right) = x$.

Here, P_0 has been computed to adjust the overall weighted and unweighted sample to be the same.

7.2.2. Combined Weighting for Nonresponse

In addition, a weighting procedure for nonresponse is also essential for self-weighting and non-self-weighting sample design outcomes. The non-response rate is calculated as

$$W_i^* = R_0 / R_i$$

where $R_i = x_i^* / x_i$ is the response rate in domain *i*.

The overall response rate (R_0) for the design can be computed as

$$R_0 = \sum_{i=1}^{I} \left(W_i\ x_i \right) \Big/ \sum_{i=1}^{I} \left(W_i\ x_i\ /\ R_i \right),$$

where R_0 is used to adjust the sample sizes to be the same, $\sum_{i=1}^{I} \left(W_i\ \ W_i^*\ x_i \right) = x$.

7.2.3. Combined Weighting for Post-Stratification

Finally, post-stratification of a complex sampling scheme requires additional weighting procedures for independent subclasses. The combined weight can be calculated by using the following weights:

$$W_i' = \left(W_i\ \ W_i^*\ X_i \right) / X ,$$

$$W_i = \left[\sum_{i=1}^{H} x_i \Big/ \sum_{i=1}^{H} \left\{ x_i / \left[(X/x)\, p_i \right] \right\} \right] \Big/ \left[(X/x)\, p_i \right] \text{ and}$$

$$W_i^* = R_0 / R_i \ ,$$

where W_i' is the post-stratification weight, W_i is the design weight, W_i^* is the non-response weight, $R_0 = \sum_{i=1}^{I} \left(W_i \ x_i \right) \Big/ \sum_{i=1}^{I} \left(W_i \ x_i / R_i \right)$ and $R_i = x_i^* / x_i$.

Consequently, $\sum_{i=1}^{I} \left(W_i \ W_i^* \ W_i' \ x_i \right) = x$ is the overall sample adjustment procedure for the combined weighted estimator. This is naturally provides the adjustment to the base variable *x* (Ayhan, 2003; Alkaya et al., 2017).

Alternative weighting adjustment procedures in multistage complex sample surveys are proposed by Ayhan et al. (2000) for adjusting the original selection probabilities by PPS procedures.

The next step in the analysis of the collected data is to compute the following error statistics for the proposed sampling design. This will provide information on how efficient the designed sample was when compared with the basic standard, which is SRS.

8. Discussion and Conclusions

In a multivariable survey design, the determination of the sample size is an important concept that has to be answered. Although there is no settled methodology, some prestigious organizations modify the formula for univariate sample size determination to be able to use it in the multivariable case. For this purpose, they included some factors such as *deff* or non-response rate in their sample size formulas. These factors have to be calculated after the sample has been collected as a data quality measure, not before. Hence, the modified univariate sample size methodologies of several survey institutions do not represent the corresponding population. The amount of bias involved in the formulas is clearly identified during the previous formulations. This paper highlights the importance of sample selection in a representative manner, to avoid the selection bias.

The ideal approach should be not to determine the sample size of the complex survey design as if it was based on the univariate case and use SRS assumptions. Consequently, representation bias enables the survey results not representing the corresponding population parameters.

For the complex designs, the suggested alternative strategy is to use weighting after SRS. For the purpose of allocation, the selected sample is reallocated to the proposed new sample design. As an alternative to the weighting independent stages, the combined weighting methods can be proposed to avoid bias for the sample estimates.

International survey organizations should be responsible for following recent developments in survey sampling theory and methods, in order to maintain themselves as reliable institutions.

REFERENCES

ADCOCK, C. J., (1997). Sample Size Determination: A Review. The Statistician, 46 (2), pp. 261–283.

ALIAGA, A., REN, R., (2006). Optimal Sample Sizes for Two-Stage Cluster Sampling in Demographic and Health Surveys. DHS Working Papers, No. 30, Demographic and Health Research, ORC Macro.

ALKAYA, A., AYHAN, H. Ö., ESIN, A., (2017). Sequential Data Weighting Procedures for Combined Ratio Estimators in Complex Sample Surveys, Statistics in Transition, 18 (2), pp. 139–162.

AYHAN, H. Ö., (1981). Sources of Nonresponse and Nonresponse Bias in 1978 Turkish Fertility Survey. Turkish Journal of Population Studies, 2-3, pp. 104–148.

AYHAN, H. Ö., (1991). Post-stratification and weighting in sample surveys. Research Symposium '91, State Institute of Statistics, Ankara. pp. 1–11.

AYHAN, H. Ö., HANCIOGLU, A., TÜRKYILMAZ A. S., ÜNALAN, T., (2000). Sample Design, Implementation and Outcome. Chapter 13 in Push and Pull Factors of International Migration: Country Report – Turkey (H. Ö. Ayhan, et. al., eds). EUROSTAT Working Papers; Population and Social Conditions 3/2000/E/n^0 8, Luxembourg: EUROSTAT Press, pp. 112–131.

AYHAN, H. Ö., (2003). Combined Weighting Procedures for Post-Survey Adjustment in Complex Sample Surveys. Bulletin of the International Statistical Institute 60, pp. 53–54.

AYHAN, H. Ö., ISLAM, M. Q., (2005). Sample Design and Allocation for Random Digit Dialling, Quality and Quantity, 39 (5), pp. 625 – 641.

BRICK, J. M., (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review, Journal of Official Statistics, 29 (3), pp. 329–353.

DELL R. B., HALLERAN, S., RAMAKRISNAN, R., (2002). Sample Size Determination, ILAR J, 43 (4), pp. 207–213.

DESU, M. M., RAGHAVARAO, D., (1990). Sample Size Methodology, Boston: Academic Press.

DHS, (2012). Demographic and Health Surveys Methodology, Sampling and Household Listing Manual. ICF International, Calverton, Maryland USA, p. 109.

HANSEN, M. H., HURWITZ, W. N., MADOW, W. G., (1953). Sample Survey Methods and Theory, New York: Wiley.

HÜNEE, (2014). 2013 Turkey Population and Health Survey, Hacettepe University, Institute of Population Studies, Ankara, Turkey.

HENRY, K. A., VALLIANT, R., (2015). A Design Effect Measure for Calibration Weighting in Single-Stage Samples, Survey Methodology, 41 (2), pp. 315–331.

MARSHALL, B., CARDON, P., PODDAR A., FONTENOT, R., (2013). Does Sample Size Matter in Qualitative Research? A Review of Qualitative Interviews in is Research, Journal of Computer Information Systems, Vol. 54 (1), pp. 11–22.

MICS, (2006). Multiple Indicator Cluster Survey Manual 2005, Division of Policy and Planning, United Nations Children's Fund (UNICEF), New York, USA, p. 595.

KALTON, G., FLORES–CERVANTES, I., (2003). Weighting Methods, Journal of Official Statistics, 19 (2), pp. 81–97.

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

KISH, L., (1982). Design Effect. Encyclopedia of the Statistical Sciences, John Wiley and Sons, Inc. Vol. 2, pp. 347–348.

KISH, L., (1992). Weighting for Unequal P_i's, Journal of Official Statistics, 8 (2), pp. 183–200.

PLACZEK, M., FRIEDE, T., (2017). Clinical trials with nested subgroups: Analysis, sample size determination and internal pilot studies, Statistical Methods in Medical Research 0 (0), pp. 1–18, DOI: 10.1177/0962280217696116.

RAO, J. N. K., SCOTT, A. J., (1984), On Chi-square Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data, Annals of Statistics 12, pp. 46–60.

RYAN, T. P., (2013). Sample Size Determination and Power, John Wiley and Sons Inc., Hoboken, New Jersey.

SAFO, S., SONG, X., DOBBIN, K. K., (2015). Sample Size Determination for Training Cancer Classifiers from Microarray and RNA-Seq Data, Ann. Appl. Stat. 9 (2), 1053—1075, DOI:10.1214/15-AOAS825, http://projecteuclid.org/euclid.aoas/1437397123.

SATHIAN, B., SREEDHARAN, J., BABOO, S., SHARAN, K., ABHILASH, E., RAJESH, E. (2010). Relevance of Sample Size Determination in Medical Research, Nepal Journal of Epidemiology, 1 (1), pp. 4–10, DOI: http://dx.doi.org/10.3126/nje.v1i1.4100.

SHORE, H., (2008). Sample-Size Determination. Encyclopedia of Statistics in Quality and Reliability, IV, John Wiley & Sons, Ltd.

SIDDIQUI, K. A., (2013). Heuristics for Sample Size Determination in Multivariate Statistical Techniques (December 6, 2013), World Applied Sciences Journal 27 (2), pp. 285–287, Available at SSRN: https://ssrn.com/abstract=2447286.

SIRKEN, M. G., (2002). Design Effects of Sampling Frames in Establishment Survey, Survey Methodology, 28 (2), pp. 183–190.

TOURANGEAU, R., CONRAD, F. G., COUPER, M. P., (2013). The Science of Web Surveys, Oxford University Press, New York.

VALLIANT, R., DEVER, J. A., KREUTER, F., (2013). Practical Tools for Designing and Weighting Survey Samples, Statistics for Social and Behavioral Sciences. Springer.

VERMA, V., (1991). Sampling Methods, Training Handbook, Statistical Institute for Asia and the Pacific, Tokyo.

WFS, (1975). WFS Basic Documentation, No. 5, Supervisor's Instruction, World Fertility Survey, London, U.K.

WFS, (1977). Guidelines for Country Report, No.1, WFS Basic Document, No. 8, World Fertility Survey, London, U.K., 186.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 219–238, DOI 10.21307/stattrans-2018-013

EFFICIENT ESTIMATORS OF POPULATION MEAN USING AUXILIARY INFORMATION UNDER SIMPLE RANDOM SAMPLING

Mir Subzar[1], Showkat Maqbool[1], Tariq Ahmad Raja[1], Surya Kant Pal[2], Prayas Sharma[3]

ABSTRACT

In the present study we have proposed an improved family of estimators for estimation of population mean using the auxiliary information of median, quartile deviation, Gini's mean difference, Downton's Method, Probability Weighted Moments and their linear combinations with correlation coefficient and coefficient of variation. The performance of the proposed family of estimators is analysed by mean square error and bias and compared with the existing estimators in the literature. By this comparison we conclude that our proposed family of estimators is more efficient than the existing estimators. To support the theoretical results, we also provide the empirical study.

Key words: Auxiliary information, Bias, Mean Square Error, Simple Random Sampling, Efficiency.

AMS Subject Classification: 62D05

1. Introduction

Since last few decades statisticians have proposed estimators for the population parameters. The concept of efficiency is a vital key word for an estimator and it depends on the Mean Square Error (MSE), thus we should know the fundamental of MSE in statistics. Therefore, according to the estimation theory the estimators with the mean square error or variance than lower that of other estimators are said to be efficient estimators and the estimators with the bias lower than that of other estimators are said to be good estimators. Hence, this efficiency in estimation theory is achieved by using the auxiliary information at the design or estimation stage or at both stages. The use of such auxiliary information is made through different methods of estimation such Ratio, Regression and Product methods. So, Cochran (1977) initiated the use of auxiliary information at estimation stage and proposed ratio estimator for population mean. It is a well-established fact that ratio type estimators provide

[1] Division of Agricultural Statistics, SKUAST-K Shalimar, Srinagar-190025, Kashmir, India.
[2] School of Studies in Statistics, Vikram University, Ujjain-456010, Madhya Pradesh P., India.
[3] Department of Decision Sciences, School of Business, UPES, Dehradun, India.
E-mail: prayassharma02@gmail.com.

better efficiency in comparison to simple mean estimator if the study variable and auxiliary variable are positively correlated and the regression line pass through origin, and if on the other side the correlation between the study variable and auxiliary variable is positive and does not pass through origin but makes an intercept, in that case the regression method provides better efficiency than ratio, simple mean and product type estimator, and if the correlation between the study variable and auxiliary variable is negative, product estimator given by Robson (1957) is more efficient than simple mean estimator.

Further improvements are also achieved on the classical ratio estimator by introducing a large number of modified ratio estimators with the use of known parameters as coefficient of variation, coefficient of kurtosis, coefficient of skewness and population correlation coefficient. For more detailed discussion, one may refer to Cochran (1977), Kadilar and Cingi (2004, 2006), Koyuncu and Kadilar (2009), Murthy (1967), Sharma and Singh (2013), Prasad (1989), Rao (1991), Singh and Tailor (2003, 2005), Singh et al. (2004), Sisodia and Dwivedi (1981), Upadhyaya and Singh (1999) and Yan and Tian (2010).

Further, Subramani and Kumarapandiyan (2012) have taken initiative by proposing modified ratio estimator for estimating the population mean of the study variable by using the population deciles of the auxiliary variable.

Recently, Subzar et al. (2016) have proposed some estimators using population deciles and correlation coefficient of the auxiliary variable, also Subzar et al. (2017) have proposed some modified ratio type estimators using the quartile deviation and population deciles of auxiliary variable, and Subzar et al. (2017) have also proposed an efficient class of estimators by using the auxiliary information of population deciles, median and their linear combination with correlation coefficient and coefficient of variation.

In this paper we have envisaged a new class of improved ratio type estimators for estimation of population mean of the study variable using the information of quartile deviation, median, non-conventional measures of dispersion and their linear combination with correlation coefficient and coefficient of variation. Let $G = \{G_1, G_2, G_3, ..., G_N\}$ be a finite population of N distinct and identifiable units. Let y and x denote the study variable and the auxiliary variable taking values y_i and x_i respectively on the i^{th} unit ($i = 1, 2, \ldots, N$). Before discussing the proposed estimators, we will mention the estimators in the literature using the notations given in the next sub-section.

1.1. Notations

N	Population size
n	Sample size
$f = n/N$	Sampling fraction
Y	Study variable
X	Auxiliary variable
$\bar{X}, \bar{Y}$	Population means

$\bar{x}, \bar{y}$	Sample means
x, y	Sample totals
S_x, S_y	Population standard deviations
S_{xy}	Population covariance between variables
C_x, C_y	Population coefficient of variation
ρ	Population correlation coefficient
$B(.)$	Bias of the estimator
$MSE(.)$	Mean square error of the estimator
$\hat{\bar{Y}}_i$	Existing modified ratio estimator of $\bar{Y}$
$\hat{\bar{Y}}_{pj}$	Proposed modified ratio estimator of $\bar{Y}$
M_d	Population median of X
β_2	Population kurtosis
β_1	Population skewness
$QD = \frac{Q_3 - Q_1}{2}$	Population quartile deviation
$HL = median((X_j + X_k)/2, 1 \le j \le k \le N)$	Hodges-Lehmann estimator
$MR = \frac{X_{(1)} + X_{(N)}}{2}$	Population mid-range
$DM = \frac{D_1 + D_2 + ... + D_9}{9}$	Decile mean
$G = \frac{4}{N-1}\sum_{i=1}^{N}\left(\frac{2i - N - 1}{2N}\right)X_{(i)}$	Gini's Mean Difference
$D = \frac{2\sqrt{\pi}}{N(N-1)}\sum_{i=1}^{N}\left(i - \frac{N+1}{2}\right)X_{(i)}$	Downton's method
$S_{pw} = \frac{\sqrt{\pi}}{N^2}\sum_{i=1}^{N}(2i - N - 1)X_{(i)}$	Probability weighted moments

Subscript

i For existing estimators

j For proposed estimators

1.2. Estimators in the literature

Kadilar and Cingi (2004) have suggested ratio type estimators for the population mean in the simple random sampling using some known auxiliary information on coefficient of kurtosis and coefficient of variation. They showed that their suggested estimators are more efficient than traditional ratio estimator in the estimation of the population mean.
Kadilar & Cingi (2004) estimators are given by

$$\hat{\bar{Y}}_1 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{\bar{x}} \bar{X},$$

$$\hat{\bar{Y}}_2 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + C_x)} (\bar{X} + C_x),$$

$$\hat{\bar{Y}}_3 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \beta_2)} (\bar{X} + \beta_2),$$

$$\hat{\bar{Y}}_4 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\beta_2 + C_x)} (\bar{X}\beta_2 + C_x),$$

$$\hat{\bar{Y}}_5 = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \beta_2)} (\bar{X}C_x + \beta_2),$$

The biases, related constants and the MSE for Kadilar and Cingi (2004) estimators are respectively as follows:

$$B(\hat{\bar{Y}}_1) = \frac{(1-f)}{n} \frac{s_x^2}{\bar{Y}} R_1^2, \qquad R_1 = \frac{\bar{Y}}{\bar{X}} \qquad MSE(\hat{\bar{Y}}_1) = \frac{(1-f)}{n} (R_1^2 S_x^2 + S_y^2 (1-\rho^2)),$$

$$B(\hat{\bar{Y}}_2) = \frac{(1-f)}{n} \frac{s_x^2}{\bar{Y}} R_2^2, \quad R_2 = \frac{\bar{Y}}{(\bar{X} + C_x)} \qquad MSE(\hat{\bar{Y}}_2) = \frac{(1-f)}{n} (R_2^2 S_x^2 + S_y^2 (1-\rho^2)),$$

$$B(\hat{\bar{Y}}_3) = \frac{(1-f)}{n} \frac{s_x^2}{\bar{Y}} R_3^2, \quad R_3 = \frac{\bar{Y}}{(\bar{X} + \beta_2)} \qquad MSE(\hat{\bar{Y}}_3) = \frac{(1-f)}{n} (R_3^2 S_x^2 + S_y^2 (1-\rho^2)),$$

$$B(\hat{\bar{Y}}_4)=\frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_4^2,\quad R_4=\frac{\bar{Y}}{(\bar{X}\beta_2+C_x)}\qquad MSE(\hat{\bar{Y}}_4)=\frac{(1-f)}{n}(R_4^2S_x^2+S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_5)=\frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_5^2,\quad R_5=\frac{\bar{Y}}{(\bar{X}C_x+\beta_2)}\qquad MSE(\hat{\bar{Y}}_5)=\frac{(1-f)}{n}(R_5^2S_x^2+S_y^2(1-\rho^2)).$$

Kadilar and Cingi (2006) developed some modified ratio estimators using known value of coefficient of correlation, kurtosis and coefficient of variation as follows:

$$\hat{\bar{Y}}_6=\frac{\bar{y}+b(\bar{X}-\bar{x})}{(\bar{x}+\rho)}(\bar{X}+\rho),$$

$$\hat{\bar{Y}}_7=\frac{\bar{y}+b(\bar{X}-\bar{x})}{(\bar{x}C_x+\rho)}(\bar{X}C_x+\rho),$$

$$\hat{\bar{Y}}_8=\frac{\bar{y}+b(\bar{X}-\bar{x})}{(\bar{x}\rho+C_x)}(\bar{X}\rho+C_x),$$

$$\hat{\bar{Y}}_9=\frac{\bar{y}+b(\bar{X}-\bar{x})}{(\bar{x}\beta_2+\rho)}(\bar{X}\beta_2+\rho),$$

$$\hat{\bar{Y}}_{10}=\frac{\bar{y}+b(\bar{X}-\bar{x})}{(\bar{x}\rho+\beta_2)}(\bar{X}\rho+\beta_2).$$

The biases, related constants and the MSE for Kadilar and Cingi [6] estimators are respectively given by

$$B(\hat{\bar{Y}}_6)=\frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_6^2,\quad R_6=\frac{\bar{Y}}{\bar{X}+\rho}\qquad MSE(\hat{\bar{Y}}_6)=\frac{(1-f)}{n}(R_6^2S_x^2+S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_7)=\frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_7^2,\quad R_7=\frac{\bar{Y}C_x}{\bar{X}C_x+\rho}\qquad MSE(\hat{\bar{Y}}_7)=\frac{(1-f)}{n}(R_7^2S_x^2+S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_8)=\frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_8^2,\quad R_8=\frac{\bar{Y}\rho}{\bar{X}\rho+C_x}\qquad MSE(\hat{\bar{Y}}_8)=\frac{(1-f)}{n}(R_8^2S_x^2+S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_9)=\frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_9^2,\quad R_9=\frac{\bar{Y}\beta_2}{\bar{X}\beta_2+\rho}\qquad MSE(\hat{\bar{Y}}_9)=\frac{(1-f)}{n}(R_9^2S_x^2+S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{10})=\frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{10}^2,\quad R_{10}=\frac{\bar{Y}\rho}{\bar{X}\rho+\beta_2}\qquad MSE(\hat{\bar{Y}}_{10})=\frac{(1-f)}{n}(R_{10}^2S_x^2+S_y^2(1-\rho^2)).$$

Abid *et al.* (2016) suggested the following ratio estimators for the population mean $\bar{Y}$ in simple random sampling using non-conventional location parameters as auxiliary information. Estimators suggested by Abid *et al.* (2016) are given as:

$$\hat{\bar{Y}}_{11} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + TM)}(\bar{X} + TM),$$

$$\hat{\bar{Y}}_{12} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + TM)}(\bar{X}C_x + TM)$$

$$\hat{\bar{Y}}_{13} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + TM)}(\bar{X}\rho + TM),$$

$$\hat{\bar{Y}}_{14} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + MR)}(\bar{X} + MR),$$

$$\hat{\bar{Y}}_{15} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + MR)}(\bar{X}C_x + MR),$$

$$\hat{\bar{Y}}_{16} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + MR)}(\bar{X}\rho + MR)$$

$$\hat{\bar{Y}}_{17} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + HL)}(\bar{X} + HL),$$

$$\hat{\bar{Y}}_{18} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + HL)}(\bar{X}C_x + HL),$$

$$\hat{\bar{Y}}_{19} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + HL)}(\bar{X}\rho + HL)$$

The biases, related constants and the mean square error (MSE) for Abid *et al.* (2016) estimators are respectively given by:

$$B(\hat{\bar{Y}}_{11}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{11}^2, \qquad R_{11} = \frac{\bar{Y}}{(\bar{X} + TM)} \qquad MSE(\hat{\bar{Y}}_{11}) = \frac{(1-f)}{n}(R_{11}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{12}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{12}^2, \qquad R_{12} = \frac{\bar{Y}C_x}{(\bar{X}C_x + TM)} \qquad MSE(\hat{\bar{Y}}_{12}) = \frac{(1-f)}{n}(R_{12}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{13}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{13}^2, \qquad R_{13} = \frac{\bar{Y}\rho}{(\bar{X}\rho + TM)} \qquad MSE(\hat{\bar{Y}}_{13}) = \frac{(1-f)}{n}(R_{13}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{14}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{14}^2, \qquad R_{14} = \frac{\bar{Y}}{(\bar{X} + MR)} \qquad MSE(\hat{\bar{Y}}_{14}) = \frac{(1-f)}{n}(R_{14}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{15}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{15}^2, \qquad R_{15} = \frac{\bar{Y}C_x}{(\bar{X}C_x + MR)} \qquad MSE(\hat{\bar{Y}}_{15}) = \frac{(1-f)}{n}(R_{15}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{16}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{16}^2, \qquad R_{16} = \frac{\bar{Y}\rho}{(\bar{X}\rho + MR)} \qquad MSE(\hat{\bar{Y}}_{16}) = \frac{(1-f)}{n}(R_{16}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{17}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{17}^2, \qquad R_{17} = \frac{\bar{Y}}{(\bar{X} + HL)} \qquad MSE(\hat{\bar{Y}}_{17}) = \frac{(1-f)}{n}(R_{17}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{18}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{18}^2, \qquad R_{18} = \frac{\bar{Y}C_x}{(\bar{X}C_x + HL)} \qquad MSE(\hat{\bar{Y}}_{18}) = \frac{(1-f)}{n}(R_{18}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{19}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{19}^2, \qquad R_{19} = \frac{\bar{Y}\rho}{(\bar{X}\rho + HL)} \qquad MSE(\hat{\bar{Y}}_{19}) = \frac{(1-f)}{n}(R_{19}^2 S_x^2 + S_y^2(1-\rho^2)).$$

Abid *et al.* (2016) suggested the following ratio estimators for the population mean $\bar{Y}$ in simple random sampling using Decile mean, with linear combination of population correlation coefficient and population coefficient of variation as auxiliary information. Estimators suggested by Abid *et al.* (2016) are given as:

$$\hat{\bar{Y}}_{20} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + DM)}(\bar{X} + DM)$$

$$\hat{\bar{Y}}_{21} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + DM)}(\bar{X}C_x + DM)$$

$$\hat{\bar{Y}}_{22} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + DM)}(\bar{X}\rho + DM)$$

The biases, related constants and the mean square error (MSE) for Abid *et al.* (2016) estimators are respectively given by:

$$B(\hat{\bar{Y}}_{20}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{20}^2, \qquad R_{20} = \frac{\bar{Y}}{(\bar{X} + DM)} \qquad MSE(\hat{\bar{Y}}_{20}) = \frac{(1-f)}{n}(R_{20}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{21}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{21}^2, \qquad R_{21} = \frac{\bar{Y}C_x}{(\bar{X}C_x + DM)} \qquad MSE(\hat{\bar{Y}}_{21}) = \frac{(1-f)}{n}(R_{21}^2 S_x^2 + S_y^2(1-\rho^2)),$$

$$B(\hat{\bar{Y}}_{22}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}}R_{22}^2, \qquad R_{22} = \frac{\bar{Y}\rho}{(\bar{X}\rho + DM)} \qquad MSE(\hat{\bar{Y}}_{22}) = \frac{(1-f)}{n}(R_{22}^2 S_x^2 + S_y^2(1-\rho^2)),$$

2. Improved ratio estimators

Motivated by the estimators mentioned in Section 1.2, in this section we suggest some improved class of estimators by using the auxiliary information median, quartile deviation, Gini's mean difference, Downton's Method, Probability Weighted Moments and their linear combinations with correlation coefficient and

coefficient of variation in survey sampling, and the suggested estimators are given below as:

$$\hat{\bar{Y}}_{p1} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \Psi_1)}(\bar{X} + \Psi_1)$$

$$\hat{\bar{Y}}_{p2} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \Psi_2)}(\bar{X} + \Psi_2)$$

$$\hat{\bar{Y}}_{p3} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \Psi_3)}(\bar{X} + \Psi_3)$$

$$\hat{\bar{Y}}_{p4} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \Psi_4)}(\bar{X} + \Psi_4)$$

$$\hat{\bar{Y}}_{p5} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \Psi_5)}(\bar{X} + \Psi_5)$$

$$\hat{\bar{Y}}_{p6} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x} + \Psi_6)}(\bar{X} + \Psi_6)$$

$$\hat{\bar{Y}}_{p7} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + \Psi_1)}(\bar{X}\rho + \Psi_1)$$

$$\hat{\bar{Y}}_{p8} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + \Psi_2)}(\bar{X}\rho + \Psi_2)$$

$$\hat{\bar{Y}}_{p9} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + \Psi_3)}(\bar{X}\rho + \Psi_3)$$

$$\hat{\bar{Y}}_{p10} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + \Psi_4)}(\bar{X}\rho + \Psi_4)$$

$$\hat{\bar{Y}}_{p11} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + \Psi_5)}(\bar{X}\rho + \Psi_5)$$

$$\hat{\bar{Y}}_{p12} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}\rho + \Psi_6)}(\bar{X}\rho + \Psi_6)$$

$$\hat{\bar{Y}}_{p13} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \Psi_1)}(\bar{X}C_x + \Psi_1)$$

$$\hat{\bar{Y}}_{p14} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \Psi_2)}(\bar{X}C_x + \Psi_2)$$

$$\hat{\bar{Y}}_{p15} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \Psi_3)}(\bar{X}C_x + \Psi_3)$$

$$\hat{\bar{Y}}_{p16} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \Psi_4)}(\bar{X}C_x + \Psi_4)$$

$$\hat{\bar{Y}}_{p17} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \Psi_5)}(\bar{X}C_x + \Psi_5)$$

$$\hat{\bar{Y}}_{p18} = \frac{\bar{y} + b(\bar{X} - \bar{x})}{(\bar{x}C_x + \Psi_6)}(\bar{X}C_x + \Psi_6)$$

where $\Psi_1 = (M_d + G)$, $\Psi_2 = (M_d + D)$, $\Psi_3 = (M_d + S_{pw})$, $\Psi_4 = (QD + G)$, $\Psi_5 = (QD + D)$ and $\Psi_6 = (QD + S_{pw})$

The bias, related constant and the MSE for the first proposed estimator can be obtained as follows:

MSE of this estimator can be found using Taylor series method defined as

$$h(\bar{x}, \bar{y}) \cong h(\bar{X}, \bar{Y}) + \frac{\partial h(c,d)}{\partial c}|_{\bar{X},\bar{Y}} (\bar{x} - \bar{X}) + \frac{\partial h(c,d)}{\partial d}|_{\bar{X},\bar{Y}} (\bar{y} - \bar{Y}) \qquad (2.1)$$

where $h(\bar{x}, \bar{y}) = \hat{R}_{p1}$ and $\boldsymbol{h(\bar{X}, \bar{Y}) = R.}$

As shown in Wolter (1985), (2.1) can be applied to the proposed estimator in order to obtain MSE equation as follows:

$$\hat{R}_{pj} - R \cong \frac{\partial((\bar{y} + b(\bar{X} - \bar{x}))/(\bar{x} + \Psi_k)}{\partial \bar{x}}|_{\bar{X},\bar{Y}} (\bar{x} - \bar{X}) + \frac{\partial((\bar{y} + b(\bar{X} - \bar{x}))/(\bar{x} + \Psi_k)}{\partial \bar{y}}|_{\bar{X},\bar{Y}} (\bar{y} - \bar{Y})$$

$$\cong -\left(\frac{\bar{y}}{(\bar{x} + \Psi_k)^2} + \frac{b(\bar{X} + \Psi_k)}{(\bar{x} + \Psi_k)^2}\right)|_{\bar{X},\bar{Y}} (\bar{x} - \bar{X}) + \frac{1}{(\bar{x} + \Psi_k)}|_{\bar{X},\bar{Y}} (\bar{y} - \bar{Y})$$

$$E(\hat{R}_{pj} - R)^2 \cong \frac{(\bar{Y} + B(\bar{X} + \Psi_k))^2}{(\bar{X} + \Psi_k)^4} V(\bar{x}) - \frac{2(\bar{Y} + B(\bar{X} + \Psi_k))}{(\bar{X} + \Psi_k)^3} Cov(\bar{x}, \bar{y}) + \frac{1}{(\bar{X} + \varphi_1)^2} V(\bar{y})$$

$$\cong \frac{1}{(\bar{X} + \Psi_k)^2}\left\{\frac{(\bar{Y} + B(\bar{X} + \Psi_k))^2}{(\bar{X} + \Psi_k)^2} V(\bar{x}) - \frac{2(\bar{Y} + B(\bar{X} + \Psi_k)}{(\bar{X} + \Psi_k)} Cov(\bar{x}, \bar{y}) + V(\bar{y})\right\}$$

where $B = \frac{s_{xy}}{s_x^2} = \frac{\rho s_x s_y}{s_x^2} = \frac{\rho s_y}{s_x}$. Note that we omit the difference of $(E(b) - B)$.

$$MSE(\bar{y}_{pj}) = (\bar{X} + \Psi_k)^2 E(\hat{R}_{pj} - R)^2 \cong \frac{(\bar{Y} + B(\bar{X} + \Psi_k))^2}{(\bar{X} + \Psi_k)^2} V(\bar{x}) - \frac{2(\bar{Y} + B(\bar{X} + \Psi_k))}{(\bar{X} + \Psi_k)} Cov(\bar{x}, \bar{y}) + V(\bar{y})$$

$$\cong \frac{\bar{Y}^2 + 2B(\bar{X} + \Psi_k)\bar{Y} + B^2(\bar{X} + \Psi_k)^2}{(\bar{X} + \Psi_k)^2} V(\bar{x}) - \frac{2\bar{Y} + 2B(\bar{X} + \Psi_k)}{(\bar{X} + \Psi_k)} Cov(\bar{x}, \bar{y}) + V(\bar{y})$$

$$\cong \frac{(1-f)}{n}\left\{\left(\frac{\bar{Y}^2}{(\bar{X} + \Psi_k)^2} + \frac{2B\bar{Y}}{(\bar{X} + \Psi_k)} + B^2\right)S_x^2 - \left(\frac{2\bar{Y}}{(\bar{X} + \Psi_k)} + 2B\right)S_{xy} + S_y^2\right\}$$

$$\cong \frac{(1-f)}{n}(R^2 S_x^2 + 2BRS_x^2 + B^2 S_x^2 - 2RS_{xy} - 2BS_{xy} + S_y^2)$$

$$MSE(\bar{y}_{pj}) \cong \frac{(1-f)}{n}(R^2 S_x^2 + 2R\rho S_x S_y + \rho^2 S_y^2 - 2R\rho S_x S_y - 2\rho^2 S_y^2 + S_y^2$$

$$\cong \frac{(1-f)}{n}(R^2 S_x^2 - \rho^2 S_y^2 + S_y^2) \cong \frac{(1-f)}{n}(R^2 S_x^2 + S_y^2(1-\rho^2)$$

Similarly, the bias is obtained as

$$Bias(\bar{y}_{pj}) \cong \frac{(1-f)}{n}\frac{S_x^2}{\bar{Y}} R_j^2$$

Thus, the bias and MSE of the proposed estimator is given below:

$$B(\hat{\bar{Y}}_{pj}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}} R_j^2, \quad R_j = \frac{\bar{Y}}{\bar{X} + \Psi_k} \qquad MSE(\hat{\bar{Y}}_{pj}) = \frac{(1-f)}{n}(R_j^2 S_x^2 + S_y^2(1-\rho^2)),$$

where $J = 1,2,...,6$ and $k = 1,2,...,6$

Similarly we can obtain the bias, constant and mean square error for the other proposed estimators as follows:

$$B(\hat{\bar{Y}}_{pj}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}} R_j^2, \quad R_j = \frac{\bar{Y}\rho}{\bar{X}\rho + \Psi_k} \qquad MSE(\hat{\bar{Y}}_{pj}) = \frac{(1-f)}{n}(R_j^2 S_x^2 + S_y^2(1-\rho^2)),$$

where $J = 7,8,...,12$ and $k = 1,2,...,6$

$$B(\hat{\bar{Y}}_{pj}) = \frac{(1-f)}{n}\frac{s_x^2}{\bar{Y}} R_j^2, \quad R_j = \frac{\bar{Y}C_x}{\bar{X}C_x + \Psi_k} \qquad MSE(\hat{\bar{Y}}_{pj}) = \frac{(1-f)}{n}(R_j^2 S_x^2 + S_y^2(1-\rho^2)),$$

where $J = 13,14,...,18$ and $k = 1,2,...,6$

3. Efficiency comparisons

From the expressions of the MSE of the proposed estimators and the existing estimators, we have derived the conditions for which the proposed estimators are more efficient than the usual and existing modified ratio estimators. They are given as follows.

3.1 Comparison with the classical ratio estimator

Modified proposed ratio estimators are more efficient than that of the classical ratio estimator if $MSE(\bar{\bar{Y}}_{pj}) \le MSE(\bar{\bar{Y}}_r)$,

$$\frac{(1-f)}{n}(R_{pj}^2 S_x^2 + S_y^2(1-\rho^2) \le \frac{(1-f)}{n}(S_y^2 + R^2 S_x^2 - 2R\rho S_x S_y),$$

$$R_{pj}^2 S_x^2 - \rho^2 S_y^2 - R^2 S_x^2 + 2R\rho S_x S_y \le 0,$$

$$(\rho S_y - RS_x)^2 - R_{pj}^2 S_x^2 \ge 0,$$

$$(\rho S_y - RS_x + R_{pj}^2)(\rho S_y - RS_x - R_{pj} S_x) \ge 0.$$

Condition I: $(\rho S_y - RS_x + R_{pj} S_x) \le 0$ and $(\rho S_y - RS_x - R_{pj} S_x) \le 0$

After solving the condition I, we get

$$\left(\frac{RS_y - RS_x}{S_x}\right) \le R_{pj} \le \left(\frac{RS_x - \rho S_y}{S_x}\right).$$

Hence,

$MSE(\bar{\bar{Y}}_{pj}) \le MSE(\bar{\bar{Y}}_r)$,

$$\left(\frac{\rho S_y - RS_x}{S_x}\right) \le R_{pj} \le \left(\frac{RS_x - \rho S_y}{S_x}\right), \text{ or}$$

$$\left(\frac{RS_x - \rho S_y}{S_x}\right) \le R_{pj} \le \left(\frac{\rho S_y - RS_x}{S_x}\right). \quad \text{Where } j = 1,2,....,18.$$

3.2 Comparisons with existing ratio estimators

$$MSE(\widehat{\overline{Y}}_{pj}) \le MSE(\widehat{\overline{Y}}_{i}),$$

$$\frac{(1-f)}{n}(R_{pj}^2 S_x^2 + S_y^2(1-\rho^2) \le \frac{(1-f)}{n}(R_i^{\,2} S_x^2 + S_y^2(1-\rho^2)),$$

$$R_{pj}^2 S_x^2 \le R_i^2 S_x^2,$$

$$R_{pj} \le R_i,$$

where $j = 1,2,....,18$ and $i = 1,2,....,22.$

4. Empirical study

The performances of the proposed ratio estimators are evaluated and compared with the mentioned ratio estimators in Section 1.2 by using the data of the three natural populations. For the population I and II we use the data of Singh and Chaudhary (1986) page 177 and for the population III we use the data of Murthy (1967) page 228, in which fixed capital is denoted by X (auxiliary variable) and output of 80 factories is denoted by Y (study variable). We apply the proposed and existing estimators to these data sets and the statistics of these populations are given in Table 1.

From Table 2a and Table 2b, we observe that the proposed estimators are more efficient than all of the estimators in the literature as their Bias, Constant and Mean Square error are much lower than the existing estimators.

The percentage relative efficiency (PRE) of the proposed estimators (*p*), with respect to the existing estimators (*e*), is computed by

$$PRE = \frac{MSE\,of\ Existing\ Estimator}{MSE\,of\ propoesdestimator} \times 100$$

These PRE values are given in Table 3, for the population I. From this table, it is clearly evident that the proposed estimators are quiet efficient with respect to the estimators in the literature. Similarly, we can calculate the PRE values for population II and population III respectively by using the same formula mentioned above.

Table 1. Characteristics of the populations

Parameters	Population 1	Population 2	Population 3
N	34	34	80
n	20	20	20
$\overline{Y}$	856.4117	856.4117	5182.637
$\overline{X}$	199.4412	208.8823	1126.463

Table 1. Characteristics of the populations (cont.)

Parameters	Population 1	Population 2	Population 3
ρ	0.4453	0.4491	0.941
S_y	733.1407	733.1407	1835.659
C_y	0.8561	0.8561	0.354193
S_x	150.2150	150.5059	845.610
C_x	0.7531	0.7205	0.7506772
β_2	1.0445	0.0978	-0.063386
β_1	1.1823	0.9782	1.050002
M_d	142.50	150	757.5
TM	89.375	162.25	931.562
MR	165.562	284.5	1795.5
HL	320	190	1040.5
QD	184	80.25	588.125
G	162.996	155.446	901.081
D	144.481	140.891	801.381
S_{pw}	206.944	199.961	791.364
DM	206.944	234.82	1150.7

Table 2a. The Statistical Analysis of the Estimators for the Populations

Estimators	Population I		Population II		Population III	
	Constant	Bias	Constant	Bias	Constant	Bias
$\widehat{\overline{Y}}_r$	4.294	4.940	4.100	4.270	4.601	60.877
$\widehat{\overline{Y}}_1$	4.294	10.0023	4.100	9.1539	4.601	36.5063
$\widehat{\overline{Y}}_2$	4.278	9.9272	4.086	9.0911	4.598	36.4577
$\widehat{\overline{Y}}_3$	4.272	9.8983	4.098	9.1454	4.601	36.5104
$\widehat{\overline{Y}}_4$	4.279	9.9303	3.960	8.5387	4.650	37.2861
$\widehat{\overline{Y}}_5$	4.264	9.8646	4.097	9.142	4.601	36.5117
$\widehat{\overline{Y}}_6$	4.284	9.9578	4.091	9.1147	4.597	36.4453
$\widehat{\overline{Y}}_7$	4.281	9..9432	4.088	9.0995	4.596	36.4251

Table 2a. The Statistical Analysis of the Estimators for the Populations (cont.)

Estimators	Population I		Population II		Population III	
	Constant	Bias	Constant	Bias	Constant	Bias
$\hat{\bar{Y}}_8$	4.258	9.8348	4.069	9.0149	4.598	36.4546
$\hat{\bar{Y}}_9$	4.285	9.9597	4.011	8.763	4.662	37.4882
$\hat{\bar{Y}}_{10}$	4.244	9.7711	4.096	9.1349	4.601	36.5106
$\hat{\bar{Y}}_{11}$	2.3463	2.986	2.3076	2.9	2.518	32.81
$\hat{\bar{Y}}_{12}$	2.0427	2.263	1.973	2.12	2.189	24.79
$\hat{\bar{Y}}_{13}$	1.4993	1.219	1.5021	1.229	2.449	31.03
$\hat{\bar{Y}}_{14}$	1.6487	1.475	1.7358	1.641	1.774	16.23
$\hat{\bar{Y}}_{15}$	1.3718	1.021	1.4185	1.096	1.473	11.23
$\hat{\bar{Y}}_{16}$	0.9329	0.472	1.0167	0.563	1.708	15.10
$\hat{\bar{Y}}_{17}$	2.233	2.706	2.147	2.51	2.392	29.59
$\hat{\bar{Y}}_{18}$	1.930	2.021	1.812	1.788	2.063	22.01
$\hat{\bar{Y}}_{19}$	1.398	1.060	1.355	1.980	2.322	27.90
$\hat{\bar{Y}}_{20}$	2.107	2.137	1.9301	2.2087	2.276	26.800
$\hat{\bar{Y}}_{21}$	1.806	1.483	1.6013	1.3964	1.949	19.650
$\hat{\bar{Y}}_{22}$	1.289	0.800	1.1703	0.7459	2.206	25.188
$\hat{\bar{Y}}_{p1}$	1.6960	1.5604	1.6651	1.5098	1.9813	20.312
$\hat{\bar{Y}}_{p2}$	1.7606	1.6815	1.7136	1.599	2.0598	21.953
$\hat{\bar{Y}}_{p3}$	1.5602	1.3205	1.5324	1.2788	2.0681	22.129
$\hat{\bar{Y}}_{p4}$	1.8955	1.9490	1.9263	2.0207	1.8608	17.916
$\hat{\bar{Y}}_{p5}$	1.9764	2.1191	1.9915	2.1598	1.9299	19.271
$\hat{\bar{Y}}_{p6}$	1.7274	1.6187	1.751	1.6696	1.9371	19.416
$\hat{\bar{Y}}_{p7}$	0.9671	0.5074	0.9633	0.5053	1.7938	16.650

Table 2a. The Statistical Analysis of the Estimators for the Populations (cont.)

Estimators	Population I		Population II		Population III	
	Constant	Bias	Constant	Bias	Constant	Bias
$\hat{\bar{Y}}_{p8}$	1.0148	0.5586	0.9997	0.5443	1.8622	17.942
$\hat{\bar{Y}}_{p9}$	0.8701	0.4107	0.8666	0.409	1.8693	18.079
$\hat{\bar{Y}}_{p10}$	1.1177	0.6777	1.1672	0.7419	1.9130	18.936
$\hat{\bar{Y}}_{p11}$	1.1818	0.7577	1.2211	0.8121	1.9909	20.508
$\hat{\bar{Y}}_{p12}$	0.9902	0.5318	1.0283	0.5758	1.9991	20.677
$\hat{\bar{Y}}_{p13}$	1.4153	1.0866	1.3533	0.9973	1.5540	12.495
$\hat{\bar{Y}}_{p14}$	1.4752	1.1806	1.3979	1.0642	1.6184	13.552
$\hat{\bar{Y}}_{p15}$	1.2908	0.9038	1.2329	0.8278	1.6252	13.666
$\hat{\bar{Y}}_{p16}$	1.6021	1.3923	1.5977	1.3901	1.6667	14.373
$\hat{\bar{Y}}_{p17}$	1.6793	1.5298	1.6603	1.5011	1.7410	15.684
$\hat{\bar{Y}}_{p18}$	1.4444	1.1317	1.4326	1.1176	1.7489	15.825

Table 2b. The Statistical Analysis of the Estimators for the Populations

Estimators	Pop I	Pop II	Pop III	Estimators	Pop I	Pop II	Pop III
	MSE	MSE	MSE		MSE	MSE	MSE
$\hat{\bar{Y}}_{r}$	10960.76	10539.27	189775.1	$\hat{\bar{Y}}_{21}$	10386.83	10030.11	116239.3
$\hat{\bar{Y}}_{1}$	17437.7	16673.5	193998.1	$\hat{\bar{Y}}_{22}$	9644.04	9472.95	144936.7
$\hat{\bar{Y}}_{2}$	17373.3	16619.6	193746.2	$\hat{\bar{Y}}_{p1}$	10208.16	10333.07	119741.5
$\hat{\bar{Y}}_{3}$	17348.6	16666.1	194019.4	$\hat{\bar{Y}}_{p2}$	10311.83	10203.59	128249.9
$\hat{\bar{Y}}_{4}$	17376.0	16146.6	198039.9	$\hat{\bar{Y}}_{p3}$	10002.72	9929.39	129161.3
$\hat{\bar{Y}}_{5}$	17319.8	16663.3	194026.4	$\hat{\bar{Y}}_{p4}$	10540.91	10564.74	107326.5
$\hat{\bar{Y}}_{6}$	17399.5	16639.9	193682.3	$\hat{\bar{Y}}_{p5}$	10686.6	10683.87	114349.5

Table 2b. The Statistical Analysis of the Estimators for the Populations (cont.)

Estimators	Pop I	Pop II	Pop III	Estimators	Pop I	Pop II	Pop III
	MSE	MSE	MSE		MSE	MSE	MSE
$\widehat{\overline{Y}}_7$	17387.1	16626.9	193577.6	$\widehat{\overline{Y}}_{p6}$	10258.09	10264.06	115098.9
$\widehat{\overline{Y}}_8$	17294.2	16554.4	193730.5	$\widehat{\overline{Y}}_{p7}$	9306.32	9266.94	100762.1
$\widehat{\overline{Y}}_9$	17401.1	16338.7	199087.0	$\widehat{\overline{Y}}_{p8}$	9350.19	9300.31	107457.3
$\widehat{\overline{Y}}_{10}$	17239.7	16654.2	194020.7	$\widehat{\overline{Y}}_{p9}$	9223.53	9184.47	108172.8
$\widehat{\overline{Y}}_{11}$	11429.08	11317.28	184446.20	$\widehat{\overline{Y}}_{p10}$	9452.18	9469.56	112609.8
$\widehat{\overline{Y}}_{12}$	10809.99	10649.40	142903.20	$\widehat{\overline{Y}}_{p11}$	9250.7	9529.64	120761.3
$\widehat{\overline{Y}}_{13}$	9915.81	9886.21	175238.70	$\widehat{\overline{Y}}_{p12}$	9327.27	9327.31	121636.1
$\widehat{\overline{Y}}_{14}$	10134.39	10239.11	98755.61	$\widehat{\overline{Y}}_{p13}$	9802.37	9688.30	79228.58
$\widehat{\overline{Y}}_{15}$	9745.79	9772.39	72582.52	$\widehat{\overline{Y}}_{p14}$	9882.86	9745.56	84709.31
$\widehat{\overline{Y}}_{16}$	9275.87	9316.02	92644.60	$\widehat{\overline{Y}}_{p15}$	9645.86	9543.10	85298.12
$\widehat{\overline{Y}}_{17}$	11189.04	10983.77	167778.60	$\widehat{\overline{Y}}_{p16}$	10064.19	10024.69	88963.27
$\widehat{\overline{Y}}_{18}$	10602.02	10365.55	128487.60	$\widehat{\overline{Y}}_{p17}$	10181.93	10119.77	95756.01
$\widehat{\overline{Y}}_{19}$	9779.43	9690.50	158990.7	$\widehat{\overline{Y}}_{p18}$	9841.01	9791.32	96489.39
$\widehat{\overline{Y}}_{20}$	10934.74	10571.58	153292.6				

Table 3. PRE of the Proposed Estimators with the Estimators in the Literature for Population I.

	$\widehat{\overline{Y}}_{p1}$	$\widehat{\overline{Y}}_{p2}$	$\widehat{\overline{Y}}_{p3}$	$\widehat{\overline{Y}}_{p4}$	$\widehat{\overline{Y}}_{p5}$	$\widehat{\overline{Y}}_{p6}$	$\widehat{\overline{Y}}_{p7}$	$\widehat{\overline{Y}}_{p8}$	$\widehat{\overline{Y}}_{p9}$
$\widehat{\overline{Y}}_r$	107.372	106.293	109.578	103.983	102.565	106.85	117.778	117.225	118.835
$\widehat{\overline{Y}}_1$	170.821	169.103	174.329	165.428	163.173	169.991	187.381	186.495	189.056
$\widehat{\overline{Y}}_2$	170.190	168.479	173.685	164.817	162.570	169.363	186.689	185.806	188.358
$\widehat{\overline{Y}}_3$	169.948	168.239	173.438	164.583	162.339	169.122	186.423	185.542	188.090

Table 3. PRE of the Proposed Estimators with the Estimators in the Literature for Population I. (cont.)

	$\widehat{\bar{Y}}_{p1}$	$\widehat{\bar{Y}}_{p2}$	$\widehat{\bar{Y}}_{p3}$	$\widehat{\bar{Y}}_{p4}$	$\widehat{\bar{Y}}_{p5}$	$\widehat{\bar{Y}}_{p6}$	$\widehat{\bar{Y}}_{p7}$	$\widehat{\bar{Y}}_{p8}$	$\widehat{\bar{Y}}_{p9}$
$\widehat{\bar{Y}}_{4}$	170.216	168.505	173.712	164.843	162.596	169.389	186.718	185.835	188.387
$\widehat{\bar{Y}}_{5}$	169.666	167.960	173.150	164.310	162.070	168.841	186.114	185.234	187.778
$\widehat{\bar{Y}}_{6}$	170.447	168.733	173.947	165.066	162.816	169.618	186.970	186.087	188.642
$\widehat{\bar{Y}}_{7}$	170.325	168.613	173.823	164.948	162.700	169.498	186.837	185.954	188.508
$\widehat{\bar{Y}}_{8}$	169.415	167.712	172.895	164.067	161.830	168.592	185.839	184.960	187.500
$\widehat{\bar{Y}}_{9}$	170.462	168.748	173.963	165.081	162.831	169.634	186.988	186.104	188.659
$\widehat{\bar{Y}}_{10}$	168.881	167.183	172.350	163.550	161.320	168.061	185.253	184.378	186.910
$\widehat{\bar{Y}}_{11}$	111.960	110.834	114.259	108.425	106.947	111.416	122.814	122.233	123.912
$\widehat{\bar{Y}}_{12}$	105.895	104.831	108.070	102.552	101.154	105.381	116.161	115.612	117.200
$\widehat{\bar{Y}}_{13}$	97.1361	96.1595	99.1311	94.0697	92.7873	96.6641	106.552	106.049	107.505
$\widehat{\bar{Y}}_{14}$	99.2773	98.2792	101.316	96.1434	94.8326	98.7949	108.901	108.387	109.875
$\widehat{\bar{Y}}_{15}$	95.4705	94.5107	97.4314	92.4568	91.1963	95.0067	104.725	104.230	105.662
$\widehat{\bar{Y}}_{16}$	90.8672	89.9536	92.7334	87.9987	86.7990	90.4257	99.6762	99.2051	100.567
$\widehat{\bar{Y}}_{17}$	109.608	108.506	111.860	106.148	104.701	109.076	120.234	119.666	121.309
$\widehat{\bar{Y}}_{18}$	103.858	102.814	105.991	100.579	99.2085	103.353	113.926	113.388	114.945
$\widehat{\bar{Y}}_{19}$	95.8001	94.8370	97.7677	92.7759	91.5111	95.3346	105.087	104.590	106.027
$\widehat{\bar{Y}}_{20}$	107.117	106.040	109.317	103.736	102.322	106.597	117.502	116.946	118.552
$\widehat{\bar{Y}}_{21}$	101.750	100.727	103.840	98.5382	97.1949	101.255	111.614	111.086	112.612
$\widehat{\bar{Y}}_{22}$	94.4738	93.5240	96.4141	91.4915	90.2442	94.0148	103.632	103.142	104.559

Table 3. PRE of the Proposed Estimators with the Estimators in the Literature for Population I. (cont.)

	$\widehat{\bar{Y}}_{p10}$	$\widehat{\bar{Y}}_{p11}$	$\widehat{\bar{Y}}_{p12}$	$\widehat{\bar{Y}}_{p13}$	$\widehat{\bar{Y}}_{p14}$	$\widehat{\bar{Y}}_{p15}$	$\widehat{\bar{Y}}_{p16}$	$\widehat{\bar{Y}}_{p17}$	$\widehat{\bar{Y}}_{p18}$
$\widehat{\bar{Y}}_r$	115.96	118.486	117.513	111.817	110.907	113.632	108.909	107.649	111.378
$\widehat{\bar{Y}}_1$	184.483	188.501	186.954	177.893	176.444	180.779	173.265	171.261	177.194
$\widehat{\bar{Y}}_2$	183.802	187.805	186.264	177.236	175.792	180.111	172.625	170.629	176.540
$\widehat{\bar{Y}}_3$	183.541	187.538	185.999	176.984	175.542	179.855	172.379	170.386	176.289
$\widehat{\bar{Y}}_4$	183.831	187.834	186.292	177.263	175.820	180.139	172.652	170.655	176.567
$\widehat{\bar{Y}}_5$	183.236	187.227	185.690	176.690	175.251	179.557	172.093	170.103	175.996
$\widehat{\bar{Y}}_6$	184.079	188.088	186.544	177.503	176.057	180.383	172.885	170.886	176.806
$\widehat{\bar{Y}}_7$	183.948	187.954	186.411	177.376	175.932	180.255	172.762	170.764	176.680
$\widehat{\bar{Y}}_8$	182.965	186.950	185.415	176.429	174.992	179.291	171.839	169.852	175.736
$\widehat{\bar{Y}}_9$	184.096	188.106	186.562	177.519	176.074	180.400	172.901	170.902	176.822
$\widehat{\bar{Y}}_{10}$	182.389	186.361	184.831	175.873	174.440	178.726	171.297	169.317	175.182
$\widehat{\bar{Y}}_{11}$	120.915	123.548	122.534	116.595	115.645	118.487	113.562	112.249	116.137
$\widehat{\bar{Y}}_{12}$	114.365	116.856	115.897	110.279	109.381	112.069	107.410	106.168	109.846
$\widehat{\bar{Y}}_{13}$	104.905	107.190	106.310	101.157	100.333	102.799	98.5257	97.3864	100.760
$\widehat{\bar{Y}}_{14}$	107.217	109.553	108.653	103.387	102.545	105.065	100.698	99.5331	102.981
$\widehat{\bar{Y}}_{15}$	103.106	105.352	104.487	99.4228	98.6131	101.036	96.8363	95.7165	99.0324
$\widehat{\bar{Y}}_{16}$	98.1347	100.272	99.4489	94.6288	93.8582	96.1643	92.1671	91.1013	94.2573
$\widehat{\bar{Y}}_{17}$	118.375	120.953	119.961	114.146	113.217	115.998	111.177	109.891	113.698
$\widehat{\bar{Y}}_{18}$	112.165	114.608	113.667	108.158	107.277	109.913	105.344	104.126	107.733
$\widehat{\bar{Y}}_{19}$	103.462	105.716	104.848	99.7660	98.9534	101.385	97.1706	96.0469	99.3743
$\widehat{\bar{Y}}_{20}$	115.685	118.204	117.234	111.552	110.643	113.362	108.650	107.394	111.114
$\widehat{\bar{Y}}_{21}$	109.888	112.282	111.360	105.962	105.099	107.682	103.206	102.012	105.546
$\widehat{\bar{Y}}_{22}$	102.030	104.252	103.396	98.3848	97.5835	99.9811	95.8253	94.7172	97.9985

5. Conclusion

From the present study we conclude that our proposed estimators perform better than the existing estimators in the literature as their mean square error and bias are much lower than that of the existing estimators. Hence, we strongly recommend that our proposed estimators are preferred over the existing estimators in practical applications.

Acknowledgements

The authors are grateful to two anonymous reviewers for giving many insightful and constructive comments and suggestions which led to the improvement of the earlier manuscript.

REFERENCES

ABID, M., ABBAS, N., NAZIR, H. Z., LIN, Z., (2016). Enhancing the mean ratio estimators for estimating population mean using non-conventional location parameters, Revista Colombiana de Estadistica, 39 (1), pp. 63–79.

ABID, M., SHERWANI, R. A. K., ABBAS, N., NAWAZ, T., (2016). Some improved modified ratio estimators based on decile mean of an auxiliary variable, Pakistan Journal of Statistic and Operation Research, 12 (4), pp. 787–797.

COCHRAN, W. G., (1977). Sampling Techniques, Third Edition, Wiley Eastern Limited, New York.

KADILAR C., CINGI, H., (2006). An improvement in estimating the population mean by using the correlation coefficient, Hacettepe Journal of Mathematics and Statistics, 35 (1), pp. 103–109.

KADILAR, C., CINGI, H., (2004). Ratio estimators in simple random sampling, Applied Mathematics and Computation, 151, pp. 893–902.

KOYUNCU, N., KADILAR, C., (2009). Efficient Estimators for the Population mean, Hacettepe Journal of Mathematics and Statistics, 38 (2), pp. 217–225.

MURTHY, M. N., (1967). Sampling Theory and Methods, 1 ed., Statistical Publishing Society, India.

Prasad, B., (1989). Some improved ratio type estimators of population mean and ratio in finite population sample surveys, Communications in Statistics-Theory and Methods, 18, pp. 379–392.

RAO, T. J., (1991). On certain methods of improving ratio and regression estimators, Communications in Statistics-Theory and Methods, 20 (10), pp. 3325–3340.

ROBSON, D. S., (1957). Application of multivariate Polykays to the theory of unbiased ratio type estimation, Journal of American Statistical Association, 52, pp. 411–422.

SHARMA, P., SINGH, R., (2013). Improved Estimators for Simple random sampling and Stratified random sampling Under Second order of Approximation. Statistics In Transition- new series, 14 (3), pp. 379–390.

SINGH H. P., TAILOR, R., (2003). Use of known correlation coefficient in estimating the finite population means, Statistics in Transition, 6 (4), pp. 555–560.

SINGH, D., CHAUDHARY, F. S., (1986). Theory and Analysis of Sample Survey Designs, 1 ed., New Age International Publisher, India.

SINGH, H. P., TAILOR, R., KAKRAN, M., (2004). Improved estimators of population mean using power transformation, Journal of the Indian Society of Agricultural Statistics, 58 (2), pp. 223–230.

SINGH, H. P., TAILOR, R., (2005). Estimation of finite population mean with known coefficient of variation of an auxiliary, STATISTICA, anno LXV, 3, pp. 301–311.

SISODIA, B. V. S., DWIVEDI, V. K., (1981). A modified ratio estimator using coefficient of variation of auxiliary variable, Journal of the Indian Society of Agricultural Statistics, 33 (1), pp. 13–18.

SUBRAMANI, J., KUMARAPANDIYAN, G., (2012). A class of modified ratio estimators using deciles of an auxiliary variable, International Journal of Statistical Application, 2, pp. 101–107.

SUBZAR, M., ABID, M., MAQBOOL, S., RAJA, T. A., SHABEER, M., LONE, B. A., (2017). A Class of Improved Ratio Estimators for Population Mean using Conventional Location parameters, International Journal of Modern Mathematical Sciences, 15 (2), pp. 187–205.

SUBZAR, M., MAQBOOL, S., RAJA, T. A., SHABEER, M., (2017). A New Ratio Estimators for estimation of Population mean using Conventional Location parameters, World Applied Sciences Journal, 35 (3), pp. 377–384.

SUBZAR, M., RAJA, T. A., MAQBOOL, S., NAZIR, N., (2016). New Alternative to Ratio Estimator of Population Mean, International Journal of Agricultural Statistical Sciences, 12 (1), pp. 221–225.

UPADHYAYA, L. N., SINGH, H., (1999). Use of transformed auxiliary variable in estimating the finite population mean, Biometrical Journal, 41 (5), pp. 627–636.

WOLTER, K. M., (1985). Introduction to Variance Estimation, Springer-Verlag.

YAN, Z., TIAN, B., (2010). Ratio method to the mean estimation using coefficient of skewness of auxiliary variable, Information Computing and Applications, 106, pp. 103–110.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 239–258, DOI 10.21307/stattrans-2018-014

MODIFIED RECURSIVE BAYESIAN ALGORITHM FOR ESTIMATING TIME-VARYING PARAMETERS IN DYNAMIC LINEAR MODELS

O. Olawale Awe[1], **A. Adedayo Adepoju**[2]

ABSTRACT

Estimation in Dynamic Linear Models (DLMs) with Fixed Parameters (FPs) has been faced with considerable limitations due to its inability to capture the dynamics of most time-varying phenomena in econometric studies. An attempt to address this limitation resulted in the use of Recursive Bayesian Algorithms (RBAs) which is also affected by increased computational problems in estimating the Evolution Variance (EV) of the time-varying parameters. In this paper, we propose a modified RBA for estimating TVPs in DLMs with reduced computational challenges.

Key words: discounted variance, dynamic models, granularity range, estimation algorithm.

1. Introduction

Generally speaking, a model is dynamic each time the variables (or parameters) are indexed by time or appear with different time lags (Ravines et al., 2006). In recent times, estimation of time-varying parameters in econometric models has become more relevant especially as the length of the observed time series increases and the series itself is subject to changes in the dynamic structure. Particular examples can be found in world economic time series where key monthly, quarterly or annual economic indicators are commonly available from the 1950s and cover periods of different economic conditions. For example, the periods of strong economic growth in the 1950s and 1960s, periods with oil crises in the 1970s, periods of major monetary policy changes in the 1980s, rapid changes of financial markets in the 1990s and the collapse of the financial and banking systems more recently (Doh and Connolly, 2013). Although, not all economic structures are subject to changes due to these developments, it is expected that the dynamic properties of longer time series require parameters that are allowed to change over time. Models with fixed parameters have been found to perform poorly for analysis of these kinds of data because basic econometric time series analysis lies in the possibility of finding a reasonable regularity in the phenomenon under study (Petris, 2010). In a dynamic economy, for instance, the relations between economic agents are subject to change. As the

[1]Department of Mathematical Sciences, Anchor University Lagos, Nigeria. E-mail: oawe@aul.edu.ng
[2]Department of Statistics, University of Ibadan, Nigeria. E-mail: pojuday@yahoo.com

knowledge of production techniques improved, as the means of transportation allow for long-distance trade, and as the society changed its preferences for certain goods or services, the structure of the economy varies accordingly. It is also natural to think that as time changes, information decays thereby necessitating the need for discounting the variance of the underlined evolution of the dynamic parameter as adopted in this work.

In a surprisingly short period of time, Markov chain Monte Carlo (MCMC) algorithms, especially the Metropolis-Hastings algorithm (Hastings, 1970) and the Gibbs sampling algorithm (Geman and Geman, 1984; Gelfand et al., 1990) have emerged as extremely popular tools for the analysis of complex statistical and econometric models. Bayesian analysis requires the evaluation of complex and often high-dimensional integrals in order to obtain posterior distributions for the unobserved quantities of interest in the model. In many such settings, alternative methodologies such as asymptotic approximation and non-recursive Monte Carlo algorithms are either infeasible or fail to provide sufficiently accurate results. Properly defined and implemented, MCMC methods enable the user to successively sample values from a Markov chain process. Important features of MCMC methods that enhance their applicability include their ability to reduce complex multidimensional problems to a sequence of much lower-dimensional ones. While MCMC algorithms allow an enormous expansion of the class of candidate models for a given data, they also suffer from a well-known and potentially serious problem: it is often difficult to decide when it is safe to terminate them and conclude their "convergence" (Zellner, 2009).

The algorithms for recursive estimation and Kalman filtering are being used increasingly in applied econometrics, but econometricians have been slower than other statisticians to explore them (Pollock, 2003). In recursive estimation, the knowledge about the parameters of a model is updated continuously as new measurements are collected. It is suitable in problems where the parameters have dynamic properties that make them change with time. Several measurements $y_1, y_2, ..., y_n$ are considered alongside their joint probability density function $f(\theta, y_1, y_2, y_3, ..., y_n)$. New measurements are received and estimation done one at a time. After measuring y_1, we construct an estimate $\hat{\theta}_i$, when y_{i+1} is received again, parameter $\hat{\theta}_i$ would be updated. This process continues recursively. To initiate the recursion, we need an initial estimate of the parameter θ and its variance-covariance matrix.

Another reason for the burgeoning popularity of the recursive approach in econometrics is the increased importance of numerical simulations in statistics and econometrics, hence, most computational algorithms rely on recursive methods . A significant breakthrough in the application of recursive methods in econometrics was achieved by several researchers including Cooley and Prescott (1973, 1976); Bertsekas (1976); Spear and Srivastava (1987); Blanchard and Fischer (1989); Abreu et al. (1990), Ng and Young (1990); Pollock (2003); Young (2011). Although, economic theory rarely provides a useful guide to distinguish between fixed and time-varying parameters, estimation of Dynamic Linear Models (DLMs) with Fixed Parameters (FPs) has been faced with considerable limitations due to its inability

to capture the dynamics of most time-varying phenomena in econometric studies. This is because the classical linear regression model with fixed parameters presumes that the relationship between the explanatory and the explained variable remains constant through the estimation period. However, there are situations when this kind of assumption becomes unreasonable and totally non-implementable because assuming time-invariant parameters and variances turn out to be quite restrictive in capturing the evolution dynamics of most economic time series. For instance, business cycle dynamics and monetary policy in United States and other major economies of the world has changed substantially over the post-war period. In addition, the introduction of time-varying parameters in linear models can lead to different levels of complexities including the fact that they gain new evolution variance parameters which are also time-varying and, in turn, need to be estimated. In literature, the choice of the evolution variance (say Ω_t) has been found to be complex and usually difficult to characterize because of a number of practical problems associated with it which includes:

1. it varies with the measurement scale of regressor variables as specified in the observation equation of DLM.
2. it can be ambiguous i.e there may not be an optimal value of Ω_t suitable for all times.
3. it is often grossly mis-specified because most modellers have great difficulty in directly quantifying its variance and covariance elements.
4. the predictive performance of the dynamic linear model depends on the choice of the evolution variance Ω_t (West and Harrison, 1997).

An attempt to address these problems resulted in the use of Recursive Bayesian Algorithms (RBAs) which is also affected by increased computational problems in estimating the Evolution Variance (Ω_t) of the Time-Varying Parameters (TVPs). Consequently, researchers require a better way of structuring the evolution variance which previous studies have failed to effectively address. The aim of this study, therefore, was to modify an existing RBA of Fúquene et al. (2015) for estimating TVPs in DLMs. Discounting is proposed as an alternative way of coping with the system evolution variance of economic series in order to portray a clearer picture of the volatility of the parameters in the model under study over time.

The proposed recursive Bayesian estimation algorithm will be useful for proper choice of discount values to represent the evolution variance which is inevitable in order to address problems (2) and (3) above with reduced computational challenges.

2. Dynamic Linear Models with Time-Varying Parameters

It has been argued severally in literature that the parameters in econometric models cannot, in general, be expected to remain constant and hence Time-Varying Parameter (TVP) models should be considered in almost all circumstances (Soloviev

et al., 2011; Primiceri, 2005; Doh and Connolly, 2013). The difficulty in estimating such models is however often exacerbated by the fact that the econometrician would have only some idea regarding the most likely value that a parameter may assume (as indicated by, say, the Ordinary Least Squares (OLS) and maximum likelihood estimators); with a range of uncertainty surrounding this nominal value and consequently, misleading policy prescriptions are likely to arise from a straightforward optimization exercise based on such a set of nominal values especially in the presence of structural breaks in the underlying economic, technological, behavioural and institutional patterns. However, because discontinuities are a crucial feature of modern economic systems, there is the need to consider models with time-varying parameters. According to literature, such TVP models can be classified into three types:
First, the parameters can vary across subsets of observations within the sample but be non-stochastic. Examples of such models include the general systematically varying parameter model of Belsley and Kuti (1973) and a variety of switching regression models with either known joint points (see McZgee and Carleton (1970); Hinkley (1971); Goldfeld and Quandt (1973)) or unknown joint points of Gallant and Fuller (1973). A second class of models is where the parameters are stochastic, and are assumed to be generated by a stationary stochastic process. Examples of such models include the pure random coefficient model of Harvey and Phillips (1982) which includes the adaptive and varying-parameter regression models of Cooley and Prescott (1973) and the stochastically convergent parameter model of Rothenberg (1973). Finally, the third class of models consists of those where the stochastic parameters are generated by a process that is not stationary. These include the mixed estimation model of Cooper (1972), the Kalman filter model of Athans (1974), the stochastic variations model of Cooley and Prescott (1976), the systematically varying parameter model of Kalaba and Tesfatsion (1980) which was then extended to the flexible least squares (FLS) approach Kalaba and Tesfatsion (1988), the recursive and optimal control model of Ng and Young (1990) which have gained tremendous popularity in literature and become more relevant in recent times. Some of the rationale behind time-varying parameter models are documented in Sarris (1973). The archetypical (existing) dynamic linear model in literature with fixed variances has the following general form:

$$y_t = F'\beta_t + v_t \qquad v_t \sim N(0,V) \tag{1}$$

$$\beta_t = G\beta_{t-1} + w_t \qquad w_t \sim N_p(0,\Omega) \tag{2}$$

$$\beta_0 \sim N_p(m_0, C_0) \tag{3}$$

where y_t is a vector of dimension $m \times 1$

Equation (1) is known as the observation equation while equation (2) is a first order Markov process called the evolution equation.

The matrices F, V, G and Ω are known as the system matrices and contain non-random elements. If they do not depend deterministically on t, the model is time invariant, otherwise it is time varying. The initial state distribution is assumed to be Normally distributed with parameters m_0 and C_0 as shown in (3) where $E(v_t\beta_t') = 0$,$E(w_t\beta_t') = 0$ for $t = 1,...,T$.

The dynamic linear model with state space approach offers attractive features with respect to their generality, flexibility and transparency. The lack of publicly available software to estimate these models has been the main reason why only relatively few economic and finance related problems have been analyzed with dynamic linear models so far. Basically, the estimation of DLM involves three stages: prediction, filtering and smoothing. Prediction has to do with forecasting future values of the time-varying state parameters. Filtering makes the best estimate of the current values of the time-varying state parameter from the record of observations including the current observation. Smoothing involves making the best estimate of past values of the states given the record of observations.

3. Model Specification and Methodology

A typically difficult problem in econometrics is to formulate a stationary model that best resembles the model dictated by economic theory, but which does not pose serious problems of estimation (Chetty, 1971). This section, lays out the specified dynamic linear model proposed in this work. It also contains details of the developed recursive Bayesian algorithm. Remove (RBA) employed for the posterior estimation of the specified dynamic linear model in the presence of discounted evolution variance.

A concrete mathematical formulation of the proposed dynamic linear model specification in this work takes the form of the two equations:

$$y_t = X_t\theta_t + v_t \qquad v_t \sim N(0,\varphi), \tag{4}$$

$$\theta_t = G_t\theta_{t-1} + w_t \qquad w_t \sim N(0,\Omega_t), \tag{5}$$

$$\theta_0 \sim N(m_0, C_0).$$

where equation (4) is known as the observation equation while equation (5) is the evolution equation. G_t is a known transition matrix of order $p \times p$ that determines how the observation and evolution equations evolve in time. Since each parameter at time t only depends on results from time $t-1$, the state parameters are time-varying and constitute a Markov chain. X_t is a matrix of observed time series of known order. θ_t is the time-varying parameter associated with the predictor matrix X_t. It is assumed that information decays arithmetically through the addition of future evolution error variance which we estimate with discount values. Parameters of interest to be estimated are the time-varying parameter θ_t, the error variances φ and Ω_t, and the one-step-ahead forecasts error f_t. φ is assumed to be distributed

inverse-gamma a priori, while we estimate Ω_t via discounting method which is explained later in this section. The difference between this model and the one stated in Fuquene et al. (2013) is that the observational variance is presumed to be fixed and the evolution variance is estimated by the method of discounting unlike the use of Wishart prior which is common in literature. Also, in contrast to the Box-Jenkins methodology, which still plays an important role in time series analysis today, the specified dynamic linear model approach allows for structural analysis of univariate as well as multivariate problems without initial differencing or log transformation of the observed series. The different components of a series, such as trend and seasonality, as well as the effects of explanatory variables can be modelled explicitly. They do not have to be removed prior to the main analysis as is the case in the Box-Jenkins methodology.

3.1. Existing Recursive Bayesian Algorithm (RBA) and Gibbs Sampler for Estimating TVPs

The recursive Bayesian algorithms in literature usually takes the following form: Let $\Theta_t = [\theta_0, \theta_1, ..., \theta_t]$, θ_t is estimated from the conditional density $p(\Theta_t|y_T)$ which is denoted by

$$p(\Theta_t, y_T) = p(Y_T|\Theta_T)p(\Theta_T)$$

where $p(y_T|\Theta_T)$ and $p(\Theta_T)$ are given by

$$p(y_T|\Theta_T) = \prod_{t=1}^{T} p(y_t|\theta_t) \tag{6}$$

and

$$p(\Theta_T) = p(\theta_0)\prod_{t=1}^{T}(\theta_t|\theta_{t-1}), \tag{7}$$

where $p(y_t|\theta_t)$ and $p(\theta_t|\theta_{t-1})$ were derived from the observational and evolution equations (4) and (5) specified above to give

$$p(y_t|\theta_t) = (2\pi\sigma^2)^{-\frac{1}{2}} exp\bigg(-\frac{1}{2\sigma^2}(y_t - x_t\theta_t)^2 \bigg)$$

where $V = \sigma^2$,

$$p(\theta_t|\theta_{t-1}) = (2\pi)^{-\frac{k}{2}}|\Omega_t|^{-\frac{1}{2}} exp\bigg(-\frac{1}{2}(\theta_t - G_t\theta_{t-1})'\Omega_t^{-1}(\theta_t - G_t\theta_{t-1}) \bigg)$$

The recursive algorithm alternatively compute the densities of the current and the future parameter θ conditional on all available observations. Using the notation $y_t = y_{1:t}$, the prediction equation is given by

$$p(\theta_{t+1}) = \int p(\theta_{t+1}, \theta_t|y_t)d\theta_t \tag{8}$$

$$= \int p(\theta_{t+1}|\theta_t)p(\theta_t|y_t)d\theta_t.$$

Applying Bayes' rule, the filtering equation gives

$$p(\theta_t|y_{1:t}) = \frac{p(y_t|\theta_t)p(\theta_t|\theta_{t-1}, y_{1:t-1})}{p(y_t|y_{1:t-1})} \tag{9}$$
$$\propto p(y_t|\theta_t)p(\theta_t|\theta_{t-1}, y_{1:t-1})$$

The denominator, $p(y_t|y_{1:t-1})$ is constant relative to θ_t and thereby ignored. These recursive equations were initialized with the density of the initial parameter

$$p(\theta_1|y_0) = p(\theta_1).$$

This posterior is then used to update the prior recursively until convergence is achieved. The forward filtering step is the standard Kalman filtering analysis to give $p(\theta_t|D_t)$ at each t, for $t = 1,...,n$. The backward sampling step uses the Markov's property to sample θ_T^* from $p(\theta_T|D_T)$ and then for $t = 1,...,T-1$, sample θ_t^* from $p(\theta_t|D_t, \theta_{t+1}^*)$ in order to generate samples from the posterior parameter structure. In particular, denote

$$p(\theta_0,...,\theta_T|y_T) = \prod_{t=0}^{T} p(\theta_t|\theta_{t+1},...,\theta_T, y_T)$$

and note that, by the Markov's property,

$$p(\theta_t|\theta_{t+1},...,\theta_t, y_t) = p(\theta_t|\theta_{t+1}, y_t) \tag{10}$$

and

$$p(\theta_t|y_t) = \int p(\theta_t, \theta_{t+1}|y_t)d\theta_{t+1}$$

$$= \int p(\theta_{t+1}|y_t)p(\theta_t|\theta_{t+1}, y_t)d\theta_{t+1}$$

$$= p(\theta_t|y_t) \int \frac{p(\theta_{t+1}|y_t)(p(\theta_{t+1}|\theta_t)d\theta_{t+1}}{p(\theta_{t+1}|y_t)} \tag{11}$$

which follows again from the recursive application of Bayes' rule and Markov property of θ_t.

Since the sampling is done from $t = T$ to $t = 0$, recursively, this procedure is referred to as recursive backward sampling.

In particular, Fuquene et al. (2013) proposed a dynamic linear model which is specified by a normal prior distribution for the p-dimensional state vector for macroeconomic modeling with prior θ_0) as follows: $\theta_0 \sim N_p(m_0, C_0)$ with the set of equations

$$y_t = F_t\theta_t + v_t, v_t \sim N_m(0, V_t) \tag{12}$$

$$\theta_t = G_t\theta_t + v_t, w_t \sim N_m(0, W_t) \tag{13}$$

with $t = 1 : T$ where F_t and G_t are known matrices of order $p \times p$ and $m \times p$, respectively. Let $\theta \sim Student - t(\mu, \tau, \nu)$ where ν is the degree of freedom, μ and τ are the location and scale parameters of the student-t density respectively. Then,

$$\pi(\theta|\tau^2) = \frac{k_1}{\tau}\left(1 + \frac{1}{\nu}\left(\frac{\theta - \mu}{\tau}\right)^2\right)^{\frac{-\nu+1}{2}} \tag{14}$$

where $\nu > 0$, $-\infty < \mu < \infty$, $-\infty < \theta < \infty$, and
$k1 = \Gamma\frac{\frac{(\nu+1)}{2}}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}}$ we have that $\pi(\theta) = \int_0^\infty \pi(\theta|\tau^2)\pi(\tau^2)d\tau^2$ Let $W_{t,i}$ denote the i^th diagonal element of the evolution variance $W_{t,i}$, $i = 1, ..., n$, the observation and evolution variances were given as $V_t^{-1} = \lambda_y w_y, t$
and $W_{t,i}^{-1} = \lambda\theta, iw\theta, t_i$

In order to obtain posterior inference on the time-varying parameter, θ_t, they used the recursive Forward Filtering Backward Sampling (FFBS) algorithm which proceeds as follows

1. Use the Kalman filter equations for (14) above.
   Let m_0 C_0 be known with $(\theta_0|D_0) \sim N(m_0, C_0)$ and $\theta_t|y_{1:t-1} \sim N(m_{t-1}, C_{t-1})$,
   The one-step predictive distribution of θ_t given $y_{1:t-1}$ is Gaussian i.e $\theta_t|D_{t-1} \sim N(a_t, R_t)$ with parameter $a_t = G_t m_{t-1}$, $R_t = G_t C_{t-1} G_t'$.
   The one step ahead predictive distribution of y_t given $y_{1:t-1}$ is Normally distributed as $(y_t|D_{t-1}) \sim N(f_t, Q_t)$ with parameters $f_t = F_t' a_t$, $Q_t = F_t' R_t F_t + V_t$.
   The filtering distribution of θ_t given $y_{1:t-1}$ is $(\theta_t|D_t) \sim N(m_t, C_t)$ with parameters $m_t = a_t + A_t e_t$, $C_t = R_t - A_t Q_t A_t'$ where $A_t = R_t F_t Q_t^{-1}$ and $e_t = y_t - f_t$.

2. At time $t = T$, sample θ_T from $N(\theta_T|m_t, C_T)$

3. For $t = T - 1$, sample θ_t from $N(\theta_t|m_t^*, C_t^*)$ where $m_t^* = m_t + b_t(\theta_{t+1} - a_{t+1})$ $C_t^* = C_t - b_t R_{t+1} b_t'$ where $b_t = C_t G_{t+1} R_{t+1}^{-1}$

This algorithm does not specify a block for the evolution variance of the time-varying parameter θ_t which is often difficult to characterize. Additionally, the algorithm presented in this work specifies a sub-algorithm for optimal selection of Average Granularity Range (AGR) of the discounting parameter λ which also plays an important role in determining convergence of the parameters. First, the observational variance φ, was specified as constant and estimated via Gibbs sampling as presented in the next section.

3.2. Recursive Estimation of Time-Varying Parameters in the Presence of Discounted Evolution Variance

In this section, we propose an algorithm to estimate the time-varying parameters in dynamic linear models in the presence of discounted evolution variance. This approach makes use of the Recursive Forward Filtering Backward Sampling algorithm within the Kalman filter framework to improve the efficiency of the adapted Gibbs

sampler by discounting the evolution variance. The main idea of this procedure is to make use of Markov's property of the specified evolution equation so that

$$P(\theta_t|\theta_{t+1}, D_t) = P(\theta_{t+1}|\theta_t, D_t) \tag{15}$$

where θ_t denotes the time-varying parameters at time t and $D_t = (y_1, ..., y_t, x_1, ..., x_t)$. Due to the Markovian structure of the time-varying parameter θ_t, it is estimated by computing the predictive and filtering distributions of θ_t recursively starting from the prior $\theta_0 \sim N(m_0, C_0)$. This recursive method allows us to draw the parameter vectors jointly. Consider a vector of unknown time-varying slope parameters $\theta_t = (\theta_1, ..., \theta_p)$, the Gibbs sampling algorithm employed proceeds by sampling recursively the conditional posterior distribution where the most recent values of the conditioning parameters are used. Following the Bayesian paradigm, the specification of the model is complete only after specifying the prior distribution of all the unknown quantities of interest in the model. We assign a distribution to θ_t at time t=0, conditional on all the information available before any observation is made. Let D_0 be the set containing all this information, then the prior distribution is $\theta_0|D_0 \sim N(m_0, C_0)$ where m_0 and C_0 are known vector and matrix respectively. Next, an update is made for θ_1 and D_0 which is also normally distributed. Based on this update, the one step-ahead forecast follows from the conditional distribution $y|\theta_0, D_0$. Once the value of y_1 at time $t = 1$ is known, the posterior distribution of θ_1 is obtained recognizing that the information available at time $t = 1$ is $D_1 = y_1, D_0$. The inference is made in this recursive fashion for every time t. The Kalman filter was used to calculate the mean and variance of the parameter θ_t, given the observations D_t. It is a recursive algorithm because the current best estimate is updated whenever a new observation is obtained. This recursive Bayesian technique of model estimation can be stated in form of prediction, filtering and update equations. The prediction and update step requires a few basic calculations of which only the conditional means and variances of the filtering and prediction density is stored in each step of the iteration.
To describe the filtering procedure, let

$$m_t = E(\theta_t|D_t) \tag{16}$$

be the optimal estimator of θ_t based on D_t and let

$$C_t = E((\theta_t - m_t)(\theta_t - m_t)'|D_t) \tag{17}$$

be the mean square error matrix of m_t. Let $\theta_{t-1}|y_{1:t-1} \sim N(m_{t-1}, C_{t-1})$, where $y_{1:t-1}$ denote all observations up to time $t-1$. Then the one-step-ahead predictive density $\theta_t|y_{1:t-1}$ is Gaussian with parameters:

$$E(\theta_t|y_{1:t-1}) = m_{t-1} \equiv A_t(say) \tag{18}$$

$$Var(\theta_t|y_{1:t-1}) = C_{t-1} + \Omega_t \equiv R_t(say) \tag{19}$$

The one-step-ahead predictive density of $y_t|y_{1:t-1}$ is Gaussian with parameters:

$$f_t = E(y_t|y_{1:t-1}) = X_t A_t \tag{20}$$

$$Q_t = Var(y_t|y_{1:t-1}) = X_t R_t X_t^{'} + V \tag{21}$$

The filtering density of θ_t given $y_{1:t}$ is Gaussian with parameters:

$$m_t = E(\theta_t|y_{1:t}) = A_t + R_t X_t^{'} Q_t^{-1} e_t \tag{22}$$

$$C_t = Var(\theta_t|y_{1:t}) = R_t - R_t X_t^{'} Q_t^{-1} X_t R_t^{\prime} \tag{23}$$

where $e_t = y_t - f_t$ is the forecast error.

3.2.1 Posterior Estimation of Unknown Observational Variance (φ) with Independent Priors

In the simulation excercise for estimating the static observational variance, φ, the following Gibbs sampler of Nakajima et al. (2011) was adopted with slight modifications: Consider the linear equation which is the observational equation specified in (4) above

$$y_t = X_t \theta_t + v_t, v_t \sim N(0, \varphi), \tag{24}$$

let $\varphi \equiv \sigma^2$ and $\theta_t = \theta$

and assume a normal prior for the parameter θ and inverse gamma prior for the parameter σ^2, to sample from $\varphi|\theta$ we impose a gamma prior on φ^{-1} and derive the posterior hyperparameters. Let $\varphi^{-1} \sim Gamma(a_0, b_0)$, then

$$\varphi^{-1}|\theta \sim Gamma(a_0 + \frac{T}{2}, b_0 + \frac{1}{2}\sum_{t=1}^{T}(y_t - X_t\theta_t)^2)$$

We start with

$$p(y|\theta, X) = (2\pi)^{-\frac{n}{2}} exp(-\frac{1}{2\sigma^2}(y - X\theta)'(y - X\theta)) \tag{25}$$

The priors are given as follows:

$$p(\theta, \varphi) = p(\theta)p(\varphi)$$

where

$$\theta \sim N(\mu_0, \varphi_0) \tag{26}$$

and

$$\varphi \sim IG(\nu_0, \tau_0) \tag{27}$$

μ_0 is the prior mean for θ and φ_0 is the prior variance- covariance matrix for θ

with

$$E(\varphi) = \frac{\tau_0}{\nu_0 - 1} \tag{28}$$

$$V(\varphi) = \frac{\tau_0^2}{(\nu_0 - 1)^2(\nu_0 - 2)} \tag{29}$$

We chose the form given in Gelman (2004) where ν_0 and τ_0 are the shape and scale parameters respectively. Using Bayes rule to combine the priors (26) and (27) above with the likelihood and dropping all unrelated terms to the parameters of interest yields the following posterior kernels:

$$p(\theta, \varphi | y, X) \propto (\sigma^2)^{\frac{-n-2\nu_0-2}{2}} exp(-\frac{1}{2\sigma^2}(2\tau_0)) \times exp\left(-\frac{1}{2}(\frac{1}{\sigma^2}(y - X\theta)'(y - X\theta) + (\theta - \mu_0)'\varphi_0^{-1}(\theta - \mu_0))\right) \tag{30}$$

First, we find the posterior density of θ, conditional on φ while treating φ as a constant.
This leaves us with the posterior kernel:

$$p(\theta | \varphi, y, X) \propto$$

$$exp(-\frac{1}{2}(\frac{1}{\varphi}(y - X\theta)'(y - X\theta) + (\theta - \mu_0)'(\varphi_0)^{-1}(\theta - \mu_0))). \tag{31}$$

Transformations

Let

$$\varphi_1 = (\varphi_0^{-1} + \frac{1}{\varphi}X'X)^{-1}$$

and

$$\mu_1 = \varphi_1(\varphi_0^{-1}\mu_0 + \frac{1}{\varphi}X'Xb) = \varphi_1(\varphi_0^{-1}\mu_0 + \frac{1}{\varphi}X'y)$$

Then from (3.34),

$$\frac{1}{\varphi}(y - X\theta)'(y - X\theta) + (\theta - \mu_0)'\varphi_0^{-1}(\theta - \mu_0)$$

$$
\begin{aligned}
&= \frac{1}{\varphi}y'y + \theta'\frac{1}{\varphi}X'X\theta - \frac{1}{\varphi}y'X\theta - \theta'\frac{1}{\varphi}X'y + \theta'\varphi_0^{-1}\theta \\
&- \mu_0'\varphi_0^{-1}\theta - \theta'\varphi_0^{-1}\mu_0 + \mu_0'\varphi_0^{-1}\mu_0 \\
&= \theta'(\varphi_0^{-1} + \frac{1}{\varphi}X'X)\theta - \theta'(\varphi_0^{-1}\mu_0 + \frac{1}{\varphi}X'y) \\
&- (\mu_0'\varphi_0^{-1} + \frac{1}{\varphi}y'X)\theta + \frac{1}{\varphi}y'y + \mu_0'\varphi_0^{-1}\mu_0
\end{aligned}
$$

$$
\begin{aligned}
&= \theta'\varphi_1^{-1}\theta - \theta'\varphi_1^{-1}\mu_1 - \mu_1'\varphi_1^{-1}\theta \\
&+ \mu_1'\varphi_1^{-1}\mu_1 - \mu_1'\varphi_1^{-1}\mu_1 + \frac{1}{\varphi}y'y + \mu_0'\varphi_0^{-1}\mu_0 \\
&= (\theta - \mu_1)'\varphi_1^{-1}(\theta - \mu_1) - \mu_1'\varphi_1^{-1}\mu_1\varphi_1^{-1}\mu_1 + \frac{1}{\varphi}y'y \\
&+ \mu_0'\varphi_0^{-1}\mu_0.
\end{aligned}
$$

Therefore, the conditional posterior kernel in (31) above can be written as :

$$p(\theta|\varphi, y, X) \propto$$

$$exp(-\frac{1}{2}(\theta - \mu_1)'\varphi_1^{-1}(\theta - \mu_1)exp(-\frac{1}{2}(\frac{1}{\varphi}y'y + \mu_0'\varphi_0^{-1}\mu_0 - \mu_1'\varphi_1^{-1}\mu_1)) \tag{32}$$

Since none of the terms in the second exponent include θ, we simplify the full conditional distribution in (32) to

$$p(\theta|\varphi, y, X) \propto exp(-\frac{1}{2}(\theta - \mu_1)'\varphi_1^{-1}(\theta - \mu_1)) \tag{33}$$

Therefore, we have again, the kernel of a multivariate normal density, and we can say that

$$\theta|\varphi, y, X \sim N(\mu_1, \varphi_1)$$

where

$$\varphi_1 = (\varphi_0^{-1} + \frac{1}{\varphi}X'X)^{-1}$$

and

$$\mu_1 = \varphi_1(\varphi_0^{-1}\mu_0 + \frac{1}{\varphi}X'y)$$

to sample from.

Posterior Inference on φ

In order to derive the conditional posterior density for φ, we return to our original expression for the joint posterior given in (30). Ignoring terms that are not related

to φ, we have :

$$p(\varphi|\theta,y,X) \propto (\varphi)^{\frac{-n-2\nu_0-2}{2}} exp(-\frac{1}{2\varphi}(2\tau_0+(y-X\theta)'(y-X\theta))) \tag{34}$$

Comparing this expression with the kernel of the inverse gamma prior specified in (29) above, we have the kernel of another inverse gamma density: Hence

$$\varphi|\theta,y,X \sim IG(\nu_1,\tau_1) \tag{35}$$

where

$$\nu_1 = \frac{2\nu_0+n}{2}$$

and

$$\tau_1 = \frac{2\tau_0+(y-X\theta)'(y-X\theta)}{2}$$

3.3. Estimation of Evolution Variance (Ω_t) with Discount Values

Consider the evolution equation in (5) above,

$$\theta_t = G_t\theta_{t-1}+w_t, w_t \sim N(0,\Omega_t) \tag{36}$$

where Ω_t is the evolution variance and other parameters are as defined earlier. Let

$$\begin{aligned} V(\theta_{t-1}|D_{t-1}) &= V(G_t\theta_{t-1}|D_{t-1}) \\ &= G_tC_{t-1}G_t' \\ &= C_{t-1} \end{aligned}$$

so that

$$V(\theta_t|D_{t-1}) = C_{t-1}+\Omega_t$$

The prior distribution for θ_{t-1} is

$$\theta_{t-1}|D_{t-1} \sim N(m_{t-1},C_{t-1})$$

where $D_{t-1} = (y_1,y_2,...,y_{t-1})$ and the prior distribution for θ_t is

$$\theta_t|D_{t-1} \sim N(m_{t-1},Q_t)$$

where

$$Q_t = C_{t-1}+\Omega_t$$

Therefore,

$$\Omega_t = Q_t - C_{t-1} \tag{37}$$

We introduce the discount factor as a quantity λ such that

$$Q_t = C_{t-1}/\lambda \tag{38}$$

can be interpreted as the percentage of information that passes from time $t-1$ to t.

Therefore, we select the discounting grid $\lambda \in [0.01, 0.99]$. We next develop a sub-algorithm to select optimal granularities of the discount values λ which enable us to conlude the convergence of the model.

4. Parsimonious Model Selection Algorithm (PMSA) for Optimal Model and Discount Value Selection

Since the choice of the evolution variance determines the forcasting performance of DLM, a sub-algorithm for optimal discount value selection with Mean Squared Prediction Error was developed as follows:

1. Init: i=0
2. Let $\lambda_i \in [0.01..0.99]$
3. Compute Ω_t in the DLM with λ_i
4. Estimate one-step ahead predictive density of the specified Bayesian DLM
5. Compute concurrent MSPE of DLM in 3 and cross-validate with the discount value of Ω_{ti}
6. Set $i = i+1$
7. Is the current MSPE lower than the previous one?
8. If No, Go To 6
9. If Yes, Go To 10
10. Stop: Pick the current discount value and DLM as the best.

4.1. Convergence Diagnostics

The convergence diagnostics of Geweke (1993) was used to compare values in the early part of the Markov chain to those in the latter part of the chain in order to detect failure of convergence. The statistic is constructed as follows: Two sub-sequences of the Markov chain θ are taken out, with $\theta_1^t : t = 1, ..., n_1$ and $\theta_2^t : t = n_a, ..., n$ where $1 \leq n_1 \leq n_a < n$.

Let $n_2 = n - n_a + 1$ and define $\bar{\theta}_1 = \frac{1}{n_1}\sum_{t=1}^{n_1}\theta^t$ and $\bar{\theta}_2 = \frac{1}{n_2}\sum_{t=na}^{n}\theta^t$. Geweke test statistics was used to test whether the mean estimates have converged by comparing means from the early and latter part of the Markov chain. Assuming the ratios

$\frac{n_1}{n}$ and $\frac{n_2}{n}$ are fixed, $\frac{n_1+n_2}{n} < 1$, then the following statistic converges to standard normal distribution as n approaches ∞ we have

$$Z_n = \frac{\bar{\theta}_1 - \bar{\theta}_2}{\sqrt{\hat{s_1}(\theta)/n_1 + \hat{s_2}(\theta)/n_2}} \tag{39}$$

where $\hat{s_1}(\theta)$ and $\hat{s_2}(\theta)$ respresent spectral density estimates at zero frequencies. This is a two-sided test and large absolute value $Z-score$ indicates rejection of the null hypothesis of non-stationarity. Effective sample size relates to autocorrelation and measures mixing of the Markov chain. Most often, much discrepancy between the effective sample size and the simulation sample size indicates poor mixing. Effective Sample Size (ESS) is defined as

$$ESS = \frac{n}{\eta} = \frac{n}{1+2\sum_{k=1}^{\infty}\rho_k(\theta)} \tag{40}$$

where n is the total sample size and $\rho_k(\theta)$ is the autocorrelation at lag k for θ. The quantity η is autocorrelation time. The Bayesian process for estimating it is to first find a cut off point k after which the autocorrelations are very close to zero and then sum all the ρ_k to that point. The cut-off point k is such that $\rho_k < 0.01$ or $\rho_k < 2s_k$ where s_k is the standard deviation defined as

$$s_k = 2\sqrt{(\frac{1}{n}(1+2\sum_{j=1}^{k-1}\rho_k^2(\theta)))} \tag{41}$$

In this method, the Lowest Average Granularity Range (AGR) of λ required for convergence and for minimum Mean Squared Prediction Error (MSPE) would be used to determine optimal performance of the DLMs.

4.2. The Modified Recursive Bayesian Algorithm

In summary, the modified recursive Bayesian algorithm for estimating time-varying parameter proceeds as follows:

1. Sample from $p(\theta_T|D_T)$ using the filtering density in section 3.2 . This distribution is assumed to be Normally distributed with parameter $N(h_t, H_t)$ where:

$$h_t = m_t + C_t G_t' R_{t+1}^{-1}(\theta_{t+1} - a_{t+1}) \tag{42}$$

$$H_t = C_t - C_t G_t' R_{t+1}^{-1} G_t C_t' \tag{43}$$

2. Sample from $p(\theta_{T-1}|\theta_T, D_T)$.

3. For the filtering algorithm to run, estimate φ using the Gibbs sampler in section 3.2.1 .

4. Given $(\theta_t|D_t)$, obtain $\Omega_t = C_t(1-\lambda)/\lambda$ via the discounting method in section 3.3

5. Proceed by sampling recursively in this manner for $t+1, t+2$, ...
6. Use the sub-algorithm in section 3.4 to determine AGR required for convergence and when to stop sampling.
7. Sample from $p(\theta_0, ..., \theta_T | D_T)$.
8. Starting from the final density sampled in equation (7) above, the smoothing recursion proceeds backwards in time, using the previously computed filtering and prediction densities.
9. Employ the convergence diagnostics discussed in section 3.4.1 to detect failure or otherwise of convergence of the Markov chain.
10. Use λ and minimum MSPE to assess the performance of the modified algorithm for DLM with FPs and TVPs for various sample sizes.

5. Conclusion

A sound theoretical exposition of how recursive Bayesian algorithms can be employed to model dynamic relationships over time in the presence of discounted evolution variance constituted a major portion of this paper. The modelling of change in the context of widely established concepts in econometrics was addressed by proposing a conceptually implementable Recursive Bayesian Algorithm (RBA) for estimating of time-varying slope parameters (θ_t) in dynamic linear model in the presence of discounted evolution variance. A fast and efficient sub-algorithm for optimal discount value and model selection was also proposed, to determine the average granularities of discount values required for convergence in estimation of time-varying parameters. Future studies will explore the application of this algorithm to simulated and real financial, economic and environmental time series data.

REFERENCES

ABREU, D., PEARCE, D., STACCHETTI, E., (1990). Toward a theory of discounted repeated games with imperfect monitoring. Econometrica, Journal of the Econometric Society, pp. 1041–1063.

ATHANS, M., (1974). The importance of kalman filtering methods for economic systems, In Annals of Economic and Social Measurement, Vol. 3, No. 1, pp. 49–64. NBER.

BELSLEY, D. A., KUTI, E., (1973). Time-varying parameter structures: An overview, In Annals of Economic and Social Measurement, Volume 2, number 4, pp. 375–379. NBER.

BERTSEKAS, D. P., (1976). Dynamic programing and stochastic control, Mathematics in Science and Engineering, Academic Press, New York.

BLANCHARD, O. J., FISCHER, S., (1989). Lectures on macroeconomics, MIT press.

CHETTY, V. K., (1971). Estimation of solowŠs distributed lag models, Econometrica: Journal of the Econometric Society, pp. 99–117.

COOLEY, T. F., PRESCOTT, E. C., (1973). Systematic (non-random) variation models: varying parameter regression: a theory and some applications, In Annals of Economic and Social Measurement, Vol. 2, No. 4, pp. 463–473. NBER.

COOLEY, T. F., PRESCOTT, E. C., (1976). Estimation in the presence of stochastic parameter variation, Econometrica: journal of the Econometric Society, pp. 167–184.

COOPER, R. L., (1972). The predictive performance of quarterly econometric models of the united states, In Econometric Models of Cyclical Behavior, Vol. 1 and 2, pp. 813–947. NBER.

DOH, T., CONNOLLY, M., (2013). The state space representation and estimation of a time- varying parameter VAR with stochastic volatility, Springer.

FUQUENE, J., ALVAREZ, M., PERICCHI, L., (2013). A robust Bayesian dynamic linear model to detect abrupt changes in an economic time series: The case of Puerto Rico, arXiv preprint arXiv:1303.6073.

FÚQUENE, J., M. ÁLVAREZ, PERICCHI, L. R., (2015). A robust Bayesian dynamic linear model for Latin-American economic time series: Şthe Mexico and Puerto Rico casesŤ, Latin American Economic Review 24 (1), pp. 1–17.

GALLANT, A. R., FULLER, W. A., (1973). Fitting segmented polynomial regression models whose join points have to be estimated, Journal of the American Statistical Associa- tion 68 (341), pp. 144–147.

GELFAND, A. E., HILLS, S. E., RACINE-POON, A., SMITH, A. F., (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling, Journal of the American Statistical Association 85 (412), pp. 972–985.

GELMAN, A., (2004). Parameterization and Bayesian modeling, Journal of the American Statistical Association 99 (466).

GEMAN, S., GEMAN, D., (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images, Pattern Analysis and Machine Intelligence, IEEE Trans- actions on (6), pp. 721–741.

GEWEKE, J., (1993). Bayesian treatment of the independent student-t linear model, Journal of Applied Econometrics 8(S1), pp. 19–40.

GOLDFELD, S. M., QUANDT, R. E., (1973). A Markov model for switching regressions, Journal of econometrics 1 (1), pp. 3–15.

HARVEY, A., PHILLIPS, G., (1982). The estimation of regression models with time-varying parameters, In Games, economic dynamics, and time series analysis, pp. 306–321. Springer.

HASTINGS, W. K., (1970). Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1), pp. 97–109.

HINKLEY, D. V., (1971). Inference in two-phase regression. Journal of the American Statistical Association 66 (336), 736–743.

KALABA, R., TESFATSION, L., (1980). A least-squares model specification test for a class of dynamic nonlinear economic models with systematically varying parameters, Journal of Optimization Theory and Applications 32 (4), pp. 537–567.

KALABA, R., TESFATSION, L., (1988). The flexible least squares approach to time-varying linear regression. Journal of Economic Dynamics and Control 12(1), 43–48.

MCZGEE, V. E., CARLETON, W. T., (1970). Piecewise regression. Journal of the American Statistical Association 65 (331), pp.1109–1124.

NAKAJIMA, J., KASUYA, M., WATANABE, T., (2011). Bayesian analysis of time-varying parameter vector autoregressive model for the Japanese economy and monetary policy, Journal of the Japanese and International Economies 25 (3), pp. 225–245.

NG, C., YOUNG, P. C., (1990). Recursive estimation and forecasting of non-stationary time series, Journal of Forecasting 9 (2), pp. 173–204.

PETRIS, G., (2010). An r package for dynamic linear models, Journal of Statistical Soft- ware 36(12), pp. 1–16.

POLLOCK, D., (2003). Recursive estimation in econometrics, Computational statistics data analysis 44 (1), pp. 37–75.

PRIMICERI, G. E., (2005). Time varying structural vector autoregressions and monetary policy, The Review of Economic Studies 72 (3), pp. 821–852.

RAVINES, R., SCHMIDT, A. M, MIGON, H. S, (2006). Revisiting distributed lag models through a Bayesian perspective, Applied Stochastic Models in Business and Industry 22 (2), pp. 193–210.

ROTHENBERG, T. J., (1973). Efficient estimation with a priori information, New Haven: Yale University Press.

SARRIS, A. H., (1973). A Bayesian approach to estimation of time-varying regression co-efficients, In Annals of Economic and Social Measurement, Vol. 2, No. 4, pp.497–520. NBER.

SOLOVIEV, V., SAPTSIN V., CHABANENKO, D., (2011). Markov chains application to the financial-economic time series prediction, arXiv preprint arXiv:1111.5254.

SPEAR, S. E., S. SRIVASTAVA, (1987). On repeated moral hazard with discounting, The Review of Economic Studies 54 (4), pp. 599–617.

WEST, M., HARRISON, P., (1997). Bayesian Forecasting and Dynamic Models (2nd ed.), New York: Springer-verlag.

YOUNG, P. C., (2011). Recursive estimation and time-series analysis: An introduction for the student and practitioner, Springer.

ZELLNER, A., (2009). Bayesian econometrics: past, present, and future. Advances in Econometrics 23, pp. 11–60.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 259–276, DOI 10.21307/stattrans-2018-015

A GENERALIZED EXPONENTIAL TYPE ESTIMATOR OF POPULATION MEAN IN THE PRESENCE OF NON-RESPONSE

Siraj Muneer[1], **Javid Shabbir**[2], **Alamgir Khalil**[3]

ABSTRACT

In this article, we propose a class of generalized exponential type estimators to estimate the finite population mean by using two auxiliary variables under non-response in simple random sampling. The proposed estimator under non-response in different situations has been studied and gives minimum mean square error as compared to all other considered estimators. Usual exponential ratio type estimator, exponential product type estimator and many more estimators are also identified from the proposed estimator. We use three real data sets to obtain the efficiencies of estimators.

Key words: auxiliary variables, bias, MSE, efficiency, non-response.

1. Introduction

In survey sampling, it is assumed that all the observations are correctly measured on the characteristics under study. But in practice, when we fail to collect the complete information on different variables, non-response is supposed to occur. Non-response occurs due to many reasons, which includes the lack of information provided by respondents, also some of the respondents refuse to answer the questionnaire, sometimes it is difficult to find out the respondents, etc. The common approach to overcome non-response problem is to contact the non-respondents and obtain maximum information as much as possible. Generally auxiliary information is used to increase the precision of the estimators when there exists a correlation between the study and the auxiliary variables. Ratio, product and regression estimators are good examples in this context. In daily life there are many situations when we are unable to access the complete information either on the study variable or the auxiliary variable or at the same time both on the study and the auxiliary variables. Hansen and Hurwitz (1946) were the first to suggest a non-response handling technique in mail surveys combined the advantages of mailed questionnaires and personal interviews. Later on, several authors, including Srinath (1971), selected the subsample of non-respondents, where the sub-sampling

[1]Department of Statistics, University of Peshawar, Peshawar, Pakistan.
E-mail: sirajmuneer1@gmail.com
[2]Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan.
E-mail: javidshabbir@gmail.com
[3]Department of Statistics, University of Peshawar, Peshawar, Pakistan.
E-mail: profalamgir@yahoo.com

fraction varied according to the non-response rates. Rao (1986) suggested a ratio type estimator for population mean, in the presence of non-response for two-phase sampling, when the population mean of the auxiliary variable x was unknown and non-response occurred on the auxiliary variable. Khare and Srivastava (1993, 1995, 1997, 2010) and Olkin (1958) suggested different ratio and product types estimators for estimation of population mean using the auxiliary information under non-response. Similarly, Singh and Kumar (2008) and Singh et al. (2010) have made significant contributions and proposed ratio, product and difference classes of estimators under non-response. El-Badry (1956), Bahl and Tuteja (1991), Kumar and Bhougal (2011), Muneer et al. (2017) and Ismail et al. (2011) suggested many estimators in two-phase sampling with sub-sampling of non-respondents in estimating the finite population mean. Khare and Sinha (2007, 2009, 2011) proposed some classes of estimators for estimating population mean in the presence of non-response using multi-auxiliary characters in different ways. For controlling the non-response bias and eliminating the need for call backs in survey sampling, Tabasum and Khan (2006), Shabbir and Nasir (2013) and references cited therein have discussed some good techniques and plans for the estimation of finite population mean followed the technique proposed by Hansen and Hurwitz (1946) using one or more auxiliary variables in the presence of non-response. Now, we explain the Hansen and Hurwitz (1946) strategy for non-response.

Suppose a finite population $U=(1,2,...,N)$ of size N units can be divided into two classes $N=N_1+N_2$. Let N_1 and N_2 be the number of units in the population that form the response and non-response classes respectively. We draw a sample of size n units from U by using a simple random sample without replacement (SRSWOR) sampling scheme. Let r_1 units respond and $r_2=(n-r_1)$ units do not respond in the first attempt. Let a subsample of size k_2 units be selected from r_2 non-respondents units, such that $r_2=k_2h,(h>1)$. Let y_i and (x_i,z_i) be the values of the study variable (y) and the auxiliary variables (x,z) respectively. Let $\bar{y}$ and $(\bar{x},\bar{z})$ be the sample means corresponding to population means $\bar{Y}$ and $(\bar{X},\bar{Z})$ respectively.

2. Notations and Symbols with Selected Estimators

To obtain the properties of estimators, we define the following symbols and notations.

Let $e_0^*=\left(\frac{\bar{y}^*-\bar{Y}}{\bar{Y}}\right), e_1^*=\left(\frac{\bar{x}^*-\bar{X}}{\bar{X}}\right), e_1=\left(\frac{\bar{x}-\bar{X}}{\bar{X}}\right)$, $e_2=\left(\frac{\bar{z}-\bar{Z}}{\bar{Z}}\right)$, $e_2^*=\left(\frac{\bar{z}^*-\bar{Z}}{\bar{Z}}\right)$ are the relative error terms, such that

$E(e_i^*)=0,(i=0,1,2)$, $E(e_i)=0,(i=1,2)$.

$E(e_0^{*2})=\lambda C_y^2+\theta C_{y(2)}^2=V_y^*, \quad E(e_1^{*2})=\lambda C_x^2+\theta C_{x(2)}^2=V_x^*,$

$E(e_2^{*2})=\lambda C_z^2+\theta C_{z(2)}^2=V_z^*, \quad E(e_0^*e_1^*)=\lambda C_{yx}+\theta C_{yx(2)}=V_{yx}^*,$

$E(e_0^*e_2^*)=\lambda C_{yz}+\theta C_{yz(2)}=V_{yz}^*, E(e_1^*e_2^*)=\lambda C_{xz}+\theta C_{xz(2)}=V_{xz}^*,$

$E(e_1^*e_1)=E(e_1^2)=\lambda C_x^2=V_x, E(e_0^*e_1)=E(e_0e_1)=\lambda C_{yx}=V_{yx},$

where

$C_y^2 = \frac{S_y^2}{\bar{Y}^2}$, $C_x^2 = \frac{S_x^2}{\bar{X}^2}$, $C_z^2 = \frac{S_z^2}{\bar{Z}^2}$, $C_{yx} = \rho_{yx}C_yC_x$, $C_{yz} = \rho_{yz}C_yC_z$,
$C_{xz} = \rho_{xz}C_xC_z$, $C_{y(2)}^2 = \frac{S_{y(2)}^2}{\bar{Y}^2}$, $C_{x(2)}^2 = \frac{S_{x(2)}^2}{\bar{X}^2}$, $C_{z(2)}^2 = \frac{S_{z(2)}^2}{\bar{Z}^2}$,
$C_{yx(2)} = \rho_{yx(2)}C_{y(2)}C_{x(2)}$, $C_{yz(2)} = \rho_{yz(2)}C_{y(2)}C_{z(2)}$, $C_{xz(2)} = \rho_{xz(2)}C_{x(2)}C_{z(2)}$,
$\lambda = (\frac{1-f}{n})$, $f = \frac{n}{N}$, $\theta = W_2(\frac{h-1}{n})$ and $W_2 = \frac{N_2}{N}$.

Now, we review some important estimators which are available in the literature.

1. Hansen and Hurwitz (1946) were the first who formulated an unbiased estimator of the population mean $\bar{Y}$ of the study variable Y in the presence of non-response. Initially they considered the mailed survey in the first attempt and personal interviews in the second attempt after the deadline was over. The estimator is given by

$$\bar{y}^* = \left(\frac{r_1}{n}\right)\bar{y}_{r_1} + \left(\frac{r_2}{n}\right)\bar{y}_{k_2}, \tag{1}$$

where $\bar{y}_{r_1} = \frac{1}{r_1}\sum_{i=1}^{r_1} y_i$ and $\bar{y}_{k_2} = \frac{1}{k_2}\sum_{i=1}^{k_2} y_i$.
The variance of $\bar{y}^*$, is given by

$$V(\bar{y}^*) = \bar{Y}^2 V_y^*. \tag{2}$$

2. The ratio and product estimators under non-response case.
When non-response exists on the study variable y as well as on the auxiliary variable x, the traditional ratio and product estimators for population mean $\bar{Y}$ are given by

$$\bar{y}_R^* = \bar{y}^*\left(\frac{\bar{X}}{\bar{x}^*}\right), \tag{3}$$

and

$$\bar{y}_P^* = \bar{y}^*\left(\frac{\bar{x}^*}{\bar{X}}\right), \tag{4}$$

where $\bar{y}^*$ and $\bar{x}^*$ are the Hansen and Hurwitz (1946) estimators for population means $\bar{Y}$ and $\bar{X}$ respectively and are defined by $\bar{y}^* = \left(\frac{r_1}{n}\right)\bar{y}_{r_1} + \left(\frac{r_2}{n}\right)\bar{y}_{k_2}$ and $\bar{x}^* = \left(\frac{r_1}{n}\right)\bar{x}_{r_1} + \left(\frac{r_2}{n}\right)\bar{x}_{k_2}$ with $(\bar{y}_{r_1}, \bar{x}_{r_1})$ and $(\bar{y}_{k_2}, \bar{x}_{k_2})$ are the sample means of (y,x) based on the samples of r_1 and k_2 units respectively.

The $MSEs$ of $\bar{y}_R^*$ and $\bar{y}_P^*$, to the first order of approximation, are given by

$$MSE(\bar{y}_R^*) \cong \bar{Y}^2(V_y^* + V_x^* - 2V_{yx}^*), \tag{5}$$

and

$$MSE(\bar{y}_P^*) \cong \bar{Y}^2(V_y^* + V_x^* + 2V_{yx}^*). \tag{6}$$

We observed that $\bar{y}_R^*$ and $\bar{y}_P^*$ perform better than $\bar{y}^*$, if $V_{yx}^* > \frac{1}{2}V_x^*$ and $V_{yx}^* < -\frac{1}{2}V_x^*$ respectively.

3. Rao (1986) suggested ratio and product estimators under no-response. When non-response exists only on the study variable y, while the complete information on the auxiliary variable x is available, the ratio and product estimators, are given by

$$\bar{y}^*_{Rao(R)} = \bar{y}^* \left(\frac{\bar{X}}{\bar{x}} \right), \tag{7}$$

and

$$\bar{y}^*_{Rao(P)} = \bar{y}^* \left(\frac{\bar{x}}{\bar{X}} \right), \tag{8}$$

where $\bar{x}$ is the sample mean $\bar{X}$ based on complete information, and $\bar{X}$ is the population mean of the auxiliary variable.

The $MSEs$ of $\bar{y}^*_{Rao(R)}$ and $\bar{y}^*_{Rao(P)}$, to the first order of approximation, are given by

$$MSE(\bar{y}^*_{Rao(R)}) \cong \bar{Y}^2(V^*_y + V_x - 2V_{yx}), \tag{9}$$

and

$$MSE(\bar{y}^*_{Rao(P)}) \cong \bar{Y}^2(V^*_y + V_x + 2V_{yx}). \tag{10}$$

Note that $\bar{y}^*_{Rao(R)}$ and $\bar{y}^*_{Rao(P)}$ perform better then $\bar{y}^*$ if $V_{yx} > \frac{1}{2}V_x$ and $V_{yx} < -\frac{1}{2}V_x$ respectively.

4. Bahl and Tuteja (1991) exponential ratio and product type estimators for population mean $\bar{Y}$, when non-response exists on the study variable y as well as on the auxiliary variable x as:

$$\bar{y}^*_{exp(R)} = \bar{y}^* \exp\left(\frac{\bar{X} - \bar{x}^*}{\bar{X} + \bar{x}^*} \right), \tag{11}$$

and

$$\bar{y}^*_{exp(P)} = \bar{y}^* \exp\left(\frac{\bar{x}^* - \bar{X}}{\bar{x}^* + \bar{X}} \right). \tag{12}$$

The $MSEs$ of $\bar{y}^*_{exp(R)}$ and $\bar{y}^*_{exp(P)}$, to the first order of approximation, are given by

$$MSE(\bar{y}^*_{exp(R)}) \cong \bar{Y}^2 \left(V^*_y + \frac{1}{4}V^*_x - V^*_{yx} \right), \tag{13}$$

and

$$MSE(\bar{y}^*_{exp(P)}) \cong \bar{Y}^2 \left(V^*_y + \frac{1}{4}V^*_x + V^*_{yx} \right). \tag{14}$$

Both the estimators $\bar{y}^*_{exp(R)}$ and $\bar{y}^*_{exp(P)}$ are more efficient than $\bar{y}^*$ if $V^*_{yx} > \frac{1}{4}V^*_x$ and $V^*_{yx} < -\frac{1}{4}V^*_x$ respectively.

5. Singh and Kumar (2008) suggested ratio, product and difference type estima-

tors in the case of non-response. They considered the situation in which the population mean of the auxiliary variable x is known, but some units fail to provide information on the study variable y and the auxiliary variable x. The estimator is given by:

$$\bar{y}^*_{SK(R1)} = \bar{y}^* \left(\frac{\bar{X}}{\bar{x}^*}\right)\left(\frac{\bar{X}}{\bar{x}}\right), \tag{15}$$

where $\bar{x}^*$ and $\bar{x}$, both are unbiased estimators of the population mean $\bar{X}$ of the auxiliary variable x.
The *MSE* of $\bar{y}^*_{SK(R1)}$, to the first order of approximation, is given by

$$MSE(\bar{y}^*_{SK(R1)}) \cong \bar{Y}^2 \left(V_y^* + V_x^* + 3V_x - 2(V_{yx}^* + V_{yx})\right). \tag{16}$$

Note that $\bar{y}^*_{SK(R1)}$ performs better than $\bar{y}^*$ if $(V_{yx}^* + V_{yx}) > \frac{1}{2}(V_x^* + V_x)$. The product estimator of the above mentioned situation is

$$\bar{y}^*_{SK(P)} = \bar{y}^* \left(\frac{\bar{x}^*}{\bar{X}}\right)\left(\frac{\bar{x}}{\bar{X}}\right). \tag{17}$$

The *MSE* of $\bar{y}^*_{SK(P)}$, to the first order of approximation, is given by

$$MSE(\bar{y}_{SK(P)}) \cong \bar{Y}^2 \left(V_y^* + V_x^* + 3V_x + 2(V_{yx}^* + V_{yx})\right). \tag{18}$$

Note that *MSE* of $\bar{y}^*_{SK(P)}$, is smaller than $\bar{y}^*$ if $V_{yx}^* + V_{yx} < -\frac{1}{2}(V_x^* + 3V_x)$.
Singh and Kumar (2008) also suggested the generalized ratio-type estimator of the above mentioned situations as

$$\bar{y}^*_{SK(R2)} = \bar{y}^* \left(\frac{\bar{X}}{\bar{x}^*}\right)^{\alpha_1}\left(\frac{\bar{X}}{\bar{x}}\right)^{\alpha_2}, \tag{19}$$

where α_1 and α_2 are constants whose values are to be determined.

The minimum *MSE* of $\bar{y}^*_{SK(R2)}$ to the first order of approximation, at optimum values of α_1 and α_2 *i.e.* $\alpha_{1opt} = \frac{(V_{yx}^* - V_{yx})}{(V_x^* - V_x)}$ and $\alpha_{2opt} = \frac{(V_x^* V_{yx} - V_{yx}^* V_x)}{V_x(V_x^* - V_x)}$, is given by

$$MSE(\bar{y}^*_{SK(R2)})_{min} \cong \bar{Y}^2 \left[V_y^* - \frac{V_x V_{yx}^{*2} + V_x^* V_{yx}^2 - 2V_{yx}V_{yx}^* V_x}{V_x(V_x^* - V_x)}\right]. \tag{20}$$

Note that $\bar{y}^*_{SK(R2)}$ performs better than $\bar{y}^*$
if $\frac{V_x V_{yx}^{*2} + V_x^* V_{yx}^2 - 2V_{yx}V_{yx}^* V_x}{V_x(V_x^* - V_x)} > 0$
Singh and Kumar (2008) also suggested a difference type estimator in the case of non-response as

$$\bar{y}^*_{SK(d)} = \bar{y}^* + d_1(\bar{x} - \bar{x}^*) + d_2(\bar{X} - \bar{x}), \tag{21}$$

where d_1 and d_2 are constants whose values are to be determined.
The minimum *MSE* of $\bar{y}^*_{SK(d)}$ to the first order of approximation, at optimum values of d_1 and d_2 *i.e.* $d_{1opt} = \frac{\bar{Y}(V^*_{yx} - V_{yx})}{\bar{X}(V^*_x - V_x)}$ and $d_{2opt} = \frac{\bar{Y}V_{yx}}{\bar{X}V_x}$, is given by,

$$MSE(\bar{y}^*_{SK(d)})_{min} \cong \bar{Y}^2 \left[V^*_y - \frac{(V^*_{yx} - V_{yx})^2}{(V^*_x - V_x)} - \frac{V^2_{yx}}{V_x} \right]. \tag{22}$$

Note that $\bar{y}^*_{SK(d)}$ performs better than $\bar{y}^*$ if $\left[\frac{(V^*_{yx} - V_{yx})^2}{(V^*_x - V_x)} + \frac{V^2_{yx}}{V_x} \right] > 0$. Also, we observed that $MSE(\bar{y}^*_{SK(d)})_{min} = MSE(\bar{y}^*_{SK(R2)})_{min}$.

6. Kumar and Bhougal (2011) proposed ratio-product-type exponential estimator for the population mean $\bar{Y}$, when non-response exists on both the study variable y and the auxiliary variable x as

$$\bar{y}^*_{KB} = \bar{y}^* \left[\alpha \exp\left(\frac{\bar{X} - \bar{x}^*}{\bar{X} + \bar{x}^*} \right) + (1 - \alpha) \exp\left(\frac{\bar{x}^* - \bar{X}}{\bar{x}^* + \bar{X}} \right) \right], \tag{23}$$

where α is a constant whose value is to be determined.
The minimum *MSE* of $\bar{y}^*_{KB}$ at optimum value of $\alpha_{opt} = \frac{1}{2}(1 + 2\frac{V^*_{yx}}{V^*_x})$, to the first order of approximation, is given by

$$MSE(\bar{y}^*_{KB})_{min} \cong \bar{Y}^2 \left(V^*_y - \frac{V^{*2}_{yx}}{V^*_x} \right). \tag{24}$$

Note that $\bar{y}^*_{KB}$ performs better than $\bar{y}^*$ if $\frac{V^{*2}_{yx}}{V^*_x} > 0$, which is always true.

3. Class of Estimators

In application, our purpose was to construct a type of general class of estimators which contains many estimators, stable and efficient. So motivated by Singh and Shukla (1993) and Shukla et al. (2012), we propose the following general class of estimators in the case of non-response exists on the study variable as well as on the two auxiliary variables. Initially Bahl and Tuteja (1991) gave the idea of exponential ratio type and product type estimators for estimating the population mean by using the single auxiliary variable. Also, we can generate many more estimators by substituting different values of $(K_i, i = 1,2,3,4)$. The proposed estimator is constructed by combining the ideas of Bahl and Tuteja (1991), Singh and Shukla (1993) and Shukla (2012), given by

$$\bar{y}^*_{prop} = \bar{y}^* \left[\exp\left(\frac{G_1 - D_1}{G_1 + D_1} \right) \exp\left(\frac{G_2 - D_2}{G_2 + D_2} \right) \right], \tag{25}$$

where "*prop*" indicates proposed. $G_1 = (A_1 + C_1)\bar{X} + fB_1\bar{x}^*$, $G_2 = (A_2 + C_2)\bar{Z} + fB_2\bar{z}^*$,

$D_1=(A_1+fB_1)\bar{X}+C_1\bar{x}^*$, $D_2=(A_2+fB_2)\bar{Z}+C_2\bar{z}^*$, $A_i=(K_i-1)(K_i-2)$, $B_i=(K_i-1)(K_i-4)$, $C_i=(K_i-2)(K_i-3)(K_i-4)$, $(i=1,2,3,4)$.
Substituting different values of K_i in (25), we can generate many more different types of estimators from our proposed class of estimators given in Table 1.

Table 1: Some members of the proposed class of family of estimators $\bar{y}^*_{prop}$

Estimators	Estimators
$K_1=1$ and $K_2=1$ $\bar{y}^*_{prop1}=\bar{y}^*\exp\left(\frac{\bar{X}-\bar{x}^*}{\bar{X}+\bar{x}^*}\right)\exp\left(\frac{\bar{Z}-\bar{z}^*}{\bar{Z}+\bar{z}^*}\right)$	$K_1=1$ and $K_2=2$ $\bar{y}_{prop2}=\bar{y}^*\exp\left(\frac{\bar{X}-\bar{x}^*}{\bar{X}+\bar{x}^*}\right)\exp\left(\frac{\bar{z}^*-\bar{Z}}{\bar{z}^*+\bar{Z}}\right)$
$K_1=1$ and $K_2=3$ $\bar{y}^*_{prop3}=\bar{y}^*\exp\left(\frac{\bar{X}-\bar{x}^*}{\bar{X}+\bar{x}^*}\right)\exp\left(\frac{n(\bar{Z}-\bar{z^*})}{2N\bar{Z}-n(\bar{z}^*+\bar{Z})}\right)\Big)$	$K_1=1$ and $K_2=4$ $\bar{y}^*_{prop4}=\bar{y}^*\exp\left(\frac{\bar{X}-\bar{x}^*}{\bar{X}+\bar{x}^*}\right)$
$K_1=2$ and $K_2=1$ $\bar{y}^*_{prop5}=\bar{y}^*\exp\left(\frac{\bar{x}^*-\bar{X}}{\bar{x}^*+\bar{X}}\right)\exp\left(\frac{\bar{Z}-\bar{z}^*}{\bar{Z}+\bar{z}^*}\right)$	$K_1=2$ and $K_2=2$ $\bar{y}^*_{prop6}=\bar{y}^*\exp\left(\frac{\bar{x}^*-\bar{X}}{\bar{x}^*+\bar{X}}\right)\exp\left(\frac{\bar{z}^*-\bar{Z}}{\bar{z}^*+\bar{Z}}\right)$
$K_1=2$ and $K_2=3$ $\bar{y}^*_{prop7}=\bar{y}^*\exp\left(\frac{\bar{x}^*-\bar{X}}{\bar{x}^*+\bar{X}}\right)\exp\left(\frac{n(\bar{Z}-\bar{z^*})}{2N\bar{Z}-n(\bar{z}^*+\bar{Z})}\right)\Big)$	$K_1=2$ and $K_2=4$ $\bar{y}^*_{prop8}=\bar{y}^*\exp\left(\frac{\bar{x}^*-\bar{X}}{\bar{x}^*+\bar{X}}\right)$
$K_1=3$ and $K_2=1$ $\bar{y}^*_{prop9}=\bar{y}^*\exp\left(\frac{n(\bar{X}-\bar{x}^*)}{2N\bar{X}-n(\bar{x}^*+\bar{X})}\right)\exp\left(\frac{\bar{Z}-\bar{z^*}}{\bar{Z}+\bar{z}^*}\right)$	$K_1=3$ and $K_2=2$ $\bar{y}^*_{prop10}=\bar{y}^*\exp\left(\frac{n(\bar{X}-\bar{x}^*)}{2N\bar{X}-n(\bar{x}^*+\bar{X})}\right)\exp\left(\frac{\bar{z}^*-\bar{Z}}{\bar{z}^*+\bar{Z}}\right)$
$K_1=3$ and $K_2=3$ $\bar{y}^*_{prop11}=\bar{y}^*\exp\left(\frac{n(\bar{X}-\bar{x}^*)}{2N\bar{X}-n(\bar{x}^*+\bar{X})}\right)\Big)\exp\left(\frac{n(\bar{Z}-\bar{z^*})}{2N\bar{Z}-n(\bar{z}^*+\bar{Z}}\right)\Big)$	$K_1=3$ and $K_2=4$ $\bar{y}^*_{prop12}=\bar{y}^*\exp\left(\frac{n(\bar{X}-\bar{x}^*)}{2N\bar{X}-n(\bar{x}^*+\bar{X})}\right)$
$K_1=4$ and $K_2=1$ $\bar{y}^*_{prop13}=\bar{y}^*\exp\left(\frac{\bar{Z}-\bar{z}^*}{\bar{Z}+\bar{z}^*}\right)$	$K_1=4$ and $K_2=2$ $\bar{y}^*_{prop14}=\bar{y}^*\exp\left(\frac{\bar{z}^*-\bar{Z}}{\bar{z}^*+\bar{Z}}\right)$
$K_1=4$ and $K_2=3$ $\bar{y}^*_{prop15}=\bar{y}^*\exp\left(\frac{n(\bar{Z}-\bar{z^*})}{2N\bar{Z}-n(\bar{z}^*+\bar{Z}}\right)\Big)$	$K_1=4$ and $K_2=4$ $\bar{y}^*_{prop16}=\bar{y}^*$

Solving $\bar{y}^*_{prop}$ given in Eq. (25) in terms of $e's$ (defined earlier in Section 2), we have

$$\bar{y}^*_{prop}\cong\bar{Y}(1+e_0^*)\left(1+\frac{1}{2}\sigma_1e_1^*-\frac{1}{4}\sigma_1\nu_1e_1^{*2}+\frac{1}{8}\sigma_1^2e_1^{*2}+....\right)\left(1+\frac{1}{2}\sigma_2e_2^*-\frac{1}{4}\sigma_2\nu_2e_2^{*2}+\frac{1}{8}\sigma_2^2e_2^{*2}+....\right), \quad (26)$$

where $\sigma_1=\frac{fB_1-C_1}{A_1+fB_1+C_1}$, $\nu_1=\frac{fB_1+C_1}{A_1+fB_1+C_1}$, $\sigma_2=\frac{fB_2-C_2}{A_2+fB_2+C_2}$, $\nu_2=\frac{fB_2+C_2}{A_2+fB_2+C_2}$.
To first order of approximation, we have

$$\bar{y}^*_{prop}-\bar{Y} \cong \bar{Y}(e_0^*+\frac{1}{2}\sigma_1 e_1^*+\frac{1}{2}\sigma_2 e_2^*+\frac{1}{2}\sigma_1 e_0^* e_1^*+\frac{1}{2}\sigma_2 e_0^* e_2^*-\frac{1}{4}\sigma_1\nu_1 e_1^{*2} -\frac{1}{4}\sigma_2\nu_2 e_2^{*2}+\frac{1}{8}\sigma_1^2 e_1^{*2}+\frac{1}{8}\sigma_2^2 e_2^{*2}). \quad (27)$$

The bias of $\bar{y}^*_{prop}$ to the first order of approximation, is given by

$$B(\bar{y}^*_{prop}) \cong \bar{Y}\left[\frac{1}{2}\sigma_1 V^*_{yx}+\frac{\sigma_1}{2}\sigma_2 V^*_{yz}+\frac{1}{8}V^*_x(\sigma_1^2-2\sigma_1\nu_1)+\frac{1}{8}V^*_z(\sigma_2^2-2\sigma_2\nu_2)\right]. \quad (28)$$

The *MSE* of $\bar{y}^*_{prop}$ to the first order of approximation, is given by

$$MSE(\bar{y}^*_{prop}) \cong \bar{Y}^2 E\left[e_0^*+\frac{1}{2}\sigma_1 e_1^*+\frac{1}{2}\sigma_2 e_2^*\right]^2.$$

Solving above equation, we have

$$MSE(\bar{y}^*_{prop}) \cong \bar{Y}^2\left[V_y^*+\frac{1}{4}\sigma_1^2 V_x^*+\frac{1}{4}\sigma_2^2 V_z^*+\sigma_1 V^*_{yx}+\sigma_2 V^*_{yz}+\frac{1}{2}\sigma_1\sigma_2 V^*_{xz}\right]. \quad (29)$$

Differentiate Eq.(29) with respect to σ_1 and σ_2, we get the optimum values of σ_1 and σ_2 i.e.

$$\sigma_{1(opt)}=\frac{2(V^*_{yz}V^*_{xz}-V^*_{yx}V^*_z)}{V^*_xV^*_z-V^{*2}_{xz}}$$

and

$$\sigma_{2(opt)}=\frac{2(V^*_{yx}V^*_{xz}-V^*_{yz}V^*_x)}{V^*_xV^*_z-V^{*2}_{xz}}.$$

Substituting the optimum values of $\sigma_{1(opt)}$ and $\sigma_{2(opt)}$ in Eq.(29), we get minimum *MSE* of $\bar{y}^*_{prop}$, given by

$$MSE(\bar{y}^*_{prop})_{min} \cong \bar{Y}^2\left[V_y^*-\frac{V^{*2}_{yx}V^*_z+V^{*2}_{yz}V^*_x-2V^*_{yx}V^*_{yz}V^*_{xz}}{V^*_xV^*_z-V^{*2}_{xz}}\right]. \quad (30)$$

4. Theoretical Comparison

A comparison of our MSE estimator and previously presented 12 different estimators is given as

• By variance of Hansen and Hurwitz (1946) estimator and our MSE estimator: $MSE(\bar{y}^*_{prop})_{min} < V(\bar{y}^*)$ if
$\frac{A_1}{A_2} > 0$, where
$A_1=(V^{*2}_{yx}V^*_z+V^{*2}_{yz}V^*_x-2V^*_{yx}V^*_{yz}V^*_{xz})$ and $A_2=(V^*_xV^*_z-V^{*2}_{xz})$.
• By MSE of Rao (1986) estimator and our MSE estimator:

$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{Rao(R)})$ if

$(V_x - 2V_{yx}) + \frac{A_1}{A_2} > 0.$

•By MSE of Rao (1986) estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{Rao(P)})$ if

$(V_x + 2V_{yx}) + \frac{A_1}{A_2} > 0.$

• By MSE of ratio estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{R})$ if

$(V^*_x - 2V^*_{yx}) + \frac{A_1}{A_2} > 0.$

• By MSE of product estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{P})$ if

$(V^*_x + 2V^*_{yx}) + \frac{A_1}{A_2} > 0.$

• By MSE of Bahl and Tuteja (1991) exponential ratio estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{exp(R)})$ if

$(\frac{1}{4}V^*_x - V^*_{yx}) + \frac{A_1}{A_2} > 0.$

• By MSE of Bahl and Tuteja (1991) exponential product estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{exp(P)})$ if

$(\frac{1}{4}V^*_x + V^*_{yx}) + \frac{A_1}{A_2} > 0.$

• By MSE of Singh and Kumar (2008) ratio type estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{SK(R1)})$ if
$[V^*_x + 3V_x - 2(V^*_{yx} + V_{yx})] + \frac{A_1}{A_2} > 0.$

• By MSE of Singh and Kumar (2008) product type estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{SK(P)})$ if
$[V^*_x + 3V_x + 2(V^*_{yx} + V_{yx})] + \frac{A_1}{A_2} > 0.$

• By MSE of Kumar and Bhougal (2011) ratio and product type estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{KB})_{min}$ if
$\frac{A_1}{A_2} - \frac{V^{*2}_{yx}}{V^*_x} > 0.$

• By MSE of Singh and Kumar (2008) chain ratio type estimator and our MSE estimator:
$MSE(\bar{y}^*_{prop})_{min} < MSE(\bar{y}^*_{SK(R2)})_{min}$ if

$\frac{A_1}{A_2} - \frac{B_1}{B_2} > 0$, where

$B_1 = (V_{yx}^{*2}V_x + V_{yx}^2 V_x^* - 2V_{yx}V_{yx}^* V_x)$ and $B_2 = V_x(V_x^* - V_x)$.

• By MSE of Singh and Kumar (2008) difference type estimator and our MSE estimator:

$MSE(\bar{y}_{prop}^*)_{min} < MSE(\bar{y}_{SK(d)}^*)_{min}$ if

$\frac{A_1}{A_2} - \frac{(V_{yx}^* - V_{yx})^2}{(V_x^* - V_x)} - \frac{V_{yx}^2}{V_x} > 0.$

Note: The proposed class of estimators performs better than all other considered estimators, if the above mentioned conditions $(i) - (Xii)$ are satisfied.

5. Numerical Comparison

To observe the performance of our proposed generalized class of estimators with respect to other considered estimators, we use the following data sets, which were earlier used by many authors in the literature. We used different values of *h*, i.e. 2, 4, 6, 8 and 16, in Tables 2-4 in our study.

1. Data set 1 [Source: Khare and Sinha (2007)]

   y: Weights of the children in kilograms.
   x: Skull circumference of the children in centimeter.
   z: Chest circumference of the children in centimeter.

   $N = 95, n = 30, W_2 = 0.25, \bar{Y} = 19.4968, \bar{X} = 51.1726, \bar{Z} = 55.8611,$
   $\rho_{yx} = 0.3280, \quad \rho_{yx(2)} = 0.4770, \quad \rho_{yz} = 0.8460, \quad \rho_{yz(2)} = 0.7290,$
   $\rho_{xz} = 0.2970, \quad \rho_{xz(2)} = 0.5700, C_y = 0.1562, \quad C_{y(2)} = 0.1207,$
   $C_x = 0.0301, \quad C_{x(2)} = 0.0247, \quad C_z = 0.0586, \quad C_{z(2)} = 0.0541.$

2. Data set 2 [Source: Khare and Sinha (2012)]

   y: Number of literate persons in the village.
   x: Number of workers in the village.
   z: Number of non-workers in the village.

   $N = 109, n = 30, W_2 = 0.25, \bar{Y} = 145.30, \bar{X} = 165.26, \bar{Z} = 259.08,$
   $\rho_{yx} = 0.81, \quad \rho_{yx(2)} = 0.78, \quad \rho_{yz} = 0.90, \quad \rho_{yz(2)} = 0.87, \rho_{xz} = 0.81,$
   $\rho_{xz(2)} = 0.74, C_y = 0.76, C_{y(2)} = 0.68, C_x = 0.68, \quad C_{x(2)} = 0.57,$
   $C_z = 0.76, \quad C_{z(2)} = 0.54.$

3. Data set 3 [Source: Khare and Sinha (2009)]

   y: Number of agricultural labors in the village.
   x: Area (in hectares) of the village.
   z: Number of cultivators in the village.

$N = 96, n = 30, W_2 = 0.25, \bar{Y} = 137.92, \bar{X} = 144.87, \bar{Z} = 185.21,$
$\rho_{yx} = 0.77, \quad \rho_{yx(2)} = 0.72, \quad \rho_{yz} = 0.78, \rho_{yz(2)} = 0.78, \rho_{xz} = 0.81,$
$\rho_{xz(2)} = 0.72, C_y = 1.32, \quad C_{y(2)} = 2.08, C_x = 0.81, \quad C_{x(2)} = 0.94,$
$C_z = 1.05, \quad C_{z(2)} = 1.48.$

We use the following expression to obtain the percent relative efficiency (PRE) for different estimators using different values of h:

$$PRE = \frac{V(\bar{y}^*)}{MSE(i) \quad \text{or} \quad MSE(i)_{min}} \times 100, \qquad i = \bar{y}^*, \bar{y}^*_{Rao(P)}, \bar{y}^*_{Rao(P)}, \bar{y}^*_R, \ldots, \bar{y}^*_{prop}.$$

Results based on three data sets are given in Tables 2, 3 and 4.

Table 2: PRE of estimators with respect to $\bar{y}^*$ for data set 1

Estimator	h=2	h=4	h=6	h=8	h=16
$\bar{y}^*$	100.000	100.000	100.000	100.000	100.000
$\bar{y}^*_R$	111.204	112.946	113.987	114.678	116.057
$\bar{y}^*_P$	84.9820	83.8472	83.200	82.782	81.974
$\bar{y}^*_{Rao(R)}$	107.908	105.704	104.460	103.662	102.134
$\bar{y}^*_{Rao(P)}$	88.1636	91.004	92.745	93.922	96.314
$\bar{y}^*_{exp(R)}$	106.369	107.189	107.672	107.991	108.621
$\bar{y}^*_{exp(P)}$	92.6901	92.032	91.654	91.407	90.929
$\bar{y}^*_{SK(R1)}$	112.749	114.116	114.927	115.465	116.532
$\bar{y}^*_{SK(P)}$	72.8894	74.829	76.007	76.799	78.397
$\bar{y}^*_{KB}$	114.506	117.822	119.946	121.417	124.495
$\bar{y}^*_{SK(R2)}$	114.819	118.348	120.506	121.961	124.915
$\bar{y}^*_{SK(d)}$	114.819	118.348	120.506	121.961	124.915
$\bar{y}^*_{prop}$	309.222	270.592	254.538	245.755	231.530

Table 3: PRE of estimators with respect to $\bar{y}^*$ for data set 2

Estimator	h=2	h=4	h=6	h=8	h=16
$\bar{y}^*$	100.000	100.000	100.000	100.000	100.000
$\bar{y}^*_R$	277.329	269.560	265.571	263.144	258.764
$\bar{y}^*_P$	31.268	31.832	32.144	32.341	32.712
$\bar{y}^*_{Rao(R)}$	203.460	155.016	137.471	128.411	114.442
$\bar{y}^*_{Rao(P)}$	36.190	44.8313	51.411	56.588	69.561
$\bar{y}^*_{exp(R)}$	205.995	201.434	199.072	197.627	195.006
$\bar{y}^*_{exp(P)}$	52.514	53.144	53.488	53.705	54.111
$\bar{y}^*_{SK(R1)}$	90.356	112.140	128.782	141.911	174.939
$\bar{y}^*_{SK(P)}$	16.087	19.056	21.147	22.701	26.275
$\bar{y}^*_{KB}$	282.247	273.501	269.048	266.350	261.505
$\bar{y}^*_{SK(R2)}$	282.309	273.589	269.133	266.427	261.559
$\bar{y}^*_{SK(d)}$	282.309	273.589	269.133	266.427	261.559
$\bar{y}^*_{prop}$	549.804	521.751	510.740	505.259	498.002

Table 4: PRE of estimators with respect to $\bar{y}^*$ for data set 3

Estimator	h=2	h=4	h=6	h=8	h=16
$\bar{y}^*$	100.000	100.000	100.000	100.000	100.000
$\bar{y}^*_R$	204.333	192.089	188.197	186.285	183.458
$\bar{y}^*_P$	47.615	50.484	51.556	52.117	52.991
$\bar{y}^*_{Rao(R)}$	142.598	118.102	111.493	108.419	104.068
$\bar{y}^*_{Rao(P)}$	59.014	73.728	80.668	84.707	91.670
$\bar{y}^*_{exp(R)}$	149.031	143.344	141.481	140.556	139.175
$\bar{y}^*_{exp(P)}$	67.732	70.041	70.875	71.305	71.967
$\bar{y}^*_{SK(R1)}$	170.523	175.322	177.041	177.925	179.283
$\bar{y}^*_{SK(P)}$	31.343	39.367	43.181	45.410	49.267
$\bar{y}^*_{KB}$	222.273	213.479	211.130	210.092	208.730
$\bar{y}^*_{SK(R2)}$	226.014	216.679	213.634	212.124	209.873
$\bar{y}^*_{SK(d)}$	226.014	216.679	213.634	212.124	209.873
$\bar{y}^*_{prop}$	289.857	289.956	290.856	291.516	292.816

5.1. Discussion and Findings

In Table 2, PRE of the proposed class of estimators $\bar{y}^*_{prop}$ and Rao (1991) ratio estimator $\bar{y}^*_{Rao(R)}$ decrease as the values of h increases from 2 to 16. On the other hand, the situation is reverse for the estimates $\bar{y}^*_R$,$\bar{y}^*_{exp(R)}$,$\bar{y}^*_{SK(R1)}$, $\bar{y}^*_{KB}$,$\bar{y}^*_{SK(R2)}$,$\bar{y}^*_{SK(d)}$.
In Table 3, the performances of the proposed estimator $\bar{y}^*_{prop}$ and all the other considered estimators decrease with an increase in the value of h.
In Table 4, PRE of the proposed class of estimators $\bar{y}^*_{prop}$ and Singh and Kumar (2008) estimator $\bar{y}^*_{SK(R1)}$, increase with an increase in the values of h. Also in this table, PRE of other estimators $\bar{y}^*_{Rao(R)}$,$\bar{y}^*_{exp(R)}$,$\bar{y}^*_{SK(R1)}$, $\bar{y}^*_{KB}$,$\bar{y}^*_{SK(R2)}$,$\bar{y}^*_{SK(d)}$ decreases with an increase in the value of h.
In Tables 2, 3 and 4, we observe that the product type estimators $\bar{y}^*_{Rao(P)}$ $\bar{y}^*_P$,$\bar{y}^*_{exp(P)}$ and $\bar{y}^*_{SK(P)}$ perform very poorly because of positive correlation in data sets 1, 2 and 3. Generally, we can use product type estimators when there exists a negative correlation between the study variable and the auxiliary variable.

6. Conclusion

We proposed a generalized class of estimators for estimating the population mean using information on two auxiliary variables under non-response in simple random sampling. Expressions for bias and MSE of the proposed generalized class of estimators are derived up to the first degree of approximation. The proposed estimator $\bar{y}^*_{prop}$ is compared with Hansen and Hurwitz (1946) estimator and other considered estimators. A numerical study is carried out to support the theoretical results. In Tables 2, 3 and 4, the proposed class of estimators performs better than all other competitor estimators under non-response in simple random sampling. The product type estimators perform poorly because of positive correlation in all data sets. Therefore, the proposed class of estimators $\bar{y}^*_{prop}$ is preferable in different situations, i.e. when no auxiliary variable, single auxiliary variable, and two auxiliary variables are used. It is observed that the Singh and Kumar (2008) estimators $\bar{y}^*_{SK(R1)}$ and $\bar{y}^*_{SK(d)}$ perform equally but $\bar{y}^*_{SK(d)}$ is preferable because of unbiasednes. All product type estimators perform poorly due to weak correlation between the study and the auxiliary variables.

Acknowledgment

Authors are thankful to referees for their valuable suggestions and comments, which led to the improvement of this article.

REFERENCES

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

EL-BADRY, M. A., (1956). A sampling procedure for mailed questionnaires. J. Amer. Statist. Assoc. 51, pp. 209–227.

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of non-response in sampling surveys. Journal of American Statistical Association. 41, pp. 517–529.

ISMAIL, M., SHAHBAZ, M. S., HANIF, M., (2011). A general class of estimator of population mean in presence of non-response. Pakistan Journal of Statistics. 27 (4), pp. 467–476.

KHARE, B. B., SRIVASTAVA, S. R., (1981). A generalized regression ratio estimator for the population mean using two auxiliary variables. The Aligarh Journal of Statistics. 1(1), pp. 43–51.

KHARE, B. B., SRIVASTAVA, S., (1993). Estimation of population mean using auxiliary character in presence of non-response. Nat. Acad. Sci. Lett. India. 16, pp. 111–114.

KHARE, B. B., SRIVASTAVA, S., (1995). Study of conventional and alternative two-phase sampling ratio, product and regression estimators in presence of non-response. Nat. Acad. Sci. Lett. India. 65, pp. 195–203.

KHARE, B. B., SRIVASTAVA, S., (1997). Transformed ratio type estimators for the population mean in presence of non-response. Communication in Statistics Theory and Method. 17, pp. 1779–1791.

KHARE, B. B., SRIVASTAVA, S., (2010). Generalized two phase estimators for the population mean in the presence of non-response. The Aligarh journal of Statistics. 30, pp. 39–54.

KHARE, B. B., SINHA, R, R., (2007). Estimation of the ratio of the two population means using multi auxiliary characteristics in the presence of non-response. In Statistical Techniques in Life Testing, Reliability, Sampling Theory and Quality Control. 1, pp. 63–171.

KHARE, B. B., SINHA, R, R., (2009). On the class of estimators for population mean using multi auxiliary characters in the presence of non-response. Statistics in Transition-New Series. 10 (1), pp. 3–14.

Khare, B. B., Sinha, R, R., (2012). Improved classes of estimators for ratio of two means with with double sampling the non-respondents. Statistica. 49 (3), pp. 75–83.

KUMAR, S., BHOUGAL, S., (2011). Estimation of the population mean in presence of non-response. Communications of the Korean Statistical Society. 18 (4), pp. 537–548.

MUNEER, S., SHABBIR, J, KHALIL, A., (2017). Estimation of finite population mean in simple random sampling and stratified random sampling using two auxiliary variables. Communication in Statistics Theory and Methods. 46 (5), pp. 2181–2192.

RAO, P. S. R. S., (1986). Ratio estimation with sub sampling the non-respondents. Survey Methodology. 12 (2), pp. 217–230.

OLKIN, I., (1958). Multivariate ratio estimation for finite populations. Biometrika, 45, pp. 154–165.

SINGH, H. P., KUMAR, S., (2008). A regression approach to the estimation of the finite population mean in the presence of non-response. Aust. N. Z. J. Stat. 50(4), pp. 395–408.

SINGH, H. P., KUMAR, S., KOZAK, M., (2010) Improved estimation of finite population mean using sub-sampling to deal with non response in two phase sampling scheme. Communication in Statistics Theory and Method. 39 (5), pp. 791–802.

SRINATH, K. P., (1971). Multiphase sampling in non-response problems. Journal of American Statistical Association. 66, pp. 583–586.

SHABBIR, J., NASIR, S., (2013). On estimating the finite population mean using two auxiliary variables in two phase sampling in the presence of non response. Communication in Statistics-Theory and Methods. 42, pp. 4127–4145.

SHUKLA, D., PATHAK, S., THAKUR, N. S., (2012) A transformed estimator for estimation of population mean with missing data in sample surveys. Journal of current engineering research. 2 (1):50-55.

SINGH, V. K., SHUKLA, D., (1993) An efficient one parameter family of factor-type estimator in sample survey. Metron. 51 (1-2), pp. 139–159.

TABASUM, R., KHAN, I. A., (2006). Double sampling for ratio estimator for the population mean in the presence of non-response. Assam Statistical Review 20, pp. 73–83.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 277–296, DOI 10.21307/stattrans-2018-016

ON MEASURING POLARIZATION FOR ORDINAL DATA: AN APPROACH BASED ON THE DECOMPOSITION OF THE LETI INDEX

Mauro Mussini[1]

ABSTRACT

This paper deals with the measurement of polarization for ordinal data. Polarization in the distribution of an ordinal variable is measured by using the decomposition of the Leti heterogeneity index. The ratio of the between-group component of the index to the within-group component is used to measure the degree of polarization for an ordinal variable. This polarization measure does not require imposing cardinality on ordered categories to quantify the degree of polarization in the distribution of an ordinal variable. We address the practical issue of identifying groups by using classification trees for ordinal variables. This tree-based approach uncovers the most homogeneous groups from observed data, discovering the patterns of polarization in a data-driven way. An application to Italian survey data on self-reported health status is shown.

Key words: polarization, ordinal data, Leti index, classification trees.
JEL: D31, C40, C46

Introduction

Surveys frequently comprise one or more questions asking a respondent to self-assess his status (e.g., health, well-being, satisfaction) by choosing a response category from a set of ordered categories. When analyzing polarization in the distribution of an ordinal variable, one approach consists in imposing cardinality on ordinal categories to calculate conventional polarization measures. However, Apouey (2007) argued that transforming ordinal data into cardinal data is a supra-ordinal assumption, and she proposed bi-polarization indices which do not require supra-ordinal assumptions. Apouey's indices measure bi-polarization in the distribution (Wolfson, 1994); that is, the disappearing of the central class induced by the distribution of the observations towards the lower and upper categories rather than around the central categories. The concept of bi-polarization differs from that of polarization, since the latter is the tendency of grouping around local poles (Deutsch et al., 2013), which can be more than two and different from the extreme categories. In this paper, we use classification and regression trees (CART) (Breiman et al., 1984) for uncovering polarization

[1] Department of Economics, University of Verona. Italy. E-mail: mauro.mussini@univr.it.

patterns when dealing with ordinal data. The use of regression trees to explore polarization in income distribution has been recently investigated by Mussini (2016). Classification trees for ordinal variables (Piccarreta, 2008) are used to handle ordinal data. To quantify the polarization uncovered from ordinal data exploration, a measure based on the decomposition of the Leti heterogeneity index by group is applied. We show that this polarization measure is coherent with a criterion for identifying groups of observations in classification trees for ordinal variables.

Polarization is a relevant topic in studies on income distribution (Esteban and Ray, 1994; Duclos et al., 2004; Palacios-González and García-Fernández, 2012; Jenkins, 1995; Bossert and Schworm, 2008; Wang and Tsui, 2000; Yitzhaki, 2010; Gigliarano and Mosler, 2009; Foster and Wolfson, 2010) and its original notion is based on the concept of identification-alienation: individuals identify themselves with those having similar income levels, whereas they feel alienated from those with different income levels. When measuring polarization for an ordinal variable whose categories describe the status of an individual, there is polarization when groups of individuals characterized by within-group homogeneity (identification) and between-group heterogeneity (alienation) are observable. A similar approach was suggested by Fusco and Silber (2014), who defined the situations with the lowest and highest levels of polarization under the assumption that groups are defined a priori. According to Fusco and Silber (2014), polarization is lowest if each group shows the same relative frequency distribution of individuals between the various ordered categories; that is, if an individual cannot identify himself with the members of his group or distinguish himself from those of the other groups. Polarization is highest if all the individuals within a group belong to one category and such category varies according to the group considered; that is, if an individual can fully identify himself with the members of his group and feel alienated from those of the other groups. This approach based on within-group homogeneity and between-group heterogeneity is in line with that suggested by Zhang and Kanbur (2001) for measuring income polarization[2], however it suffers from the practical limitation that groups must be defined a priori (Duclos et al., 2004). We overcome this limitation by identifying groups through data exploration. We show that groups can naturally emerge from data by using classification trees to recursively partition individuals into groups. We assume that the ordinal variable is the response variable and some variables describing respondents (e.g., earned income, age, gender, education) are the explanatory variables. The population is recursively partitioned to maximize the between-group heterogeneity, which is equivalent to searching for the partition maximizing the gain in homogeneity within groups. A classification tree can uncover groups of homogeneous respondents in a data-driven way by selecting the explanatory variables which play a role in the polarization of the distribution of the response variable. Thus, polarization is examined on the basis not only of the response variable distribution but also of the socio-demographic characteristics of individuals, as suggested by Permanyer and D'Ambrosio (2015).

The classification tree is obtained by applying the ordinal Gini-Simpson criterion proposed by Piccarreta (2008), which is based on a measure of

[2] Given an inequality index (e.g. the Theil index), Zhang and Kanbur (2001) suggested measuring polarization by the ratio of the between-group component of the index to the within-group component.

heterogeneity for ordinal variables that can be expressed as a function of the between-group component of the Leti index of heterogeneity for ordinal variables (Leti, 1983). Grilli and Rampichini (2002) decomposed the Leti index of heterogeneity into two components: a within-group component measuring heterogeneity within groups, and a between-group component measuring heterogeneity between groups.[3] Building on the Zhang and Kanbur approach to the measurement of polarization for numerical variables, polarization in the distribution of an ordinal variable is measured by the ratio of the between-group component of the Leti index to the within-group component. Since both the recursive partition and the polarization measure depend on the between-group component of the Leti index, this link is used to define a procedure for measuring polarization which consists of two phases. First, the most homogeneous groups are identified by using classification trees for ordinal variables. Second, polarization is measured by breaking down the Leti index into between-group and within-group components.

We measure polarization in self-reported health data for a sample of Italian householders interviewed in the Survey on Household Income and Wealth in 2010 (Banca d'Italia, 2012). Our findings show that polarization is low and that the interaction effect of income and age contributes to explaining the polarization pattern.

The paper is organized as follows. Section 2 introduces the measure of polarization for ordinal variables. Section 3 outlines the procedure to recursively partition individuals into homogeneous groups. In section 4, an application to Italian household data on self-reported health status is shown. Section 5 concludes.

2. Measuring Polarization for Ordinal Variables

We briefly review the Leti heterogeneity index and its decomposition by group (subsection 2.1); we then introduce the measure of polarization based on the decomposition of the Leti index (subsection 2.2).

2.1 The Leti Index and Its Decomposition

Suppose that Y is an ordinal variable with k ordered categories $y_1, \cdots, y_j, \cdots, y_k$. Let n be the number of individuals and $n_1, \cdots, n_j, \cdots, n_k$ be the frequencies observed for the k ordered categories of Y. Let $F(y_j)$ be the cumulative relative frequency of y_j:

$$F(y_j) = \frac{\sum_{i=1}^{j} n_i}{n}. \quad (1)$$

The Leti index (Leti, 1983, pp. 290-297) is

$$L = 2\sum_{j=1}^{k-1} F(y_j)[1 - F(y_j)], \quad (2)$$

[3] Shorrocks (1980) defined a class of decomposable inequality measures for the measurement of inequality in the distribution of a numerical variable. Shorrocks (1984) also studied the properties of the inequality measures which can be decomposed by population subgroups.

and measures the degree of heterogeneity in the distribution of Y. The Leti index equals 0 if frequencies are concentrated in one category. The Leti index equals $(k-1)/2$ if heterogeneity is highest; that is, when frequencies are equally split between the lowest category y_1 and the highest category y_k. The Leti index can be normalized by dividing L by $(k-1)/2$.[4] Building on the conceptualization of maximum heterogeneity for an ordinal variable suggested by Leik (1966), Blair and Lacy (1996, 2000) developed a measure of heterogeneity for ordinal variables, which is equivalent to the normalized version of the Leti index. This index was used by Reardon (2009) to measure segregation in the case of an ordinal variable. In addition, the index is a member of a class of inequality measures for ordinal data that was axiomatically derived by Lv et al. (2015).

Grilli and Rampichini (2002) showed that the Leti index is decomposable by groups. Suppose the n individuals are split into h groups. Let $n_{j,g}$ be the frequency observed for category y_j within group g (with $g = 1, \cdots, h$) and n_g be the size of group g. Let $F(y_j|g)$ be the cumulative relative frequency of y_j within group g:

$$F(y_j|g) = \frac{\sum_{i=1}^{j} n_{i,g}}{n_g}. \tag{3}$$

The heterogeneity within group g can be measured by using the Leti index:

$$L_g = 2\sum_{j=1}^{k-1} F(y_j|g)\left[1 - F(y_j|g)\right]. \tag{4}$$

$p_g = n_g/n$ being the population share of group g, the within-group component of the Leti index is

$$L^W = \sum_{g=1}^{h} p_g L_g. \tag{5}$$

The between-group component of the Leti index is

$$L^B = 2\sum_{g=1}^{h} p_g \sum_{j=1}^{k-1} F(y_j|g)\left[F(y_j|g) - F(y_j)\right]. \tag{6}$$

L^B in eq. (6) measures the heterogeneity between the cumulative relative frequency distribution in the population and the cumulative relative frequency distributions in the various groups.

Since $F(y_j) = \sum_{g=1}^{h} p_g F(y_j|g)$, L^B can be rewritten as

$$L^B = 2\sum_{g=1}^{h}\sum_{j=1}^{k-1} p_g F(y_j|g)\left[\sum_{i\neq g} p_i\, F(y_j|g) - \sum_{i\neq g} p_i F(y_j|i)\right]. \tag{7}$$

Hence, after simple manipulations, an alternative expression for L^B is obtained:

$$L^B = 2\sum_{g=1}^{h}\sum_{j=1}^{k-1} p_g F(y_j|g)\left\{\sum_{i\neq g} p_i\left[F(y_j|g) - F(y_j|i)\right]\right\}$$

[4] When n is odd, the maximum value of the Leti index is $\frac{k-1}{2}\left(1 - \frac{1}{n^2}\right)$ instead of $\frac{k-1}{2}$. However, this difference is negligible when n is sufficiently large.

$$L^B = 2\sum_{g=1}^{h}\sum_{j=1}^{k-1}\left\{\sum_{i\neq g} p_i\, p_g F(y_j|g)\left[F(y_j|g) - F(y_j|i)\right]\right\}$$

$$L^B = 2\sum_{g=1}^{h}\sum_{i\neq g} p_g p_i \sum_{j=1}^{k-1} F(y_j|g)\left[F(y_j|g) - F(y_j|i)\right]$$

$$L^B = 2\sum_{g=1}^{h}\sum_{i=g+1}^{h} p_g p_i \sum_{j=1}^{k-1}\left[F(y_j|g) - F(y_j|i)\right]^2$$

$$L^B = 2\sum_{g=1}^{h}\sum_{i=g+1}^{h} p_g p_i D_{gi}. \qquad (8)$$

In eq. (8), D_{gi} measures the heterogeneity between the cumulative relative frequency distributions of groups g and i. If the two groups have the same cumulative relative frequency distribution, then $D_{gi} = 0$. L^B in eq. (8) is expressed as a function of the pairwise differences between the within-group cumulative relative frequency distributions. In this respect, there is a similarity between L^B and an index of inequality in life chances suggested by Silber and Yalonetzky (2011).[5] When all groups have the same cumulative relative frequency distribution, D_{gi} is 0 for every $g, i = 1, \cdots, h$ (with $g \neq i$) and L^B equals 0 since there is no heterogeneity between the cumulative relative frequency distributions of different groups. L^B coincides with L if the frequencies are concentrated in one category within every group; that is, when heterogeneity is fully explained by the between-group heterogeneity.

Originally, Grilli and Rampichini (2002) interpreted the ratio of L^B to L as the share of heterogeneity explained by a generic variable X used to form groups (Grilli and Rampichini, 2002, pp. 114). In the next section, we show that the ratio of the between-group component to the within-group component can be seen as a measure of polarization for ordinal variables.

2.2 A Measure of Polarization for Ordinal Variables

Polarization is the tendency of individuals to concentrate around local poles, forming groups of reasonable size in which every individual can identify himself with the members of his group and feel alienated from those of the other groups (Esteban and Ray, 1994; Duclos et al., 2004). The concept of polarization has been applied to studies on income distribution in which the original notion of identification-alienation has been adapted to the topic: individuals identify themselves with those having similar income levels, whereas they feel alienated from those with different income levels. This idea of polarization can be extended to the distribution of an ordinal variable by observing that there is polarization if groups have different relative frequency distributions and the relative frequency distribution within each group tends to converge towards a single category; that

[5] Silber and Yalonetzky (2011) proposed a set of new indices for measuring inequality in life chances in the case of an ordinal variable. One of these indices is based on pairwise comparisons between the within-group cumulative relative frequency distributions of the ordinal variable.

is, polarization occurs if groups are characterized by within-group homogeneity (identification) and between-group heterogeneity (alienation). For example, Figure 1 shows the relative frequency distribution of an ordinal variable with five response categories (ranging from “Very Poor” to “Excellent”). If we suppose that the individuals belonging to the same response category form a group of respondents with the same characteristics, we can say that the population is “polarized” in line with the Esteban and Ray general idea of polarization (Esteban and Ray, 1994). Fusco and Silber (2014) defined the situations with the lowest and highest levels of polarization for an ordinal variable, under the assumption that groups are pre-established. Polarization is lowest if each group has the same relative distribution of individuals between the various ordered categories; that is, if an individual cannot identify himself with the members of his group and distinguish himself from those of the other groups. Polarization is highest if the individuals within a group belong to a single category, and this category varies according to the group considered; that is, if an individual can fully identify himself with the members of his group and feel alienated from those of the other groups.

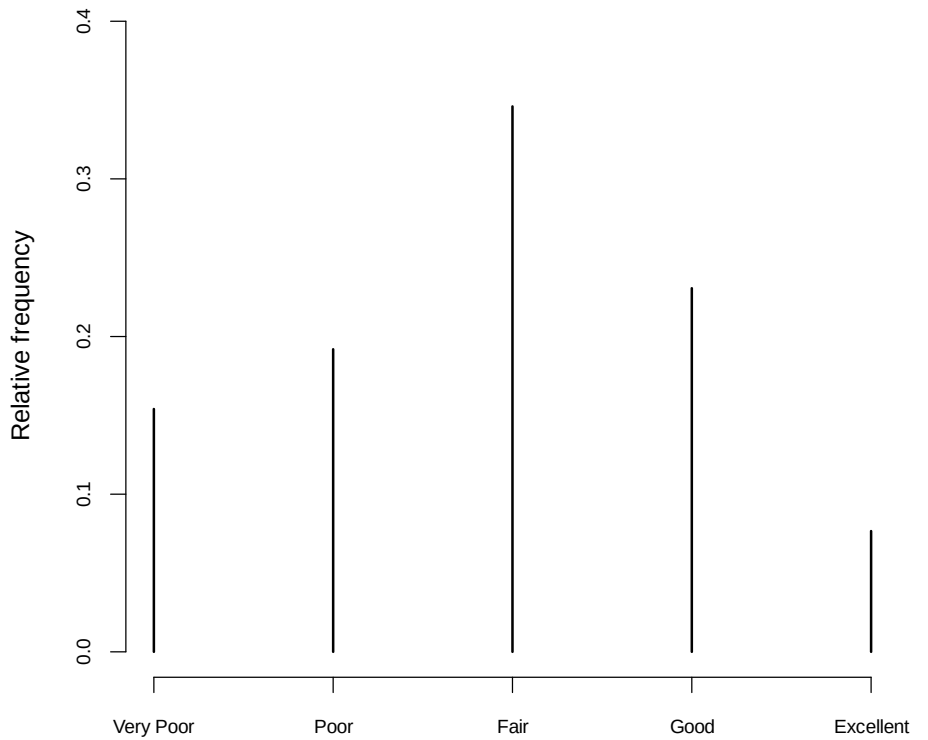


Figure 1. Relative frequency distribution of an ordinal variable

In this framework, we establish a link between the measurement of polarization and the decomposition of the Leti index. Since the between-group component measures between-group heterogeneity and the within-group component measures within-group heterogeneity, we note that polarization increases as the share of the Leti index attributable to the between-group component increases. The higher the between-group heterogeneity, the lower the within-group heterogeneity. In line with the Zhang and Kanbur approach (2001), the ratio of the between-group component to the within-group component can be interpreted as a measure of polarization:

$$PO = \frac{L^B}{L^W} = \frac{2\sum_{g=1}^{h}\sum_{i=g+1}^{h} p_g p_i D_{gi}}{\sum_{g=1}^{h} p_g L_g}. \quad (9)$$

PO equals 0 if the cumulative relative frequency distribution within each group is the same; that is, the cumulative relative frequency distribution within each group is equal to that of the whole population. In this case, polarization is lowest since there is no between-group heterogeneity. PO increases as the share of overall heterogeneity due to the between-group heterogeneity increases. While the index equals 0 in the case of minimum polarization, there is no upper limit for the index. In this respect, PO differs from conventional inequality indices, which usually range from 0 (perfect equality) to 1 (maximum inequality). The polarization index satisfies the principle of population size invariance, which is a desirable property for inequality indices. This property states that the value of the index does not change if every individual is replicated m times.[6]

The formulation of PO takes the between-group heterogeneity, within-group homogeneity and group population shares into account; that is, the three main features of polarization (Esteban and Ray, 1994, p. 824) are included in the polarization measure. While the role of between-group heterogeneity is clear, those of the other two features deserve some additional explanations. The role of within-group homogeneity is considered by the within-group heterogeneity component in the denominator of the ratio in eq. (9). The higher the within-group homogeneity, the lower the denominator. Therefore, a gain in within-group homogeneity increases polarization, all other things being equal. From eq. (9), we see that smaller groups carry less weight in the measurement of polarization than larger groups. In addition, considering groups g and i and holding the sum of their population shares constant, the more similar their population shares, the greater the weight assigned to the heterogeneity between their cumulative relative frequency distributions. In eq. (9), D_{gi} is weighted by the product $p_g p_i$, which increases as p_g and p_i become closer, holding the sum of the population shares of the two groups constant.

[6] Silber and Yalonetzky (2011) introduced an alternative property linked to population replication for indices measuring inequality in the case of ordinal data, named population composition invariance. The property of population composition invariance states that the value of the index is unchanged if every individual within a group g is replicated m times. This property is not satisfied by PO since the population share of group g and those of other groups would change if the population of group g were replicated a certain number of times.

To apply the Leti-based measure of polarization, the partition of individuals into groups is required. However, assuming that groups are pre-established does not necessarily reflect the actual polarization in the distribution of an ordinal variable. Moreover, the choice of the criterion to form groups is a practical issue to be addressed (Duclos et al., 2004). We overcome these issues by letting homogeneous groups be formed in a data driven way. To uncover the most homogeneous groups, we use classification trees for ordinal variables (Piccarreta, 2008), in which the recursive partition relies on a heterogeneity measure that can be expressed as a function of the between-group component of the Leti index. In the next section, we show that classification trees are useful to detect the most homogenous groups, since each group is composed of individuals who have the same characteristics (e.g. age, gender, occupational attainment, education) and are similar in terms of ordinal response categories. In fact, the classification tree procedure includes some individuals in the same group if they are similar in terms of a set of variables and the variable values characterizing that group differ from those characterizing the other groups, in line with the original idea of polarization proposed by Esteban and Ray (1994).

3. Using Classification Trees for Detecting Homogenous Groups

Classification and regression trees (Breiman et al., 1984) are nonparametric methods for exploring data or predicting new observations. If the response variable is categorical (numerical), a classification (regression) tree is produced. In a classification tree, the variation of a response categorical variable is explained by a set of explanatory variables. The classification tree is produced by recursively partitioning individuals into more homogeneous groups, each of which is characterized by both the within-group distribution of the response variable and the values of explanatory variables describing the members of the group. When dealing with an ordinal response variable, the conventional criteria for partitioning individuals into groups may not lead to the best partition (Piccarreta, 2008). Piccarreta (2008) extended the classification tree method and introduced splitting criteria to deal with an ordinal response variable. Here, we use ordinal classification trees as an explorative statistical tool for uncovering the relationships between an ordinal response variable and a set of individual's characteristics.

3.1 Classification Trees for Ordinal Variables

Let $(Y, \boldsymbol{X}): \Omega \to \left(S_Y \times S_{X_1} \times \cdots \times S_{X_p}\right) \equiv S$ be a vector random variable on the probability space (Ω, F, P), where Y is an ordinal variable and $\boldsymbol{X} = \{X_1, \cdots, X_m, \cdots, X_p\}$ are p explanatory variables. Assume that Y is the response variable, with k ordered categories $(y_1, \cdots, y_j, \cdots, y_k)$, and $\boldsymbol{X}$ is the vector collecting p individual's characteristics. The classification tree is built by recursively partitioning the space S into disjoint subsets, such that each subset includes individuals who are as homogeneous as possible in terms of Y. Initially, all individuals are included in one set, called the root node, and then are split into

subsets, called nodes. The degree of heterogeneity of the response variable within a node is measured by defining an impurity measure. In the case of an ordinal response variable, impurity can be measured by using the Gini index of heterogeneity of an ordinal variable (Gini, 1954):

$$I_t(Y) = \sum_{j=1}^{k} F(y_j|t)\left[1 - F(y_j|t)\right], \tag{10}$$

where t is a generic node, which coincides with the root node at the beginning of the recursive partitioning procedure. The partitioning procedure starts by splitting a parent node (the root node) into two descendent nodes according to a cut-off value chosen among all the observed values of the explanatory variables $\boldsymbol{X}$. Such a cut-off value is selected to maximize the decrease in the impurity measure in eq. (10). In the next step, each descendent node is split into two further subsets according to the partition maximizing the decrease in impurity. In each step of the splitting procedure, the decrease in impurity is measured by subtracting the impurity within the descendent nodes from the impurity of the parent node. To explain the criterion for partitioning a parent node into two descendent nodes, consider a generic node t with n_t individuals. Without loss of generality, we may assume that X_m is a numerical explanatory variable. Let $c \in S_{X_m}|t$ stand for a value of X_m, with the domain of X_m restricted to node t. Let t_l and t_r be the descendent nodes obtained by splitting t at the cut-off c. Let $n_{t_l} = \sum_{i=1}^{n_t} I_{\{X_{m,i} \le c\}}$ and $n_{t_r} = \sum_{i=1}^{n_t} I_{\{X_{m,i} > c\}}$ be the numbers of individuals in nodes t_l and t_r, respectively. The decrease in impurity obtained by splitting t into two nodes, t_l and t_r, at c is

$$\Delta_t(Y, c) = I_t(Y) - \frac{n_{t_l}}{n_t} I_{t_l}(Y) - \frac{n_{t_r}}{n_t} I_{t_r}(Y), \tag{11}$$

where $I_{t_l}(Y) = \sum_{j=1}^{k} F(y_j|t_l)\left[1 - F(y_j|t_l)\right]$ and $I_{t_r} = \sum_{j=1}^{k} F(y_j|t_r)\left[1 - F(y_j|t_r)\right]$ are the impurity measures calculated for nodes t_l and t_r, respectively. After simple manipulations, eq. (11) can be rewritten as

$$\Delta_t(Y, c) = \frac{n_{t_l} n_{t_r}}{n_t^2} \sum_{j=1}^{k} \left[F(y_j|t_l) - F(y_j|t_r)\right]^2. \tag{12}$$

Piccarreta (2008) suggested the use of the expression in eq. (12) for measuring the decrease in impurity due to splitting t into t_l and t_r, with the exclusion of the comparison between $F(y_k|t_l)$ and $F(y_k|t_r)$:

$$\Delta_t^*(Y, c) = \frac{n_{t_l} n_{t_r}}{n_t^2} \sum_{j=1}^{k-1} \left[F(y_j|t_l) - F(y_j|t_r)\right]^2. \tag{13}$$

For node t, the splitting variable and the variable threshold c are selected from all the observed values of the explanatory variables to maximize the impurity reduction in eq. (13). This splitting procedure recursively runs until a stopping rule establishes that no further partition is useful since it does not produce any important gain in terms of within-group homogeneity and between-group heterogeneity. At the end of the procedure, the individuals in a subset (terminal node) constitute a group characterized by the distribution of Y within the group and the combination of the values of the explanatory variables which identifies that group.

3.2 Linking the Decomposition of the Leti Index with the Splitting Criteria for a Classification Tree

We show that maximizing $\Delta_t^*(Y,c)$ is equivalent to searching for the breakdown maximizing the between-group component of the Leti index calculated for node t. The Leti heterogeneity index for t is

$$L_t = 2\sum_{j=1}^{k-1} F(y_j|t)\left[1 - F(y_j|t)\right]. \tag{14}$$

Supposing that t is split into t_l and t_r, the decomposition of L_t is

$$L_t = L_t^W + L_t^B, \tag{15}$$

where the within-group component is

$$L_t^W = \frac{n_{t_l}}{n_t} 2\sum_{j=1}^{k-1} F(y_j|t_l)\left[1 - F(y_j|t_l)\right] + \frac{n_{t_r}}{n_t} 2\sum_{j=1}^{k-1} F(y_j|t_r)\left[1 - F(y_j|t_r)\right] = p_{t_l} L_{t_l} + p_{t_r} L_{t_r} \tag{16}$$

and the between-group component is

$$L_t^B = 2\left\{\frac{n_{t_l}}{n_t}\sum_{j=1}^{k-1} F(y_j|t_l)\left[F(y_j|t_l) - F(y_j|t)\right] + \frac{n_{t_r}}{n_t}\sum_{j=1}^{k-1} F(y_j|t_r)\left[F(y_j|t_r) - F(y_j|t)\right]\right\}$$

$$L_t^B = 2p_{t_l}p_{t_r}\sum_{j=1}^{k-1}\left[F(y_j|t_l) - F(y_j|t_r)\right]^2. \tag{17}$$

Irrespective of the multiplicative factor 2 in eq. (17), the comparison of eq. (17) and (13) leads to the conclusion that the decrease in heterogeneity produced by splitting node t is measured by the between-group component of the Leti index calculated for that subset. The partitioning procedure iteratively searches for the breakdown maximizing the between-group component of the Leti index.

The splitting procedure can be repeated until the terminal nodes are very small, resulting in an overlarge tree that could be difficult to interpret. Therefore, a stopping rule is needed to select the optimal tree size. A tree pruning procedure (Breiman et al., 1984) is used to find the best tree. Pruning can be performed by minimizing the following cost-complexity function for a tree T:

$$R_\alpha(T) = R(T) + \alpha \cdot |T|. \tag{18}$$

In eq. (18), $|T|$ is the tree size (i.e. the number of terminal nodes), α is a complexity parameter ranging within the interval $(0, \infty)$, and $R(T)$ is the resubstitution error. The functional form of $R(T)$ depends on the nature of the response variable Y. If Y is ordinal, $R(T)$ may coincide with either the total misclassification rate or the total misclassification cost. A misclassification occurs when the true response category of an individual is different from that assigned to him by the tree. Following Galimberti et al. (2012), the response category assigned to an individual is equal to the median category of the terminal node in which the individual is included. The total misclassification rate is equal to ratio of the number of misclassified individuals to the total number of individuals. The total misclassification rate is commonly used when dealing with a nominal variable. Piccarreta (2008) suggested assigning a cost to each misclassification given that

the response variable is ordinal instead of nominal. The misclassification cost is set equal to the number of categories separating the true response category of an individual from the response category assigned to him by the tree: for example, if the two categories are adjacent, the misclassification cost equals 1; if the true response category of an individual is y_j and the response category assigned to him is y_{j-2}, then the misclassification cost equals 2. The total misclassification cost is equal to the sum of misclassification costs.

As shown in Breiman et al. (1984), for any α there is a unique smallest tree minimizing eq. (18), therefore, finding the best tree reduces to selecting the optimal tree size. Since $R(T)$ in eq. (18) is always minimized by the largest tree, Breiman et al. (1984) suggested using V-fold cross-validation to improve the reliability of misclassification error estimates. V-fold cross-validation is performed in various steps: (i) individuals are divided into V (usually V is set equal to 10) subsets of approximately equal size; (ii) each subset in turn is left out, a tree of size $|T|$ is built by using the remaining subsets and this tree is used to predict the response categories for the members of the omitted subset; (iii) the misclassification costs are calculated for each omitted subset; (iv) the misclassification costs calculated for the V subsets are added up and the cross-validated total misclassification cost is obtained, $R^{CV}(T)$; (v) steps (i)-(iv) are repeated for every tree size. Then, $R(T)$ is replaced with $R^{CV}(T)$ in eq. (18) to select the optimal tree size.

After pruning the classification tree, the terminal nodes identify groups characterized by within-group homogeneity and between-group heterogeneity in terms of a set of variables comprising the response ordinal variable and the explanatory variables used to produce the tree. Different from the Silber and Fusco (2014) approach, groups are directly identified through data exploration by clustering individuals who are similar. Therefore, using classification trees, polarization patterns can be naturally uncovered in a data driven way. A further advantage of the tree-based approach to the identification of groups is the selection of the most important explanatory variables in determining between-group heterogeneity, since only the explanatory variables producing an appreciable decrease in impurity are shown in the classification tree.

4. Application to Data on Self-Reported Health Status

We measure the polarization in the distribution of data on self-reported health status (hereafter, SRHS) collected by the Survey on Household Income and Wealth (henceforth, SHIW) carried out by the Bank of Italy in 2010 (Banca d'Italia, 2012). SRHS data include respondents' perceptions of their general health condition, with the response categories ranging from "Very Poor" to "Excellent". The use of SRHS is very common in epidemiological surveys since it is a good predictor of mortality (Allison and Foster, 2004); moreover, socio-economic surveys frequently ask SRHS to investigate the relationship between health status and socio-economic status (Kakwani et al., 1997; Idler and Benyamini, 1997). In our analysis, polarization in SRHS is measured by exploring the relationship between SRHS and a set of explanatory socio-economic variables collected in the 2010 SHIW. First, we run the classification tree procedure to partition

respondents into homogeneous groups. Second, we measure polarization in the SRHS distribution by using PO.

The 2010 SHIW collected information on income, wealth and socio-economic variables for a sample of 7,951 households. In addition, the survey asked each householder to assess his health status and that of each household member. We focus our attention on the householder SRHS and 7,950 householders are considered[7]. Table 1 shows the description and coding for the ordinal response variable and explanatory variables. SRHS is measured with an ordinal variable having five response categories: “Very Poor”, “Poor”, “Fair”, “Good”, “Excellent”.

Table 1. Variable description and coding.

Response variable			
name	description	type	ordered categories
SRHS	self-reported health status	ordinal	"Very Poor", "Poor", "Fair", "Good", "Excellent";
Explanatory variables			
name	description	type	categories coding (for categorical variables) or range (for numerical variables)
AGE_CLASS	age class	ordinal	up to 34 years, 35-44, 45-54, 55-64, more than 64 years
AREA	geographical area of residence	nominal	N="North", C="Centre", S="South and Islands"
INCOME	household income	numerical	(0,∞)
EMPLOYMENT	employment status	nominal	(BC="blue-collar worker", OW="office worker or school teacher", M="cadre or manager", P="sole proprietor/member of the arts or professions", SE="other self-employed", R="retired", NE="other not-employed")
EDUCATION	educational qualification	ordinal	N="none", P="primary school certificate", LS="lower secondary school certificate", VS="vocational secondary school diploma", US="upper secondary school diploma", B="3-year university degree", G="5-year university degree", PG="postgraduate qualification"
ACTIVITY	sector of activity	nominal	A="agriculture, fishing", I="industry", G="general government", O="other", NA="do not know"
GENDER	gender	dichotomous	F="Female"
SIZE_TOWN	size of the town of residence	ordinal	ST="0-20,000 inhabitants", MT="20,000-40,000", LT="40,000-500,000", C="more than 500,000 inhabitants"

Figure 2 shows the relative frequency distribution of SRHS data. We observe that the median category is “Good” and that the relative frequencies in the upper categories (“Good” and “Excellent”) are greater than those in the others. We initially run the recursive partitioning procedure by setting a small value of the

[7] SRHS is not available for one of the surveyed householders; therefore, he is excluded from the empirical analysis. In all calculations, we use the sample weights provided by the SHIW.

complexity parameter (CP=0.01) to produce a large tree.[8] An overlarge tree avoids that the interaction effects between explanatory variables are not discovered because none of the associated main effects produces a split with an appreciable decrease in terms of misclassification costs.[9]

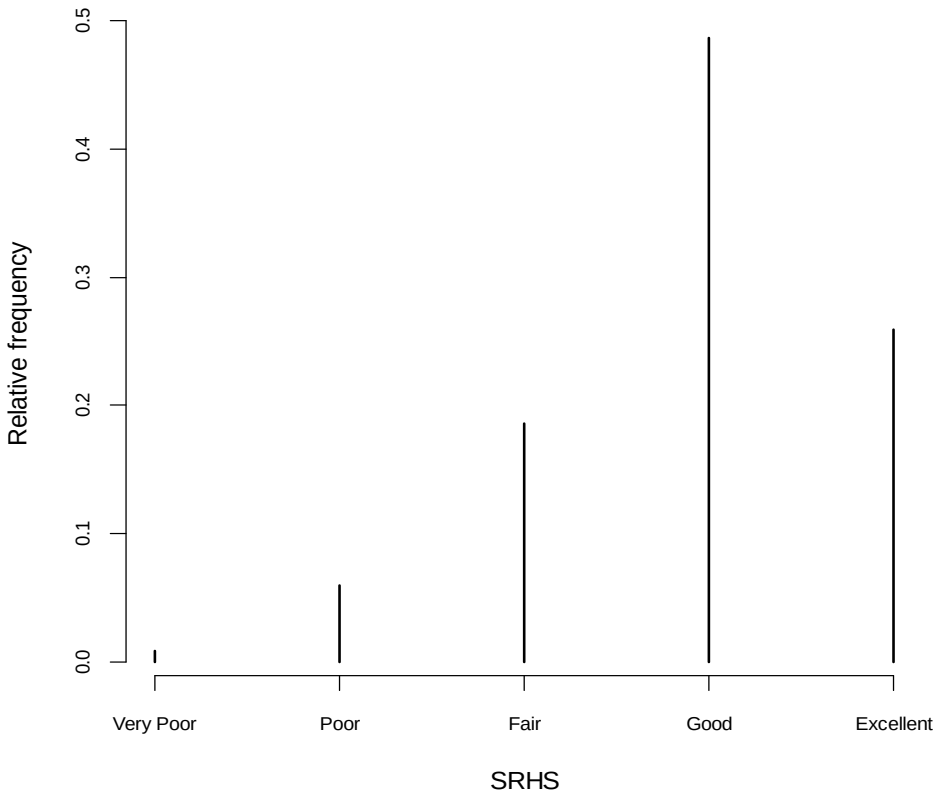


Figure 2. Relative frequency distribution of SRHS

8 We use the R package rpartScore (Galimberti et al., 2012) for recursive partitioning and we set the complexity parameter equal to the default value CP=0.01. The CP value in rpartScore is directly linked to α in eq. (18), since CP is equal to the ratio of α to the total misclassification cost calculated for the tree with no splits (i.e. the tree having no subsets). Therefore, α can be determined by setting CP.

9 Setting a large CP value serves the scope of excluding a split if it does not produce an appreciable reduction in total misclassification cost. However, if that split is made, one of the descendent subsets may be split in a way to produce an appreciable decrease in total misclassification cost. This can occur when a split based on the interaction between variables produces an appreciable decrease in total misclassification cost but none of the associated variable main effects produces an appreciable misclassification cost reduction.

Table 2 shows the tree size $|T|$ (column 1), the minimum CP value for a tree of size $|T|$ (column 2), the total misclassification cost (column 3), the 10-fold cross-validated total misclassification cost (column 4), and the standard error of the 10-fold cross-validated total misclassification cost (column 5).

Table 2. Tree size, total misclassification cost and 10-fold cross-validated total misclassification cost.

$\|T\|$	CP	$R(T)$	$R^{CV}(T)$	SE
1	0.0491	1.0000	1.0000	0.0172
3	0.0170	0.9018	0.9047	0.0174
6	0.0100	0.8507	0.8566	0.0220

Table 2 shows that the tree is not particularly successful in classifying individuals since the 10-fold cross-validated total misclassification cost is 0.8566. However, our aim is not finding a tree performing a good classification but exploring whether there are homogenous groups emerging from the data. From this standpoint, we need to handle the trade-off between the gain in within-group homogeneity and the tree size increase. We observe that passing from three to six terminal nodes does not imply a remarkable reduction of misclassification cost; that is, increasing the number of groups from three to six produces a small gain in terms of within-group homogeneity. Hence, we prune the tree by setting a complexity parameter greater than 0.01 to reduce the tree size. Figure 3 shows that the pruned tree has three terminal nodes in which the householders are split (groups 2, 6 and 7 in Figure 3). Figure 3 shows the size and the median category for each group. AGE_CLASS and INCOME are the explanatory variables playing a role in the partition of householders into groups. As expected, age has an effect on SRHS. SRHS of householders aged 65 years or older (group 3) is lower than SRHS of those younger than 65 years (group 2). Among householders aged 65 years or older (group 3), SRHS is better for householders with household income higher than 20,960.5 euros (group 7).

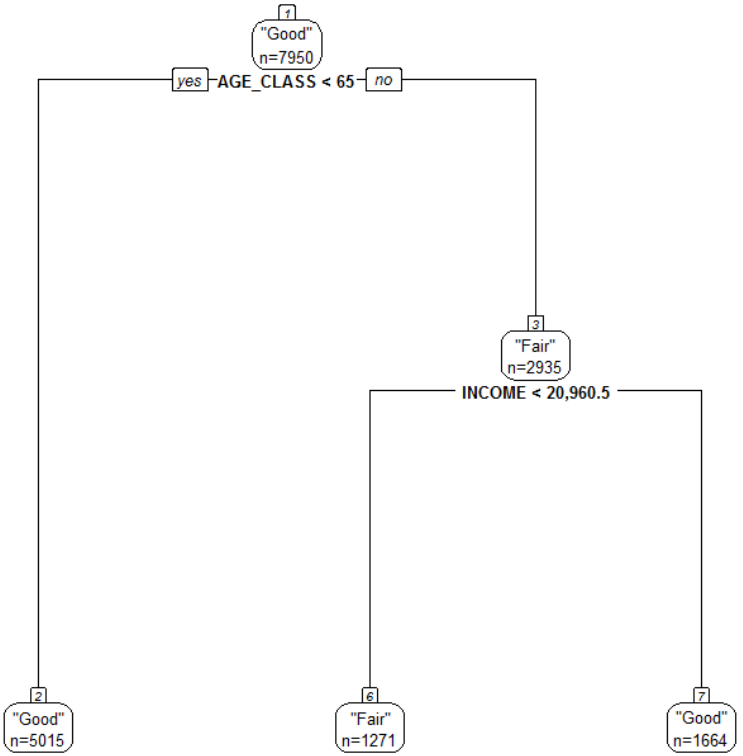


Figure 3. Classification tree for SRHS

Figure 4 shows the relative frequency distribution within each of the three groups. Although the median category of groups 2 and 7 is the same, the relative frequencies are concentrated in the upper two categories within group 2 whereas the relative frequencies spread towards the middle category within group 7. We observe that the SRHS distribution within group 2 is quite different from that within group 6, however group 6 is not very homogeneous in terms of SRHS. The normalized Leti index of the overall SRHS distribution equals 0.4542, indicating an intermediate level of heterogeneity. We break down the Leti index by group and we find that the within-group component is 0.3880 while the between-group component is 0.0662. The polarization measure PO is equal to 0.1706 and indicates that polarization is low. This means that the groups are not particularly characterized by within-group homogeneity and between-group heterogeneity with respect to SRHS and the socio-economic variables considered.

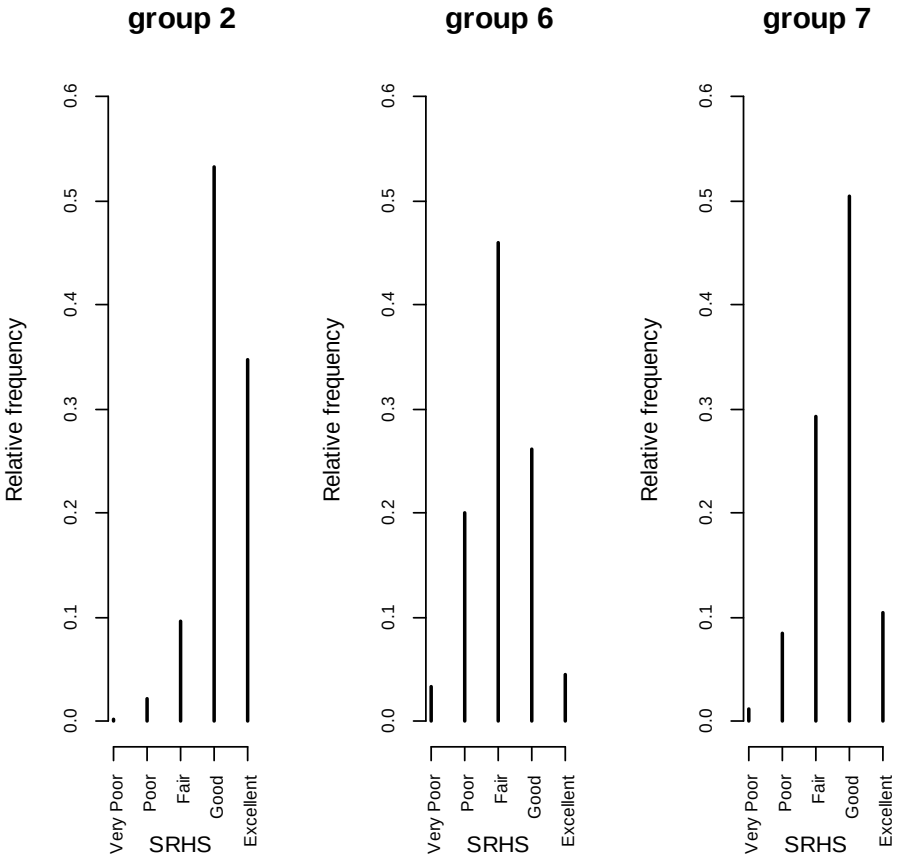


Figure 4. Relative frequency distributions of SRHS by group

5. Conclusion

This article deals with the measurement of polarization for ordinal variables. The contribution of the article is two-fold. First, we propose a synthetic measure of polarization based on the decomposition of the Leti heterogeneity index by group. Given a set of individuals split into groups by a certain criterion, the ratio of the between-group component of the Leti index to the within-group component indicates the extent to which the distribution of the ordinal variable is homogeneous within each group and heterogeneous between groups. If the within-group distributions are equal, the members of a group cannot distinguish themselves from those of the other groups. In this case, the measure of polarization equals 0, indicating that polarization is minimum. If the ordinal variable distribution within each group is mainly concentrated in a single category and this category varies according to the group considered, the within-group homogeneity is high. In this case, each member of a group can identify himself

with the members of his group and feel alienated from those belonging to the other groups. The greater the within-group homogeneity, the greater the measure of polarization. An advantage of this polarization measure is that it does not require imposing cardinality on the ordered categories of an ordinal variable. Indeed, imposing cardinality is a supra-ordinal assumption altering the original variable type.

The second contribution of the article is the use of a tree-based approach to partition individuals into homogeneous groups when exploring polarization in the distribution of an ordinal variable. As noted by Duclos et al. (2004), a practical issue in polarization studies is finding groups characterized by within-group homogeneity and between-group heterogeneity in terms of a set of variables. We show that the between-group component of the Leti index is equivalent to the impurity measure used in the process generating a classification tree for an ordinal response variable. Using classification trees, we can uncover whether individuals are naturally split into homogeneous groups, each of which comprises individuals who are similar in terms of the ordinal response variable and a set of explanatory variables. In addition, this approach is useful for selecting the explanatory variables which play a role in the polarization of the ordinal variable. Since the recursive partitioning procedure also explores the interaction effects between the explanatory variables, analysts can discover polarization patterns which cannot be assumed a priori.

We measure the polarization of SRHS data for a sample of Italian householders interviewed in the 2010 SHIW. The polarization measure is equal to 0.1706, indicating that polarization is low. The classification tree for SRHS shows that the age and household income of respondents are the most important variables in the partition of householders in terms of SRHS. All other explanatory variables, like employment status, educational qualification or gender, do not play an important role in the polarization of SRHS.

Acknowledgements

The author thanks two anonymous reviewers for their valuable comments.

REFERENCES

ALLISON, R. A., FOSTER J., (2004). Measuring health inequality using qualitative data, Journal of Health Economics 23, pp. 505–524.

APOUEY, B., (2007). Measuring health polarization with self-assessed health data, Health Economics 16, pp. 875–894.

BANCA D'ITALIA, Survey on Household Income and Wealth 2010, Rome (2012). http://www.bancaditalia.it/statistiche/indcamp/bilfait/boll_stat;internal&action=_setlanguage.action?LANGUAGE=en.

BLAIR, J., LACY, M. G., (1996). Measures of Variation for Ordinal Data as Functions of the Cumulative Distribution, Perceptual and Motor Skills 82, pp. 411–418.

BLAIR, J., LACY, M. G., (2000). Statistics of Ordinal Variation, Sociological Methods & Research 28, pp. 251–280.

BOSSERT, W., SCHWORM, W., (2008). A Class of Two-Group Polarization Measures, Journal of Public Economic Theory 10, pp. 1169–1187.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. , STONE, C. J., (1984). Classification and regression trees. Chapman & Hall/CRC press, Boca Raton.

DEUTSCH, J., FUSCO, A., SILBER, J., (2013). The BIP trilogy (Bipolarization, Inequality and Polarization): one saga but three different stories. Economics: The Open-Access, Open-Assessment E-Journal 7, pp. 2013–2022. http://www.economics-ejournal.org/economics/journalarticles/2013-22.

DUCLOS, J. Y., ESTEBAN, J. M., RAY, D., (2004). Polarization: Concepts, measurement, estimation, Econometrica 72, pp. 1737–1772.

ESTEBAN, J. M., RAY, D., (1994). On the measurement of polarization, Econometrica 62, pp. 819–851.

FOSTER, J. E., WOLFSON, M. C., (2010). Polarization and the decline of the middle class: Canada and the U.S., Journal of Economic Inequality 8, pp. 247–273.

FUSCO, A., SILBER, J., (2014). On social polarization and ordinal variables: the case of self-assessed health, European Journal of Health Economics 15, pp. 841–851.

GALIMBERTI, G., SOFFRITTI, G., DI MASO, M., (2012). Classification Trees for Ordinal Responses in R: The rpartScore Package, Journal of Statistical Software 47, pp. 1–25.

GIGLIARANO, C., MOSLER, K., (2009). Constructing indices of multivariate polarization, Journal of Economic Inequality 7, pp. 435–460.

GINI, C., (1954). Variabilità e concentrazione, Veschi, Rome.

GRILLI, L., RAMPICHINI, C., (2002). Scomposizione della dispersione per variabili statistiche ordinali, Statistica 62, pp. 111–116.

IDLER, E., BENYAMINI, Y., (1997). Self-rated health and mortality: a review of twenty-seven community studies, Journal of Health and Social Behaviour 38, pp. 21–37.

KAKWANI, N., WAGSTAFF, A., VAN DOORSLAER, E., (1997). Socioeconomic inequalities in health: Measurement, computation, and statistical inference, Journal of Econometrics 77, pp. 87–103.

JENKINS, S. P., (1995). Did the Middle Class Shrink during the 1980s? UK Evidence from Kernel Density Estimates, Economics Letters 49, pp. 407–413.

LEIK, R. K., (1966). A Measure of Ordinal Consensus, The Pacific Sociological Review 9, pp. 85–90.

LETI, G., (1983). Statistica descrittiva, il Mulino, Bologna.

LV, G., WANG, Y., XU, Y., (2015). On a new class of measures for health inequality based on ordinal data, Journal of Economic Inequality 13, pp. 465–477.

MUSSINI, M., (2016). On measuring income polarization: an approach based on regression trees, Statistics in Transition – new series 17, pp. 221–236.

PALACIOS-GONZÁLEZ, F., GARCÍA-FERNÁNDEZ, R. M., (2012). Interpretation of the coefficient of determination of an ANOVA model as a measure of polarization, Journal of Applied Statistics 39, pp. 1543–1555.

PERMANYER, I., D'AMBROSIO, C., (2015). Measuring Social Polarization with Ordinal and Categorical Data, Journal of Public Economic Theory 17, pp. 311–327.

PICCARRETA, R., (2008). Classification trees for ordinal variables, Computational Statistics 23, pp. 407–427.

REARDON, S. F., (2009). Measures of ordinal segregation, in Y. Flückiger, S. F., Reardon and J., Silber (eds.), Occupational and Residential Segregation, Research on Economic Inequality 17, pp. 129–155, Emerald, Bingley.

SHORROCKS, A. F., (1980). The Class of Additively Decomposable Inequality Measures, Econometrica 48, pp. 613–625.

SHORROCKS, A. F., (1984). Inequality Decomposition by Population Subgroups, Econometrica 52, pp. 1369–1385.

SILBER, J., YALONETZKY, G., (2011). On Measuring Inequality in Life Chances when a Variable is Ordinal, in Juan Gabriel Rodríguez (ed.) Inequality of Opportunity: Theory and Measurement, Research on Economic Inequality 19, pp.77–98, Emerald, Bingley.

WANG, Y. Q., TSUI, K. Y., (2000). Polarization orderings and new classes of polarization indices, Journal of Public Economic Theory 2, pp. 349–363.

WOLFSON, M. C., (1994). When Inequalities Diverge?, American Economic Review 84, pp. 353–358.

YITZHAKI, S., (2010). Is there room for polarization?, The Review of Income and Wealth 56, pp. 7–22.

ZHANG, X., KANBUR, R., (2001). What difference do polarization measures make? An application to China, Journal of Development Studies 37, pp. 85–98.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 297–314, DOI 10.21307/stattrans-2018-017

A NEW METHOD FOR COVARIATE SELECTION IN COX MODEL

Ujjwal Das[1], **Nader Ebrahimi**[2]

ABSTRACT

In a wide spectrum of natural and social sciences, very often one encounters a large number of predictors for time to event data. An important task is to select right ones, and thereafter carry out the analysis. The ℓ_1 penalized regression, known as "least absolute shrinkage and selection operator" (LASSO) became a popular approach for predictor selection in last two decades. The LASSO regression involves a penalizing parameter (commonly denoted by λ) which controls the extent of penalty and hence plays a crucial role in identifying the right covariates. In this paper we propose an information theory-based method to determine the value of λ in association with the Cox proportional hazards model. Furthermore, an efficient algorithm is discussed in the same context. We demonstrate the usefulness of our method through an extensive simulation study. We compare the performance of our proposal with existing methods. Finally, the proposed method and the algorithm are illustrated using a real data set.

Key words: Bhattacharya distance, index of resolvability, Kullback-Leibler measure, ℓ_1 penalty, proportional hazards model, time to event data.

1. Introduction

The statistical analysis of time to event data is very common in several applied fields, such as biology, medicine, economics, engineering and social sciences. Typical examples of such an event may be the onset of a disease, death of a subject under study, occurrence of default of a corporate bond, malfunctioning of a system, etc. It is very frequent to adjust the analysis of those event times by incorporating the information available from covariates. One of the popular ways of analysing time to event data is based on the hazard rate function, and a common way of modelling the hazard rate function with covariate matrix Z is to write it as the product of the baseline hazard and some function of Z. This model referred to as 'proportional hazards' or the 'Cox model', can connect the covariates with time to event in a parametric or semi-parametric fashion. Mathematically, from Cox (1972) we have

$$h(t|Z) = h_0(t)\exp(Z'\beta), \tag{1.1}$$

where $h_0(t)$ is called the baseline hazard rate, $Z'\beta = \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p$ and $\exp(Z'\beta)$ describes how the hazard rate varies in response to covariates. One may

[1]Operations Management, Quantitative Methods and Information Systems Area, Indian Institute of Management, Udaipur 313001, Rajasthan, India. E-mail: ujjwal.das@iimu.ac.in.
[2]Division of Statistics, Northern Illinois University, Dekalb, IL 60115, USA.
E-mail: nader@math.niu.edu.

assume some parametric form for $h_0(t)$ and then (1.1) reduces to a parametric model. If no parametric form is assumed for $h_0(t)$ then the model (1.1) is semi-parametric. In practice, the estimation and inference from the Cox model is based on the partial likelihood function. But for our purpose we use the full likelihood function.

In some practical studies such as genetics, researchers may have a large number of covariates (p) from fewer number of observations n, and they may need to select only few of those many covariates. An example includes a typical microarray data set that consists of thousands of genes from hundred subjects. Traditional selection methods such as stepwise deletion or best subset selection though useful but may perform poorly in high dimensional $(p >> n)$ situations. The limitations of the existing methods of model selection are mentioned in Breiman (1996) and Fan and Li (2001). As a unified method of variable selection for both low and high dimension, the penalized approach has gained increasing popularity in recent years. The penalized methods with some conditions on the penalty functions, not only retain the good properties of the old methods but also enjoy theoretical justifications. Among the convex penalty functions, the least absolute shrinkage and selection operator or LASSO proposed by Tibshirani (1996) has gained enormous attention from the researchers. LASSO is defined as the ℓ_1 norm of the parameters: $\lambda||\beta||_1$, where β is the vector of regression coefficients and λ is the tuning parameter or penalizing parameter. The penalizing parameter plays an influential role for variable selection. A larger value of λ exerts a higher penalty on regression coefficients, resulting in inclusion of fewer variables in the model. Conversely, a small value of λ leads to less penalty, and hence inclusion of many variables. Commonly, a sequence of λ values is generated and then variables are detected for each value of the series. Thereafter, a value of λ is chosen by k-fold cross validation, and corresponding set of predictors are included in the model. Tibshirani (1997) used generalized cross validation for the Cox model. More recently, Simon et al. (2011) developed an R-package for variable selection in Cox model via LASSO with λ selected thorough cross-validation. Li and Barron (2000) developed the concept of information theoretically valid ℓ_1-penalty by extending the work of Grunwald (2007). Using a similar risk analysis Barron *et al.* $(2008a)$ and, Barron and Luo (2008) developed the concept of information theoretically valid ℓ_1 norm penalty function for linear models. They obtained a lower bound on the penalizing parameter which makes the LASSO penalty information theoretically valid. Recently, Das and Ebraimi (2017) extended the concept for accelerated failure time model. In this paper, we introduce the information theory for time to event data under the model (1.1) and obtain the bound for λ. The nonlinear structure of the model (1.1) makes the results more intricate than linear models. We will use the lower bound as the value of the initial penalizing parameter. In addition to that, we propose an efficient algorithm for the Cox proportional hazards model for variable selection following Barron et al. (2008). Any software that performs constrained optimization, can be used to implement the proposed algorithm.

The paper is organized as follows. A brief description on information theory

along with related concepts and, the determination of the bound on penalizing parameter for the Cox model are given in subsections 2.1 and 2.2, respectively. Section 3 deals with the algorithm and its accuracy. Section 4 ensures the usefulness of the proposed methodology through extensive simulation studies. The results are presented in a tabular format for different combinations of n and p with different censoring proportions. The performance of our method is compared with existing methods of selecting the tuning parameter in the immediate section. In Section 6 we illustrate our proposed method using a real world data, and compare the results with other methods. Finally, some concluding remarks complete the paper in Section 7.

2. Method

2.1. Preliminaries

This section provides a brief summary of different information measures. For a detailed discussion one can see Ebrahimi *et al.* (2010). The most well-known and widely used measure of uncertainty is Shannon's entropy (Shannon, 1948). For a random variable X with a domain S, its entropy $H(X)$ is defined as $-\int_S \log p(x) dP(x)$, where $P(x)$ is the cumulative distribution function and $p(x)$ is the probability density (mass) function of X. As a measure of information discrepancy between two probability distribution functions P and Q, we use Kullback-Leibler (KL) divergence (Kullback, 1959) given by $D(P,Q) = E_p \log(\frac{p}{q}) = \int_S \log(\frac{p(x)}{q(x)}) dP(x)$, provided P is absolutely continuous with respect to Q on the support S. Bhattacharya distance is an alternative way to discriminate between two distribution functions P and Q, and it is given by

$$d(P,Q) = -2\log \int \sqrt{p(x)q(x)}dx, \tag{2.1}$$

see Bhattacharya, (1943). Throughout this paper, Bhattacharya distance is used as the loss function to judge the accuracy of the estimate.

Index of Resolvability: Let L_f be the likelihood characterized by f and f^* is the true value of f. Then, the index of resolvability is defined as

$$R_n(f^*) = \min_{f \in \mathscr{F}} \left\{ \frac{1}{n} D(L_{f^*}, L_f) + \frac{1}{n} pen(f) \right\}, \tag{2.2}$$

where f is a candidate to estimate unknown f^*, $\mathscr{F}$ is the set of all possible values of f and $pen(f)$ denotes some penalty function. We use this index to upper-bound the statistical risk, associated with the estimates obtained by achieving the following minimization

$$\min_{f \in \mathscr{F}} \left\{ \frac{1}{n} \log(\frac{1}{L_f}) + \frac{1}{n} pen(f) \right\}. \tag{2.3}$$

The estimator obtained from (2.3) is called minimal complexity estimator. It can be shown that the expression under minimization in (2.3), converges in probability to

index of resolvability plus a constant (entropy), which ensures that the minimization in (2.3) is equivalent to the minimization of the resolvability index, $R_n(f^*)$ in (2.2). For more details see Barron et al. (2008).

From (1.1) f^* is the linear predictor given by $Z'\beta^*$. Let $\hat{f}$ be the minimal complexity estimator of f^*. Then we measure the associated risk of $\hat{f}$ by $E[d(L_{f^*}, L_{\hat{f}})]$. We choose the penalizing parameter of LASSO such that

$$E\left(\bar{d}(L_{f^*}, L_{\hat{f}})\right) \leq \inf_{\beta \in \mathscr{R}^p} \{\bar{D}(L_{f^*}, L_f) + \frac{\lambda}{n} \sum_{j=1}^{p} |\beta_j|\}. \tag{2.4}$$

where $\bar{d}(L_{f^*}, L_{\hat{f}}) = d(L_{f^*}, L_{\hat{f}})/n$ and $\bar{D}(L_{f^*}, L_f) = D(L_{f^*}, L_f)/n$ are the average Bhattacharya distance and Kullback-Leibler measure respectively, when averaged across the n independent subjects. In the next subsection we provide a lower bound of λ so that the risk bound in (2.4) holds for Cox model.

2.2. Determination of the bound on penalizing parameter

We consider survival studies in which n individuals are put on test and data of the form (v_i, δ_i, z_i) for $i = 1, 2, ..., n$, are collected. Here, v_i is the minimum of the exact failure time X_i and the censoring time C_i of the i^{th} individual, $\delta_i = I(X_i \leq C_i)$ is an indicator variable that represents the failure status, and z_i is the corresponding covariate that may be a vector. In addition, the survival function of the i^{th} individual is $S(t|z_i) = P(X_i > t|z_i)$. The corresponding density function is $f(t|z_i)$, where X_i is the exact failure time. Furthermore, we assume that the censoring time C_i of the i^{th} individual is a random variable with survival and density functions $G(t|z_i)$ and $g(t|z_i)$ respectively, and that given $z_1, ..., z_n$, the $C_1, ..., C_n$ are stochastically independent of each other and of the independent failure times $X_1, ..., X_n$. Therefore, the full likelihood function of the data (t_i, δ_i, z_i), conditional on $z_1, ..., z_n$, is

$$L_f(v_1, v_2, ..., v_n|\delta_1, \delta_2, ..., \delta_n, Z) = \prod_{i=1}^{n} (f(x_i|Z_i)G(x_i|Z_i))^{\delta_i} (S(C_i|Z_i)g(C_i|Z_i))^{1-\delta_i}$$

Since the censoring time is noninformative, the full likelihood function can be rewritten as

$$\begin{aligned} L_f(v_1, v_2, ..., v_n|\delta_1, \delta_2, ..., \delta_n, Z) &\propto \prod_{i=1}^{n} (f(x_i|Z_i))^{\delta_i} (S(C_i|Z_i))^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left(h_0(x_i) \exp[-H_0(x_i) \exp(f_i)]\, e^{f_i}\right)^{\delta_i} \\ &\quad (\exp[-H_0(C_i) \exp(f_i)])^{1-\delta_i} \end{aligned} \tag{2.5}$$

Under the above likelihood we have the following bound on λ:
Result 1: The ℓ_1 penalized likelihood estimator $\hat{f} = f_{\hat{\beta}} = Z_i'\hat{\beta}$ obtained by

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \frac{\delta_i}{n} \left[H_0(x_i) e^{Z_i'\beta} - Z_i'\beta \right] + \sum_{i=1}^{n} \left[(1-\delta_i) \frac{H_0(C_i)}{n} e^{Z_i'\beta} \right] + \frac{\lambda}{n} ||\beta||_1 \right\} \tag{2.6}$$

attains the risk bound

$$E\bar{d}(L_{f^*}, L_{\hat{f}}) \leq \inf_{\beta} \left\{ \bar{D}(L_{f^*}, L_f) + \frac{\lambda}{n} \sum_{j=1}^{p} |\beta_j| \right\}$$

for every sample size provided that

$$\lambda \quad \geq \quad 2\sqrt{\log 2p} \sqrt{\left[\sum_{i=1}^{n} \left\{ \delta_i \left(H_0(x_i) e^{f_i} - \frac{1}{2} \right) + (1-\delta_i) \left(H_0(C_i) e^{f_i} - \frac{1}{2} \right) \right\} \right]}. \tag{2.7}$$

In practice, f_i is replaced by $\hat{f}_i$ obtained from (2.6).
Proof: The proof is outlined in the Appendix.
Remark: Under certain conditions (2.7) may not work. In that case, the bound will be

$$\lambda \quad \geq \quad 2\sqrt{\log 2p} \sqrt{\sum_{i=1}^{n} [\delta_i e^{f_i} H_0(x_i) + (1-\delta_i) e^{f_i} H_0(C_i)]}. \tag{2.8}$$

The condition is discussed in the Appendix.

3. The Algorithm

We propose an algorithm for the detection of regression parameters in the Cox model following Barron et al. (2008). For $(p < n)$ we fit the Cox-model to the data and use the point estimates as initial estimate for the algorithm. For $(p > n)$ we begin with $\beta^0 = \mathbf{0}$. Then, we estimate the cumulative baseline hazard by using the Breslow-type estimator. With these, next we estimate λ by using (2.7) or (2.8) according to the necessity. For any $t \geq 1$ we will move from $(t-1)^{th}$ step to t^{th} step of iteration by: $\beta^t = \alpha\beta^{t-1} + \gamma I_l$, where the parameters: $\alpha \in [0,1]$, $\gamma \in (-\infty, \infty)$, and I_l is a vector of zero except for l^{th} component which is 1. Combining this with (2.6)

the likelihood, as a function of α and γ, becomes

$$\begin{aligned} W^t(\alpha,\gamma,l) &= \frac{1}{n}\sum_{i=1}^{n}\delta_i\left[H_0(x_i)\exp\left(\alpha\sum_{j=1}^{p}Z_i'\beta^{t-1}+\gamma Z_{il}\right)-\left(\alpha\sum_{j=1}^{p}Z_i'\beta^{t-1}+\gamma Z_{il}\right)\right] \\ &\quad+\frac{1}{n}\sum_{i=1}^{n}\left[(1-\delta_i)H_0(C_i)\exp\left(\alpha\sum_{j=1}^{p}Z_i'\beta^{t-1}+\gamma Z_{il}\right)\right] \\ &\quad+\frac{\lambda}{n}\left(\alpha\sum_{j=1}^{p}|\beta_j^{t-1}|+|\gamma|\right). \end{aligned} \tag{3.1}$$

for every coordinate $l=1,2,...,p$. Now, we minimize (3.1) with respect to α and γ and obtain the value of the objective function for each $l=1,2,...,p$. At t^{th} iteration the optimal α_t, γ_t and $I_{l(t)}$ are those for which the value of the objective function is minimum. We change that coordinate(s) and set others to zero. At the end of each iteration the estimates of λ and cumulative baseline hazards are also updated for the next iteration. The process is repeated until no new covariate is detected and the absolute difference between the estimates from two consecutive iterations is less than some preassigned small number. Any standard software can be used for performing the constrained optimization. We call R-routine *'constrOptim'* with the option *'Nelder-Mead'* method for its suitability to optimization of non-smooth functions. The R-code can be available from the corresponding author upon request.

3.1. Convergence of the Algorithm

Let L_f be the likelihood function with unknown parameters (or linear combination of parameters) f as given in (2.5), estimated by $\hat{f}_{(k)}$ at k^{th} iteration. Then we have
Result 2: Let $L_{\hat{f}}$ be the minimal complexity estimate of L_{f^*} and $L_{\hat{f}_k}$ be the estimate from k^{th} iteration obtained by our proposed algorithm. Then,

$$\frac{1}{n}\log\frac{1}{L_{\hat{f}_{(k)}}(x)}+\lambda u_{(k)}\leq\inf_f\left\{\frac{1}{n}\log\frac{1}{L_f(x)}+\lambda U_f+\frac{4U_f^2}{k+1}\right\}, \tag{3.2}$$

where $u_{(k)}=\sum_{j=1}^{p}|\hat{\beta}_{j,(k)}|$ and $U_f=\sum_{j=1}^{p}|\beta_j|$ with $\hat{\beta}_{j,(k)}$ is the estimate of β_j at k^{th} iteration.
Proof: The proof is given in the Appendix.

4. Numerical Studies

We investigate the performance of the proposed λ along with the algorithm through simulations. We will use the lower bound of λ as its value, for all numerical investigations. First, we create a matrix of 100 rows and 1000 columns by randomly drawing 1000 observations from a 100-dimensional multivariate normal distribution with mean **0** and pairwise correlation 0.1. Throughout the simulation study, we keep

this matrix fixed and use appropriate number of rows and columns as design matrix under four different scenarios: (a) n=100, p=50, (b) n=1000, p=100 for low dimension, and (c) n=50, p=100, (d) n=100, p=1000 for high dimension. For (a) we use the first 50 columns of the matrix, for (b) we transpose the matrix, and for (c) we consider the first 100 columns with their first 50 rows for numerical studies. Let β denote the true vector of regression coefficients. So, β is a vector of length 50 for (a), of length 100 for (b) and (c), and of length 1000 for (d). In each case, we randomly choose seven elements of β and set them to unity, and rest of the elements are all zero. Let Z be the design matrix of appropriate order. We model the baseline hazard of Cox regression assuming Weibull distribution to generate data. In this way we get a closed form expression for the survival function. We equate the survival function with random numbers generated from uniform(0,1) distribution, and then invert the survival function to get the time to event. We choose the scale and shape parameters of the Weibull distribution as 1 and 1.2 respectively. For a detailed discussion on the methods to generate data from the Cox model, one can see Bender, Augustin and Blettner (2005). Except for (n=50, p=100), for remaining pairs of (n, p) we vary the censoring proportion from 5% to 40$\%$ with an increment of 5%. In this way, we generate 1000 data sets for every combination of n, p and censoring percentage. Before analysis, for all the eight covariates we subtract the respective mean and then divide them by the respective standard deviation. Then, the variables are selected through the algorithm discussed in Section 3. The simulation results are summarized in Table 1, where **n** represents the number of subjects, **p** is the number of covariates as candidate of the model, **Cens. Pcnt.** gives percent of censoring, **TMDR** is the true model detection rate defined as the percentage of replications where the full model (all correct seven covariates) is detected, **Median** and **Mean** the number of correct variables detected, and **Avg. Incln.** is the average model size, from the 1000 replications.

From Table 1 we find that the method is working well for detecting the correct set of variables except for $n = 50, p = 100$. Along with the median, the average number of correct variables included is also higher than 6 for both $n = 100, p = 50$ and $n = 100, p = 1000$. We note that the average model size is not far from the average number of correct covariates detected, in all cases considered here. The phenomena indicates the inclusion of fewer false variables. More specifically, for $n = 1000, p = 100$, the entire correct model is identified always without any error for all censoring percentages. We observe that the convergence was achieved equally faster whether the initial estimate was $\mathbf{0}$ or taken from the Cox model fitting, for low dimension.

5. Comparison

We compare our proposed method of tuning parameter selection with cross-validation (CV), generalized cross-validation (GCV) and BIC. We use **R**-package *glmnet*. For a detailed discussion on the *glmnet*, its algorithm and convergence, see Simon *et al.* (2011). We reconsider the simulated data sets from Section 4, and reanalyse them

Table 1: Summary of the Simulation Studies

n	p	Cens. Pcnt.	TMDR	Median	Mean	Avg. Incln.
50	100	5.71	15.01	5	5.39	5.81
		10.88	10.1	5	4.89	5.34
		14.19	6.6	5	4.66	5.02
		19.68	1.1	4	4.12	4.66
		25.31	0	4	3.78	4.73
		29.19	0	4	3.72	4.81
100	50	5.27	90.1	7	6.87	7.05
		10.16	84.1	7	6.75	7.06
		15.22	80.8	7	6.71	7.1
		20.27	76.7	7	6.65	7.08
		25.02	74.9	7	6.64	7.09
		31.2	72.8	7	6.62	7.15
		34.87	68.1	7	6.54	7.06
		39.42	66.5	7	6.52	6.83
100	1000	5.76	84.4	7	6.79	7.58
		10.61	80.9	7	6.76	7.62
		14.85	77.2	7	6.74	7.53
		20.25	74.3	7	6.72	7.12
		25.25	71.8	7	6.65	7.24
		30.11	69.4	7	6.59	7.19
		35.71	65.9	7	6.57	7.22
		41.48	62.4	7	6.44	7.61
1000	100	5.02	100	7	7	7
		10.11	100	7	7	7
		15.01	100	7	7	7
		20.09	100	7	7	7
		25.19	100	7	7	7
		30.15	100	7	7	7
		34.9	100	7	7	7
		40.11	100	7	7	7

by *glmnet* in conjunction with *10-fold* CV, GCV and BIC. *10-fold* CV was performed using the R-function *cv.glmnet*, when for the other two we fit the Cox model with the selected predictors for every value of λ and then obtain the GCV and BIC values for each model. From a sequence of λ values, we pick the one as the value of the penalizing parameter and the corresponding model, for which the desired criterion (CV, GCV or BIC) attains its minimum. We compare the average number of variables detected for different values of n, p and censoring percentages, from all the methods. Table 2 provides the average number of predictors identified as non-zero from 1000 replications for 5% to 40% censoring. For $n = 50$ and $p = 100$ we perform the simulation up to 30% censoring. From Table 2 we see that cross-validation

Table 2: Average number of variables detected by Our Method and cross-validation

n	p	Method	5%	10%	15%	20%	25%	30%	35%	40%
50	100	Our method	5.81	5.34	5.02	4.66	4.73	4.81		
		CV	15.51	15.87	15.93	15.27	15.43	15.14		
		GCV	15.56	15.25	16.09	15.36	15.34	15.39		
		BIC	14.15	13.98	14.53	14.07	14.14	14.01		
100	50	Our method	7.05	7.06	7.1	7.08	7.09	7.15	7.06	6.83
		CV	20.15	19.63	19.92	18.99	20.19	18.75	20.19	19.09
		GCV	20.57	20.08	19.93	19.69	20.74	20.14	21.01	20.68
		BIC	15.21	15.17	14.94	14.61	15.33	15.18	15.66	14.88
100	1000	Our method	7.58	7.62	7.53	7.12	7.24	7.19	7.22	7.61
		CV	29.39	30.12	30.67	31.76	30.51	30.32	30.48	30.47
		GCV	30.15	31.87	31.29	32.27	31.52	31.11	31.61	31.83
		BIC	24.25	25.07	24.79	24.82	25.33	25.11	25.02	24.90
1000	100	Our method	7	7	7	7	7	7	7	7
		CV	38.93	37.71	38.13	37.82	36.57	36.93	37.12	36.67
		GCV	39.51	38.96	39.09	38.76	38.03	37.84	37.94	38.02
		BIC	30.01	29.71	29.37	29.58	29.12	29.17	29.93	29.66

tends to select more covariates compared to our method. For both $n = 100, p = 50$ and $n = 100, p = 1000$ the average model size by our method is near 7, whereas from cross-validation and GCV these are around 19 and 30 respectively. Similarly, for $n = 1000$ and $p = 100$ our method detects all the covariates up to 40% censoring without any false inclusion whereas the average model size from cross-validations is more than 35. The BIC tends to select fewer variables than CV and GCV but higher than our proposed method. We note that *glmnet* is almost always able to identify the correct set of covariates for the simulated data sets. For example, for $n = 50$, $p = 100$ and with 30% censoring, the true model detection rate (**TMDR**) was higher than 95% when our proposed method was unable to find all the correct covariates in a single instance. So, the cross-validations and BIC tend to select an entire set of right predictors at the cost of larger model size. Additionally, the coordinate descent algorithm seems to be faster than our algorithm. In general, we see that the proposed method may not always detect the full model, but the inclusion of a false covariate is small compared to the cross-validations and BIC, for all the scenarios we considered here.

6. Real data analysis

We analyse data on survival of the patients with advanced lung cancer. The study was conducted by North Central cancer treatment group, and described in Loprinzi et al. (1994). After some cleaning we are left with survival time on 167 subjects with information on 8 covariates. The covariates are: institution code, age in years, gender, ECOG performance score, Karnofsky performance score rated by physicians, Karnofsky performance score rated by the patients, calories consumed at meals, and weight loss in last six months. We analyse the data in three different ways, and as before, calculate the Bayesian information criterion (BIC) of the final selected model from each method for comparison. First, we select the model by BIC. The resulting model includes only two covariates: gender and ECOG performance score. BIC of this model is 1006.99, when the BIC value of the full model is 1023.48. Next, we fit the ℓ_1-penalized Cox proportional hazard model with penalizing

parameter chosen by 10-fold CV, GCV and BIC by using R-package *glmnet*. The CV and GCV identify seven out of eight covariates (exclude the variable calories consumed), and BIC after fitting the Cox model with these seven selected covariates is 1018.36. When the penalizing parameter was selected through BIC, two more predictors (age and Karnofsky performance score rated by the patients) are dropped from the model. BIC of the Cox model with these five predictors is 1011.38. Finally, we analyse the data by our proposed method. As mentioned before, we use the lower bound from (2.7) as the value of λ. Our method detects three covariates: institution code, gender and ECOG performance score. The BIC of the Cox model with these three variables comes out as 1008.66. We note that the p-values of gender and ECOG performance score are significant at 5% level when the same for the institution code (p-value$= 0.0675$) is significant at 10% level. The result seems to be consistent with our finding in Section 5.

7. Conclusion

The selection of appropriate penalty parameter has great influence on variable selection. Cross-validation is a widely used approach for choosing the parameter. Leng, Lin and Wahba (2006) suggested to go with some method other than cross-validations or BIC when covariate selection is of primary importance. The numerical results show that the model resulted from our method always includes fewer non-active covariates. From that perspective our method may be thought of as an alternative route to choose the penalizing parameter. Certainly, the proposed method is not a panacea for variable selection when event time is the outcome of interest. We have seen in Sections 2 and 7 that for low dimension our method yields the model with second smallest BIC value. But BIC-based model selection cannot be performed in high dimension where penalized regression is the only tool for variable selection. In general, for many of the situations we study in this paper, our method shows promising results. Together with these, our method may be a good candidate when covariate selection is the primary goal.

REFERENCES

BARRON, A. R., COHEN, A., DAHMEN, W., DEVORE, R., (2008). Approximations and learning by greedy algorithms, The Annals of Statistics. 36, pp. 64–94.

BARRON, A. R., HUANG, C., LI, J. Q., LUO, X., (2008a). The MDL principle, penalized likelihoods, and statistical risk, In Festschrift for Jorma Rissanen, Tampere University Press.

BARRON, A. R., LUO, X., (2008). MDL Procedures with ℓ_1 Penalty and their Statistical Risk, Proceedings Workshop on Information Theoretic Methods in Science and Engineering.

BENDER, R., AUGUSTIN, T., BLETTNER, M., (2005). Generating survival times to simulate Cox proportional hazards models, Statistics in Medicine, 24 (11), pp. 1713–1723.

BHATTACHARYA, A., (1943). On a measure of divergence between two statistical populations defined by probability distributions, Bulletin of Calcutta Mathematical Society. 35, pp. 99–109.

BREIMAN, L., (1996). Heuristics of instability and stabilization in model selection, The Annals of Statistics. 24, pp. 2350–2383.

COX, D. R., (1972). Regression models and life tables (with discussions), Journal of Royal Statistical Society, Series B. 34, pp. 187–220.

DAS, U., EBRAHIMI, N., (2017). Covariate selection for accelerated failure time data, Communications in Statistics: Theory and Methods. 46, pp. 4051–64.

EBRAHIMI, N., SOOFI, E. S., SOYER, R., (2010). Information measures in perspective, International Statistical Review. 78, pp. 383–412.

FAN, J.„ LI, R., (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, Journal of American Statistical Association. 456, pp. 1348-1360.

GRUNWALD, P., (2007). The minimum description length principle. MIT Press, Cambridge, MA.

KULLBACK, S., (1959). Information theory and Statistics. Wiley, New York; (reprinted in 1968 by Dover).

LENG, C., LIN, Y., WAHBA, G., (2006). A note on the lasso and related procedures in model selection, Statistica Sinica, 16, pp. 1273–1284.

LI, J. Q., BARRON, A. R., (2000). Mixture density estimation, S Solla, T Leen and K.R. Muller (Eds.), Advances in Neural Processing Information System, 12, pp. 279–285.

LOPRINZI, C. L., LAURIE, J. A., WIEAND, H. S., KROOK, J. E., NOVOTNY, P. J., KUGLER, J. W., BARTEL, J., LAW, M., BATEMAN, M., KLATT, N. E. et al., (1994). Prospective evaluation of prognostic variables from patient-completed

questionnaires. North Central Cancer Treatment Group. Journal of Clinical Oncology, 12 (3), pp. 601–607.

LUO, X., (2009). Penalized Likelihoods: Fast algorithms and risk bounds, Ph.D. Thesis, Statistics Department, Yale University.

ROSSI, P. H., BERK, R. A., LENIHAN., K. J., (1980). Money, Work and Crime: Some Experimental Results. New York: Academic Press.

SHANNON, C. E., (1948). A mathematical theory of communication, Bell Sys. Tech. J.. 27, pp. 379–423.

SIMON, N., FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R., (2011). Regularization paths for Cox's proportional hazards model via coordinate descent, Journal of Statistical Software, 39 (5), pp. 1–13.

Tibshirani, R., (1996). Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B, 58, pp. 267–288.

Tibshirani, R., (1997). The lasso method for variable selection in the Cox model, Statistics in Medicine, 16, pp. 385–395.

APPENDIX

Here, we outline the proof of Result 1 and show the convergence of the proposed algorithm.

Proof of Result 1:

From Barron et al. (2008) the condition on penalty function is

$$pen(f) \geq \log(\frac{L_f(X)}{L_{\tilde{f}}(X)}) - 2\log \frac{E\sqrt{\frac{L_f(X)}{L_{f^*}(X)}}}{E\sqrt{\frac{L_{\tilde{f}}(X)}{L_{f^*}(X)}}} + 2\mathscr{L}(\tilde{f}), \tag{7.1}$$

Using the full likelihood and the fact that $\tilde{f}_i \xrightarrow{p} f_i$ we get,

$$\begin{aligned}\frac{L_f(v_1,v_2,...,v_n)}{L_{\tilde{f}}(v_1,v_2,...,v_n)} &= \prod_{i=1}^{n} \frac{\left(h_0(v_i)\exp\left[-H_0(v_i)e^{f_i}\right]e^{f_i}\right)^{\delta_i}\left(\exp\left[-H_0(C_i)e^{f_i}\right]\right)^{1-\delta_i}}{\left(h_0(v_i)\exp\left[-H_0(v_i)e^{\tilde{f}_i}\right]e^{\tilde{f}_i}\right)^{\delta_i}\left(\exp\left[-H_0(C_i)e^{\tilde{f}_i}\right]\right)^{1-\delta_i}} \\ &= \prod_{i=1}^{n}\left(\frac{\exp\left[-H_0(v_i)e^{f_i}\right]e^{f_i}}{\exp\left[-H_0(v_i)e^{\tilde{f}_i}\right]e^{\tilde{f}_i}}\right)^{\delta_i}\left(\frac{\exp\left[-H_0(C_i)e^{f_i}\right]}{\exp\left[-H_0(C_i)e^{\tilde{f}_i}\right]}\right)^{1-\delta_i}.\end{aligned}$$

Thus, by Taylor expansion up-to order 2, we have

$$\begin{aligned}\log\left(\frac{L_f(v_1,v_2,...,v_n)}{L_{\tilde{f}}(v_1,v_2,...,v_n)}\right) &= \sum_{i=1}^{n}[\delta_i\left\{H_0(v_i)(e^{\tilde{f}_i}-e^{f_i})+(f_i-\tilde{f}_i)\right\}+(1-\delta_i)H_0(C_i) \\ &\quad (e^{\tilde{f}_i}-e^{f_i})] \\ &= \sum_{i=1}^{n}\left[\delta_i H_0(v_i)e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}+(1-\delta_i)H_0(C_i)e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\right] \\ &= \sum_{i=1}^{n}e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\left[\delta_i H_0(v_i)+(1-\delta_i)H_0(C_i)\right],\end{aligned}$$

Next, consider the expectation from (7.1),

$$\begin{aligned}&E\sqrt{\frac{L_f(V_1,V_2,...,V_n)}{L_{f^*}(V_1,V_2,...,V_n)}} \\ &= \prod_{i=1}^{n}(\int_0^{C_i}\sqrt{\frac{\exp\left[-H_0(v_i)e^{f_i}\right]e^{f_i}}{\exp\left[-H_0(v_i)e^{f_i^*}\right]e^{f_i^*}}}h_0(v_i)\exp\left[-H_0(v_i)e^{f_i^*}\right]e^{f_i^*}dv_i + \\ &\sqrt{\frac{\exp\left[-H_0(C_i)e^{f_i}\right]}{\exp\left[-H_0(t_i)e^{f_i^*}\right]}}\exp\left[-H_0(C_i)e^{f_i^*}\right]) \\ = &\prod_{i=1}^{n}\left\{\int_0^{C_i}h_0(v_i)\exp\left[-\frac{H_0(v_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]e^{\frac{f_i+f_i^*}{2}}dv_i+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]\right\} \\ = &\prod_{i=1}^{n}\left\{\frac{2e^{\frac{f_i+f_i^*}{2}}}{e^{f_i}+e^{f_i^*}}\left[1-\exp\left(-H_0(C_i)\frac{e^{f_i}+e^{f_i^*}}{2}\right)\right]+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]\right\}.\end{aligned} \tag{7.2}$$

Hence, after some algebra the ratio of the expectation in (7.1) reduces to

$$\begin{aligned}
&\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}} \\
= &\prod_{i=1}^{n}\frac{\frac{2e^{\frac{f_i+f_i^*}{2}}}{e^{f_i}+e^{f_i^*}}\left[1-\exp\left(-H_0(C_i)\frac{e^{f_i}+e^{f_i^*}}{2}\right)\right]+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{f_i}+e^{f_i^*}\right)\right]}{\frac{2e^{\frac{\tilde{f}_i+f_i^*}{2}}}{e^{\tilde{f}_i}+e^{f_i^*}}\left[1-\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)\right]+\exp\left[-\frac{H_0(C_i)}{2}\left(e^{\tilde{f}_i}+e^{f_i^*}\right)\right]} \\
= &\prod_{i=1}^{n}\left\{1+\frac{f_i-\tilde{f}_i}{2}+\frac{(f_i-\tilde{f}_i)^2}{8}\right\}\left(\frac{e^{f_i}+e^{f_i^*}+(\tilde{f}_i-f_i)e^{f_i}}{e^{f_i}+e^{f_i^*}}\right) \\
&\left\{\frac{1-\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)+(f_i-\tilde{f}_i)\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)H_0(C_i)\frac{e^{f_i}}{2}}{1-\exp\left(-H_0(C_i)\frac{e^{\tilde{f}_i}+e^{f_i^*}}{2}\right)}\right\} \\
= &\prod_{i=1}^{n}\left\{1+\frac{(f_i-\tilde{f}_i)^2}{8}\right\}. \qquad (7.3)
\end{aligned}$$

We expand f_i around $\tilde{f}_i$ up-to first order by Taylor series, and since $\tilde{f}_i \xrightarrow{p} f_i$ then by the fact that for x close to 0, $e^x = 1+x+\frac{x^2}{2}$. Taking $\log$ on both sides of (7.3) we get

$$\begin{aligned}
\log\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}} &= \sum_{i=1}^{n}\log\left\{1+\frac{(f_i-\tilde{f}_i)^2}{8}\right\} \\
&= \sum_{i=1}^{n}\frac{(f_i-\tilde{f}_i)^2}{8}. \qquad (7.4)
\end{aligned}$$

As a result, the second expression in (7.1) may be approximated as

$$-2\log\left(\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}\right) = -\sum_{i=1}^{n}\frac{(f_i-\tilde{f}_i)^2}{4}. \qquad (7.5)$$

Hence, together with (7.2) and (7.5) the condition (7.1) is equivalent to

$$\begin{aligned} pen(f) &\geq \sum_{i=1}^{n}\left[e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\{\delta_i H_0(v_i)+(1-\delta_i)H_0(C_i)\}-\frac{(\tilde{f}_i-f_i)^2}{4}\right]+2\mathscr{L}(\tilde{f}) \\ &= \sum_{i=1}^{n}\frac{(\tilde{f}_i-f_i)^2}{2}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]+2\mathscr{L}(\tilde{f}). \end{aligned} \tag{7.6}$$

Using the facts that $(\tilde{f}_i-f_i)^2 \xrightarrow{p} E(\tilde{f}_i-f_i)^2 = Var(\tilde{f}_i)$ and this variance has an upper bound $\frac{UU_f}{K}$ i.e. $Var(\tilde{f}_i) \leq \frac{UU_f}{K}$. This upper bound together with the fact that $\mathscr{L}(\tilde{f}) = K\log 2p$ yield an upper bound for the right-hand side of (7.6). Replacing these in (7.6) we obtain

$$pen(f) \geq \frac{UU_f}{2K}\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]+2K\log 2p. \tag{7.7}$$

Differentiating (7.7) with respect to K and then equating it to zero we get

$$K=\frac{\sqrt{UU_f}}{2\sqrt{\log 2p}}\sqrt{\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}.$$

Then, replacing that value of K in (7.7) with the choice of $U = U_f$ we get

$$\begin{aligned} pen(f) &\geq 2\sqrt{UU_f\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}\sqrt{\log 2p} \\ \Rightarrow \lambda U_f &\geq 2\sqrt{UU_f\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}\sqrt{\log 2p} \end{aligned}$$

which is equivalent to

$$\frac{\lambda}{n} \geq 2\frac{\sqrt{\log 2p}}{n}\sqrt{\sum_{i=1}^{n}\left[\delta_i\left(e^{f_i}H_0(v_i)-\frac{1}{2}\right)+(1-\delta_i)\left(e^{f_i}H_0(C_i)-\frac{1}{2}\right)\right]}. \tag{7.8}$$

This completes the proof of the theorem.
There is a chance that the sum in (7.7) can be negative. Then, we cannot proceed further with that negative sum. In that situation, we adopt a slightly modified route

to overcome this difficulty. From (7.5) we have

$$-2\log\left(\frac{E\sqrt{\frac{L_f(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}{E\sqrt{\frac{L_{\tilde{f}}(V_1,V_2,\ldots,V_n)}{L_{f^*}(V_1,V_2,\ldots,V_n)}}}\right) \approx -\sum_{i=1}^{n}\frac{(f_i-\tilde{f}_i)^2}{4}\leq 0. \tag{7.9}$$

With the bound in (7.9) and by the facts that $(\tilde{f}_i-f_i)^2 \xrightarrow{p} E(\tilde{f}_i-f_i)^2 = Var(\tilde{f}_i) \leq \frac{UU_f}{K}$ and $\mathscr{L}(\tilde{f}) = K\log 2p$, the condition (7.1) reduces to

$$\begin{aligned} pen(f) &\geq \sum_{i=1}^{n}\left[e^{f_i}\frac{(\tilde{f}_i-f_i)^2}{2}\{\delta_i H_0(v_i)+(1-\delta_i)H_0(C_i)\}\right]+2\mathscr{L}(\tilde{f}) \\ &= \frac{UU_f}{2K}\sum_{i=1}^{n}\left[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)\right]+2K\log 2p. \end{aligned} \tag{7.10}$$

Then, we minimize the right-hand side of (7.10) with respect to K and choose $U = U_f$ as before. Now, the equation (7.10) reduces to

$$\begin{aligned} pen(f) &\geq 2\sqrt{UU_f\sum_{i=1}^{n}[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)]}\sqrt{\log 2p} \\ \Rightarrow \lambda U_f &\geq 2\sqrt{UU_f\sum_{i=1}^{n}[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)]}\sqrt{\log 2p} \\ \Rightarrow \frac{\lambda}{n} &\geq 2\frac{\sqrt{\log 2p}}{n}\sqrt{\sum_{i=1}^{n}[\delta_i e^{f_i}H_0(v_i)+(1-\delta_i)e^{f_i}H_0(C_i)]}. \end{aligned} \tag{7.11}$$

From (7.11) it is clear that the penalty function is still information theoretically valid since it satisfies the condition (7.1).

Proof of Result 2:

Let $e_k = \frac{1}{n}\log\frac{L_{\hat{f}}(x)}{L_{\hat{f}_{(k)}}(x)} + \lambda(u_{(k)} - U_f)$. Then, using the full likelihood we get

$$\begin{aligned} e_k &= \frac{1}{n}\log\frac{L_{\hat{f}}(x)}{L_{\hat{f}_{(k)}}(x)} + \lambda(u_{(k)} - U_f) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left[\left(\frac{p_{\hat{f}}(x_i)}{p_{\hat{f}_{(k)}}(x_i)}\right)^{\delta_i}\left(\frac{\bar{P}_{\hat{f}}(C_i)}{\bar{P}_{\hat{f}_{(k)}}(C_i)}\right)^{1-\delta_i}\right] + \lambda(u_{(k)} - U_f) \\ &= \frac{1}{n}\sum_{i=1}^{n}\log\left[\delta_i\left\{\hat{f}_i - \hat{f}_{i,(k)} + H_0(v_i)(e^{\hat{f}_{i,(k)}} - e^{\hat{f}_i})\right\} + (1-\delta_i)\log\frac{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)}\right] \\ &\quad + \lambda(u_{(k)} - U_f), \end{aligned} \tag{7.12}$$

where $\hat{f}_i = Z_i'\hat{\beta}$ and $\hat{f}_{i,(k)} = Z_i'\hat{\beta}_{(k)}$ with $\hat{\beta}_{(k)}$, obtained at k^{th} iteration, is the estimate of β. To prove the theorem we need to show that

$$e_k \leq (1-\alpha)e_{k-1} + \frac{1}{2}\alpha^2 U_f^2. \tag{7.13}$$

It is clear that to have the inequality (7.13), we only need to tackle the ratio of the survival functions from (7.12). For the i^{th} subject we rewrite the ratio of the survival functions from (7.12) in the following way

$$\begin{aligned} &\frac{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)} \\ &= \left[\frac{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)}\right]^{\bar{\alpha}} \frac{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\alpha}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\bar{\alpha}}}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)}. \end{aligned} \tag{7.14}$$

So, to prove (7.13) we need to show

$$\frac{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\alpha}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\bar{\alpha}}}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)} \leq 1. \tag{7.15}$$

Then, using the updating rule that $\hat{f}_{i,(k)} = \bar{\alpha}\hat{f}_{i,(k-1)} + \gamma Z_{il}$ we rewrite (7.15) in the following way

$$\begin{aligned} & \frac{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\alpha}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\bar{\alpha}}}{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k)}}\right)} \\ = & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\bar{\alpha}}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\alpha}\exp\left(-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\gamma Z_{il}}\right)} \\ . & \end{aligned} \tag{7.16}$$

We choose customarily some α and $\gamma = \alpha U_f$ in such a way that $\gamma Z_{il} \xrightarrow{p} \alpha f_i$, which is estimated by $\alpha\hat{f}_i$. For more details regarding the customary choices and the convergence see [2]. Using these facts (7.16) reduces to

$$\begin{aligned} & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\left\{\exp\left(-H_0(C_i)e^{\hat{f}_i}\right)\right\}^{\bar{\alpha}}\left\{\exp\left(-H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}^{\alpha}\exp\left(-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}\right)} \\ = & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\left\{\exp\left(-\bar{\alpha}H_0(C_i)e^{\hat{f}_i}\right)\right\}\left\{\exp\left(-\alpha H_0(C_i)e^{\hat{f}_{i,(k-1)}}\right)\right\}\exp\left(-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}\right)} \\ = & \frac{\exp\left(-H_0(C_i)\left\{e^{\hat{f}_i}+e^{\hat{f}_{i,(k-1)}}\right\}\right)}{\exp\left(-\bar{\alpha}H_0(C_i)e^{\hat{f}_i}-\alpha H_0(C_i)e^{\hat{f}_{i,(k-1)}}-H_0(C_i)e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}\right)}. \end{aligned} \tag{7.17}$$

We denote the numerator and denominator of (7.17) by D_n and D_e, respectively. We see that for $\alpha = 0$ and 1, $D_n = D_e$ which reduces (7.17) to 1 and hence log of (7.17) becomes 0. For $\alpha \in (0,1)$ we study the nature of D_e. We have

$$\frac{\partial \log D_e}{\partial \alpha} = H_0(C_i)\left(e^{\hat{f}_i} - e^{\hat{f}_{i,(k-1)}} - e^{\bar{\alpha}\hat{f}_{i,(k-1)}+\alpha\hat{f}_i}(\hat{f}_i - \hat{f}_{i,(k-1)})\right)$$

and

$$\frac{\partial^2 \log D_e}{\partial \alpha^2} = -H_0(C_i)\exp\left\{\bar{\alpha}\hat{f}_{i,(k-1)} + \alpha\hat{f}_i\right\}(\hat{f}_i - \hat{f}_{i,(k-1)})^2 \tag{7.18}$$

From (7.18) it is clear that $\log D_e$ and hence, D_e is strictly concave function. As a result, D_e cannot attain its maximum for $\alpha = 0$ or 1 since in that case D_e will be a constant. So, (7.17) is less than or equal to 1 for $0 < \alpha < 1$, indicating that its log is negative. This completes the proof of the result.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 315–330, DOI 10.21307/stattrans-2018-018

COHORT PATTERNS OF FERTILITY IN POLAND BASED ON STAGING PROCESS – GENERATIONS 1930-1980

Wioletta Grzenda[1], **Ewa Frątczak**[2]

ABSTRACT

As a transition country in the region of Central and Eastern Europe, Poland has experienced unprecedented changes in the fertility. Currently, the total fertility rate level is very low, ca. 1.3 children per woman, which is below the replacement level. Many studies have described changes in fertility based on the cross-sectional approach. However, the changes of cohort fertility have been described not quite sufficiently. Our paper complements this gap by the assessment of stochastic fertility tables, calculated for five-year generations of women born in the period 1930-1980. The main goal of this study is to analyse changes in the cohort patterns of female fertility in Poland.

Key words: cohort fertility, stochastic process, stage probabilities.

1. Introduction

Fertility behaviour of women is influenced by numerous factors and is constantly changing over time, therefore fertility tables are the best tool for their analysis. Fertility tables are derived from life tables, which are one of the oldest tools of demographic analysis. The contemporary methodology of constructing life tables, based on the probability theory, was introduced by C.L. Chiang in 1968 (Chiang, 1968). He was also one of the first authors of stochastic fertility tables (Chiang, 1984). There are also numerous publications on this subject in Polish (Frątczak and Ptak-Chmielewska, 2011a; 2011b; Frątczak, 1996; Bolesławski, 1974) and foreign literature (Cigno, 1994; Namboodiri, 1991; Chiang and Van Den Berg, 1982; Namboodiri and Suchindran, 1987). The bases for constructing such tables are stochastic processes, because the events from a life course of a given individual, such as for example births, can be treated as the implementation of these processes. Therefore, often tables constructed in this manner are called stochastic tables. The events considered are localized in time, therefore the stochastic fertility tables can be used for analysing the changes of the level and pace of this phenomenon.

[1] Warsaw School of Economics, Collegium of Economic Analysis, Institute of Statistics and Demography. Poland. E-mail: wgrzend@sgh.waw.pl.
[2] Warsaw School of Economics, Collegium of Economic Analysis, Institute of Statistics and Demography. Poland. E-mail: ewaf@sgh.waw.pl.

The main aim of this paper is the analysis of cohort patterns changes of female fertility in Poland using stochastic fertility tables. There are various methods of constructing such stochastic fertility tables. The most popular and the least complicated approach is to treat the birth of each child as a single event and investigate single episode model. In this paper stochastic fertility tables have been constructed based on staging process (Willekens, 1991) using multi episode models. The construction of such tables necessitates taking into account the time of waiting for an event, which is the birth of a child. This requires the use of event history analysis methods (Graunt, 1962; Liu, 2012) in the modelling. This approach made it possible to comprehensively analyse successive births as sequences of events. The cohort approach is very common in fertility behaviour studies, because it allows the researchers to compare the fertility of each generation of women in a particular moment of their lives. One publication (Frejka and Calot, 2001), describing the fertility patterns in low fertility countries, studied women born in the years 1930-1970. In the majority of the 27 countries analysed, the total fertility rate decreased for each subsequent birth cohorts. In order to reverse this trend, women who are at the beginning of their fertility period should adopt vastly different fertility patterns than women born in 1960s and 1970s. The analysis of fertility patterns for groups of women born in the same year can be also found in many other publications (Frejka and Sardon, 2004; Sobotka, 2003).

Properly constructed fertility tables provide important information regarding women's fertility behaviour. They allow us to answer questions such as: when did a given woman give birth, at what age, and how many children does she have. A complementary character of such information is very important; therefore studies offer various methods of supplementing the missing information for cohort fertility schedules (Cheng and Lin, 2010).

This paper aims to analyse the fertility behaviour of women in Poland, based on the stochastic fertility tables constructed for five-year generations, from 1931-1935 to 1976-1980. Five-year age groups are widely used in the studies of female fertility (for example: Lee, 1974), which makes our results comparable with the results of other similar works. The data used for constructing the fertility tables come from "Fertility of Women" study conducted along with the National Census of Population and Housing in Poland in 2002. The basis for this study, received from the Central Statistical Office, was created for the "Epidemiology of fertility dangers in Poland – multi-centre, prospective cohort study" research project by pairing the information from Form D "Women's fertility" with selected results from Form A "National Census of Population and Housing 2002" [3]. The results of

[3] Research project: Epidemiology of fertility dangers in Poland – multi-centre, prospective cohort study / Ministry of Science and Higher Education ordered grant, decision K 140/P01/2007, Repro_PL, project director: Professor Wojciech Hanke, MD, PH.D., J. Nofer Occupational Medicine Institute in Łódź. Project implementation: 2007 – 2011.

Within the abovementioned project framework, the Event History and Multilevel Analysis Unit of the Institute of Statistics and Demography conducted two research assignments:

Research Assignment 1.1.1 *Demographic and socio-economic reasons for low fertility and total fertility rate in Poland (postponing the childbearing decisions – descriptive and modelling analyses). Past, present, perspectives.*

Research assignment 1.1.2 Late fertility and childbearing diagnosis (postponing the childbearing decisions; plans and preferences – cohort prospective study (quantitative and qualitative) of demographic, socio-economic and health factors. For the purpose of research assignment 1.1.1, the research team received the relevant National Census of 2002 data from the Central Statistical Office.

descriptive fertility tables for single cohorts 1911-1986, where the birth of each child is treated as a single-episode process and not a staging process, based on the results of "Women's Fertility" study conducted along the Census of 2002, are included in the work (Frątczak and Ptak-Chmielewska, 2011a, 2011b). Moreover, the results of the preliminary analysis of this data set are contained in (Frątczak and Grzenda, 2011). This text is a continuation of the latter research on the subject of cohort fertility based on the "Women's Fertility" data.

The scope of the analysis conducted allows us to verify numerous research hypotheses regarding the changes of fertility behaviour of Polish women after World War II. The reasons for these changes can be linked to the Second Demographic Transition (Van de Kaa, 1987), which relates to the demographic changes from the beginning of 1990s in Central-Eastern Europe. Regarding fertility, these changes are characterized mostly by a decrease in fertility rate and postponing the decision of first birth.

Analysing births in Poland (Bolesławski, 1974; 1975; Paradysz, 1992) we can conclude that the baby boom in 1970s and 1980s is an echo of the post-war baby boom of 1950s. Taking into consideration other publications dealing with the issue of fertility analysis we have posed two research hypotheses. The oldest birth cohorts: 1931-1935 and 1936-1940 are characterized by the largest probability of forth and subsequent births. The 1951-1955 and 1956-1960 birth cohorts exhibit a high staging probability of second and third births. At the end of the 20th century we saw the lowest level of fertility in virtually every European country (Frejka and Sardon, 2004). In this study we will verify the hypothesis that the youngest birth cohorts: 1971-1975 and 1976-1980, are characterized by the greatest probability of remaining childless and the lowest rate of successive births. The calculations of stochastic fertility tables for investigated cohorts not only allow verifying these hypotheses but also allow determining the exact differences in the fertility for these cohorts.

2. Research Method

The majority of methods used for population research are based on the probability theory because births, migrations or mortality for each individual can be considered an event. Each event is a transition from one state into another, and each individual at risk of each event has a certain positive probability of experiencing the transition between states (Namboodiri, 1991). We are interested not only in the time of occurrence of a given event, but also how often this event occurred during an individual's life and what the probability of such event occurring in a given timeframe is. The analysis of random events is based on the analysis of random variables. These variables are indexed with a certain parameter, interpreted as time; therefore the result is a stochastic process (Chiang, 1980).

The staging process is a sequence of certain events, generated by a random mechanism, bearing in mind that a multiple occurrence of the same event at the same time is impossible. It differs from the Markov process in that it has a certain sequence and is irreversible, while in Markov process there are recurring stages. Examples of staging processes in survivability analysis can be found in (Chiang,

1985), in medical applications and in fields related to fertility. Therefore, an example of a staging process is the fertility process:

0 *children*→ 1 *child* → 2 *child* → 3 *child* → … → *n–th child*.

An elementary process is a process which generates one event – the occurrence of the event in a given time is the end of the elementary process. A chain of independent elementary processes defines a staging process: the first elementary process generates the first event and triggers the second elementary process, which in turn generates the second event, etc.

In order to describe the staging process we assume the following notation:

μ_n – intensity of the n-th event,

X_n – the time of occurrence of the *n*-th event; the times are ordered increasingly: $X_0 < X_1 < X_2 < \ldots$,

U_n – the period of waiting for the occurrence of the *n*-th event: $U_n = X_n - X_{n-1}$,

N_x – the number of events.

The schema of staging process with 5 events can be defined as follows:

The process intensity:

$$0 \xrightarrow{\mu_1} 1 \xrightarrow{\mu_2} 2 \xrightarrow{\mu_3} 3 \xrightarrow{\mu_4} 4 \xrightarrow{\mu_5} 5$$

Waiting time:

$$X_0 \xrightarrow{U_1} X_1 \xrightarrow{U_2} X_2 \xrightarrow{U_3} X_3 \xrightarrow{U_4} X_4 \xrightarrow{U_5} X_5$$

Please note that the waiting period for the first event, which is the birth of the first child, is a random variable with no memory parameter. This parameter, also known as the Markov's parameter, is characteristic of an exponential distribution. After the first event has occurred everything starts again. Therefore, we assume that successive periods of waiting for childbirth are independent random variables with exponential distribution.

The theoretical basis for constructing fertility table is the Poisson process, where the times between two successive events (waiting periods) U_n are independent random variables with exponential distributions, but with different parameters.

In the analysed process, the measure that has been found is the intensity (hazard) function. For the *n*-th event it is expressed by the formula:

$$\mu_n = \lim_{\Delta x \to 0} \frac{P\left(x < X_n \le x + \Delta x \mid X_n \ge x\right)}{\Delta x}.$$

It is the probability of the occurrence of the *n*-th event within the time interval, assuming that the event with the number *n*-1 had occurred before the process reached time *x*.

Distribution function of the waiting period for *n*+1 event is not dependant on the occurrence of the n-th event:

$$F_{n+1}(u) = P(U_{n+1} \le u) = 1 - \exp(-\mu_{n+1}u) = 1 - \exp(-\mu_{n+1}u).$$

The density function for the time period between *n* and *n+1* event is given as:

$$f_{n+1}(u) = P(U_{n+1} = u) = \frac{\partial P(U_{n+1} \le u)}{\partial u} = \mu_{n+1}\exp(-\mu_{n+1}u).$$

The probability that the event does not occur before the time *x,* i.e. survival function, is: $P(N_x = 0) = P(X_1 > x) = \exp(-\mu_1 x).$

The probability that the process being at the first stage at the time *x* is exposed to the risk of experiencing the first event is called stage probability and labelled as: $S_1(x)$.

The probability that the event occurs exactly once in the interval $(0, x)$ is denoted by $P(N_x = 1)$. It is stage probability $S_2(x)$, which means that the process at the time *x* is at the stage 2. If the dependant events process $\mu_1 \neq \mu_2$, then the probability is determined as follows:

$$S_2(x) = P(N_x = 1) = \frac{\mu_1}{\mu_1 - \mu_2}\left(\exp(-\mu_2 x) - \exp(-\mu_1 x)\right).$$

The probability that the event occurs exactly *n*-1 times within the $(0, x)$ interval is calculated as:

$$S_n(x) = P(N_x = n-1) = (-1)^{n-1}\left[\prod_{j=1}^{n-1}\mu_j\right]\sum_{j=1}^{n}\frac{\exp(-\mu_j x)}{\prod_{\substack{k=1 \\ k\neq j}}^{n}(\mu_j - \mu_k)}.$$

While constructing the fertility tables we calculate such staging probabilities for successive births. In the staging process analysis, we quite often identify the moment from which we start measuring the duration of a given process – this study assumes a period of 15 years, and the time axis covers the following age of women: 15 – 49 years.

Based on the formulas presented, using the exponential distribution, we have calculated the hazard value for each event μ_n, $n = 1,2,\ldots,5$ and then the staging probability values for successive births. This study presents the staging probabilities for five-year age groups: 15-19, 20-24,…, 45-49.

3. Estimation Results

Based on the formulas presented in the previous section we have created an original program for calculating the characteristics of fertility tables. The estimation of all models was conducted in SAS systems. The estimation of fertility tables was conducted taking into consideration weights, therefore the results may be generalized for the entire population of women. In this chapter we present and interpret the resulting parameters of stochastic fertility tables.

In tables 1-7 we include the staging probabilities of successive births for five-year age groups of women: 15-19, 20-24,…, 45-49, determined for five-year generations, from 1931-1936 to 1976-1980.

Table 1. The stage probabilities of successive births – women aged 15-19

Birth cohorts	0 children	1 child	2 children	3 children
1931-1935	0.8696	0.0941	0.0307	0.0051
1936-1940	0.8536	0.1080	0.0349	0.0031
1941-1945	0.8471	0.1161	0.0306	0.0057
1946-1950	0.8757	0.0972	0.0239	0.0030
1951-1955	0.8921	0.0844	0.0210	0.0024
1956-1960	0.8836	0.0869	0.0259	0.0033
1961-1965	0.8685	0.0969	0.0288	0.0039
1966-1970	0.8625	0.0998	0.0323	0.0052
1971-1975	0.8681	0.1005	0.0274	0.0034
1976-1980	0.8994	0.0817	0.0171	0.0016

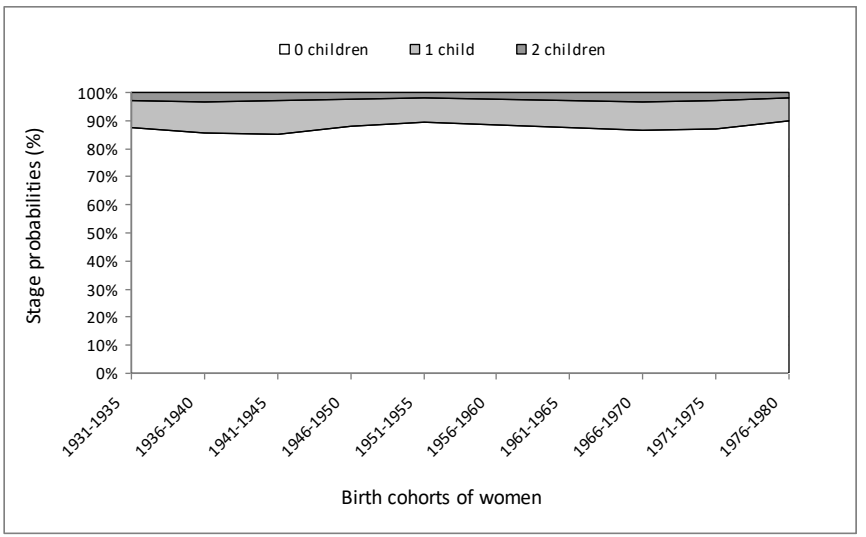


Figure 1. Distribution of stage probabilities of births – women aged 15-19

In Table 1, we present the stage probabilities for women aged 15-19. The stage probabilities of successive births for all generations are similar (Figure 1). In the case of first births, the highest probability is observed for 1936-1940, 1941-1945 and 1971-1975 cohorts.

Table 2 includes the values of stage probabilities for women aged 20-24. The youngest cohorts: 1971-1975 and 1976-1980 are the cohorts with the highest percentage of childless women, which is clearly visible in Figure 2. The stage probability values for first births for all generations are similar, peaking for generation 1946-1950 and 1971-1975. The stage probabilities of second births for generations from 1931-1940 to 1966-1970 are similar, reaching the highest value for 1961-1965 cohort, while the probability value is lowest for the two youngest cohorts. It is important to note a steady drop in higher-order births for each subsequent birth cohort.

Table 2. The stage probabilities of successive births – women aged 20-24

Birth cohorts	0 children	1 child	2 children	3 children	4 children	5+ child.
1931-1935	0.4462	0.1849	0.2033	0.1061	0.0412	0.0183
1936-1940	0.4284	0.2077	0.2264	0.0956	0.0319	0.0100
1941-1945	0.4252	0.2380	0.2199	0.0835	0.0252	0.0082
1946-1950	0.4471	0.2477	0.2086	0.0703	0.0201	0.0062
1951-1955	0.4478	0.2363	0.2221	0.0709	0.0173	0.0056
1956-1960	0.4320	0.2194	0.2355	0.0855	0.0209	0.0067
1961-1965	0.4137	0.2154	0.2485	0.0895	0.0266	0.0063
1966-1970	0.4299	0.2289	0.2341	0.0788	0.0214	0.0069
1971-1975	0.5029	0.2511	0.1818	0.0502	0.0114	0.0026
1976-1980	0.6680	0.2247	0.0887	0.0163	0.0021	0.0002

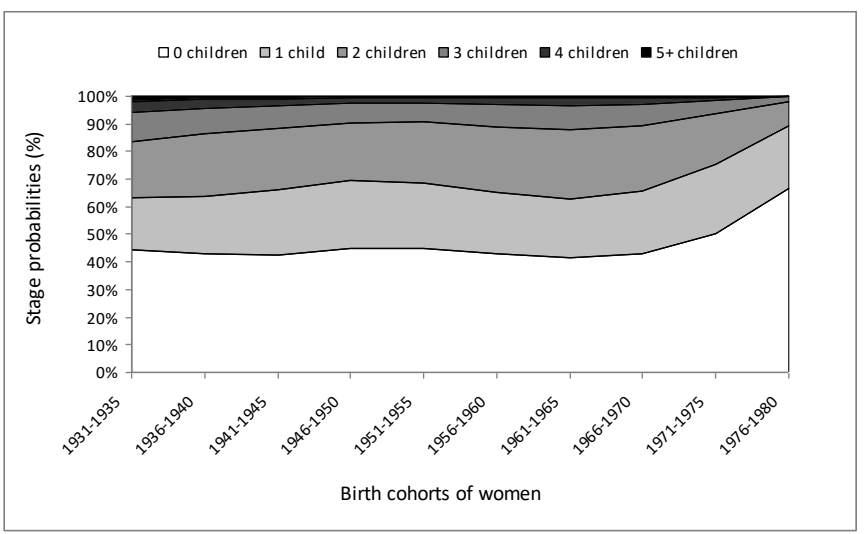


Figure 2. Distribution of stage probabilities of births – women aged 20-24

Table 3 presents the values of stage probability for women aged 25-29. The highest probability of remaining childless is observed for the youngest birth cohort: 1971-1975, but at the same time this cohort is characterized by the highest probability of first births. A higher probability of first births is also visible for generations 1941-1945, 1946-1950 and 1966-1970. When it comes to second births, they are at a similar level, bearing in mind that they are the lowest for the youngest and oldest cohorts, and the highest for 1956-1960 and 1961-1965 cohorts. We can also see that the oldest cohort has the highest value of third births, while this probability is lowest for the youngest cohort (Figure 3).

Table 3. The stage probabilities of successive births – women aged 25-29

Birth cohorts	0 children	1 child	2 children	3 children	4 children	5+ child.
1931-1935	0.2584	0.1560	0.2694	0.1831	0.0838	0.0493
1936-1940	0.2468	0.1791	0.3070	0.1662	0.0639	0.0370
1941-1945	0.2453	0.2133	0.3085	0.1487	0.0569	0.0273
1946-1950	0.2536	0.2123	0.3229	0.1409	0.0487	0.0216
1951-1955	0.2496	0.1976	0.3357	0.1455	0.0495	0.0221
1956-1960	0.2374	0.1734	0.3515	0.1616	0.0519	0.0242
1961-1965	0.2257	0.1778	0.3503	0.1655	0.0565	0.0242
1966-1970	0.2434	0.2123	0.3404	0.1443	0.0427	0.0169
1971-1975	0.3270	0.2828	0.2841	0.0812	0.0192	0.0057

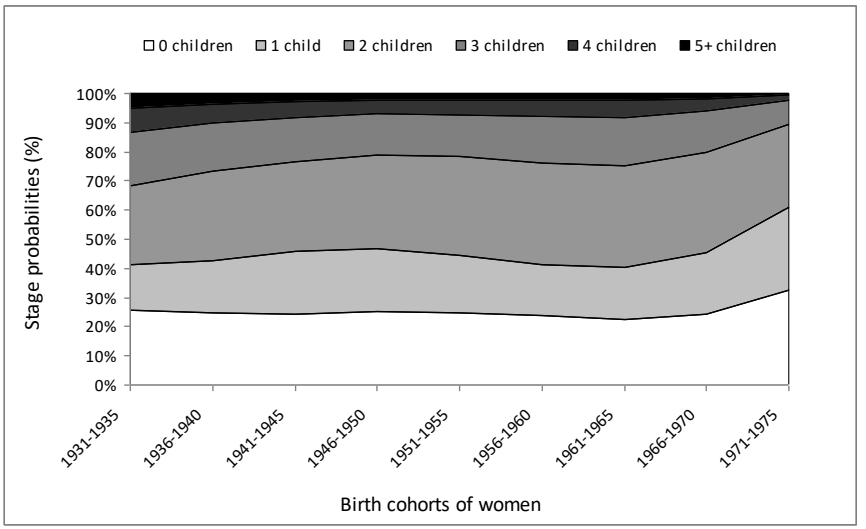


Figure 3. Distribution of stage probabilities of births – women aged 25-29

Analysing Table 4 and Figure 4 we can see that for women aged 30-34 the values of each probability are similar for "middle" cohorts. It is worth noting that the highest value of fourth births is observed for the oldest cohort. There is also a significantly lower probability of first births in favour of second births for generations 1956-1960 and 1961-1965, as well as a higher value of first births probability for cohorts 1941-1945, 1946-1950 and 1966-1970.

Table 4. The stage probabilities of successive births – women aged 30-34

Birth cohorts	0 children	1 child	2 children	3 children	4 children	5+ child.
1931-1935	0.1718	0.1289	0.3027	0.2230	0.1080	0.0656
1936-1940	0.1624	0.1501	0.3485	0.2025	0.0872	0.0493
1941-1945	0.1576	0.1803	0.3559	0.1925	0.0741	0.0396
1946-1950	0.1631	0.1788	0.3744	0.1858	0.0654	0.0325
1951-1955	0.1588	0.1601	0.3900	0.1930	0.0657	0.0324
1956-1960	0.1499	0.1409	0.4021	0.2053	0.0691	0.0327
1961-1965	0.1403	0.1450	0.4045	0.2075	0.0713	0.0314
1966-1970	0.1575	0.1942	0.4078	0.1754	0.0482	0.0169

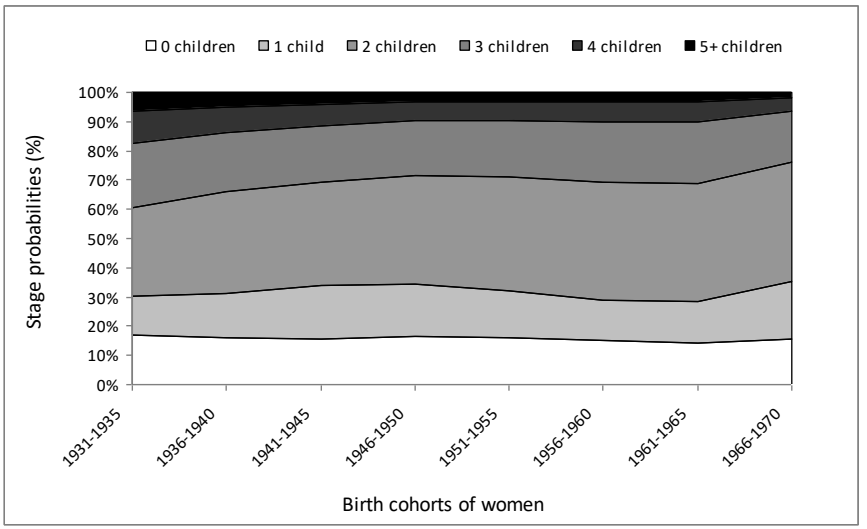


Figure 4. Distribution of stage probabilities of births – women aged 30-34

Analysing the values of stage probability of successive births for women aged 35-39, based on Table 5 and Figure 5, we can observe the highest probability of second births. However, for subsequent cohorts, beginning with the oldest one, we observe a drop in the number of higher-order births.

Table 5. The stage probabilities of successive births – women aged 35-39

Birth cohorts	0 children	1 child	2 children	3 children	4 children	5+ child.
1931-1935	0.1202	0.1116	0.3264	0.2542	0.1195	0.0681
1936-1940	0.1117	0.1313	0.3793	0.2265	0.0988	0.0524
1941-1945	0.1080	0.1596	0.3922	0.2175	0.0815	0.0412
1946-1950	0.1107	0.1593	0.4119	0.2127	0.0718	0.0336
1951-1955	0.1063	0.1382	0.4319	0.2166	0.0744	0.0326
1956-1960	0.0996	0.1206	0.4379	0.2323	0.0759	0.0337
1961-1965	0.0925	0.1252	0.4476	0.2317	0.0738	0.0292

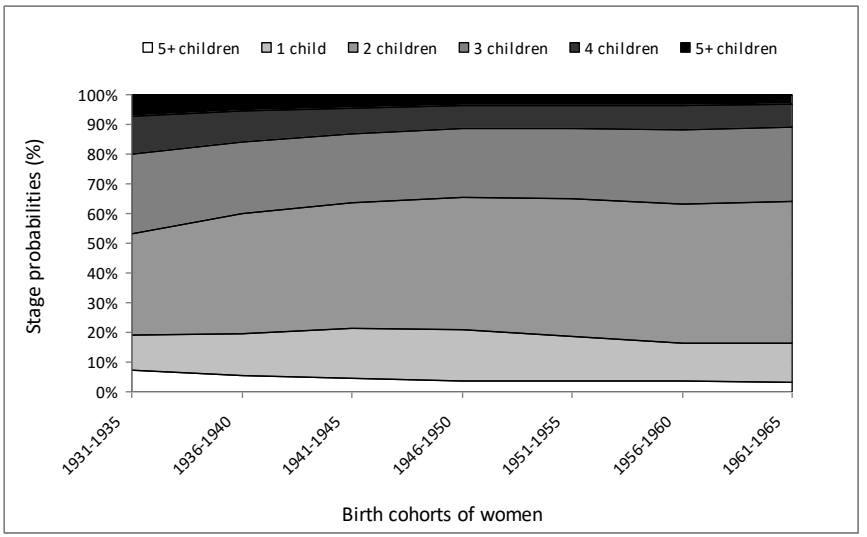


Figure 5. Distribution of stage probabilities of births – women aged 35-39

The results presented in Tables 6 and 7 are not significantly different than the ones presented in Table 5. Therefore, the figures for the women aged 40-44 and aged 44-49 have been omitted. We can observe that the percentage of childless women is very similar for all generations analysed, and the most common number of births is 2.

Table 6. The stage probabilities of successive births – women aged 40-44

Birth cohorts	0 children	1 child	2 children	3 children	4 children	5+ child.
1931-1935	0.0870	0.0992	0.3486	0.2764	0.1249	0.0639
1936-1940	0.0799	0.1182	0.4076	0.2451	0.1017	0.0475
1941-1945	0.0769	0.1467	0.4250	0.2332	0.0814	0.0368
1946-1950	0.0781	0.1468	0.4462	0.2289	0.0712	0.0288
1951-1955	0.0742	0.1240	0.4675	0.2321	0.0743	0.0279
1956-1960	0.0687	0.1063	0.4719	0.2490	0.0756	0.0285

Table 7. The stage probabilities of successive births – women aged 45-49

Birth cohorts	0 children	1 child	2 children	3 children	4 children	5+ child.
1931-1935	0.0648	0.0895	0.3664	0.2937	0.1271	0.0585
1936-1940	0.0592	0.1081	0.4295	0.2584	0.1023	0.0425
1941-1945	0.0568	0.1359	0.4497	0.2446	0.0808	0.0322
1946-1950	0.0569	0.1367	0.4725	0.2396	0.0696	0.0247
1951-1955	0.0534	0.1123	0.4953	0.2427	0.0727	0.0236

Based on the presented stage probability results for successive birth cohorts we can single out two cohorts: the post-war baby boom cohort 1951-1955 and the younger analysed cohort 1971-1975. For birth cohort 1971-1975 we can see that the percentage of childless women is larger than in the 1951-1955 cohort. Moreover, we can observe that the number of successive births is getting lower. The greatest value of state probabilities for first births for women aged 20-24 was obtained for birth cohort 1971-1975. This value for cohort 1951-1955 is similar to other cohorts. Similarly for women aged 25-29, but this difference between birth cohort 1971-1975 and other cohorts is much deeper. Based on the stage probabilities of second births values for 1951-1955 and 1971-1975 cohorts there is a visible decline for women aged 20-24 and 25-29.

Finally, we compare the stage probability values for 1951-1955 and 1971-1975 cohorts, for third births. For 1951-1955 cohort we can observe an increasing trend for third birth for each age group, starting with women aged 15-19. For the youngest birth cohort 1971-1975 the trends are the total opposite of the ones exhibited by 1951-1955 cohort.

Analysing the stage probability values for first births we can see that the first birth occurs most frequently among women aged 20-24 and 25-29, regardless of

their year of birth. We can see that the highest probability of first births in the age group 25-29 occurs in the youngest generation of women 1971-1975, and the lowest probability occurs in the oldest generation. Moreover, the highest probability of first births for this cohort is also in the age group 20-24. Therefore, it can be concluded that among women aged 20-24, the women of birth cohort 1971-1975 with the highest probability gave birth to the first child, compared to other cohorts. On the other hand, the women of birth cohort 1971-1975 aged 20-24 begun decline in successive births.

5. Conclusions and Discussion

The goal of this study was to analyse the fertility behaviour of women in Poland, based on stochastic fertility tables constructed for 5-year generations from 1931-1935 to 1976-1980. We used stochastic fertility tables based on the staging process. This approach allows considering in modelling the sequence of events, not only each event separately.

Based on the presented results (Tables 1-7), it was found that there is no indication to reject the investigated hypotheses. It follows from the Tables 1-7 that the oldest birth cohorts: 1931-1935 and 1936-1940 are characterized by the largest probability of forth and subsequent births. Moreover, the 1951-1955 and 1956-1960 birth cohorts are characterized by a high staging probability of second and third births. Furthermore, the highest changes in the values of stage probabilities can be observed in the case of the last two generations: 1971-1975 and 1976-1980. They are the result of the reaction to the socio-economic and cultural transformation in Poland after 1989. The results of probability estimation clearly show that in every remaining cohort the probability of higher-order births is decreasing (Figure 1-5). The observed changes in fertility have been conditioned by various economic and socio-cultural factors, including migration. The analysis of these factors requires a different approach and is the subject of analysis performed by Polish and foreign researchers (Kotowska et al., 2008; Olah and Frątczak, 2013; Frejka et al., 2016).

There are various theories which can help explain the transformation changes of cohort fertility patterns in Poland. We must agree with McDonald's theory (McDonald, 2006; 2008), who argued that the emergence of low fertility is associated with two waves of social change that have profound effects upon family formation behaviour in the past 40 years. The first wave of change beginning in the 1960s was an expansion of social liberalism (the so-called reflective modernization) and the second wave beginning in the 1980s was an expansion of economic deregulation, the so-called new capitalism, but the most important is the labour market deregulation. While in the period of socialism in Poland, that is until 1989, these waves could not act with full force, for example because of government regulation of labour market, similarly to other socialist countries, their effect and importance became much more intense from the beginning of the transformation period. This translates into drastic changes in younger analysed cohorts: 1971-1975, 1976-1980. The labour market participation of women and their fertility is also the subject of many studies. In (Kotowska et al., 2008), authors indicated the connections between women's employment and their fertility. They hypothesize that female employment would

have decline more if the women had not reduced their fertility. Moreover, the authors suggest that since the beginning of the 1990s the significance of educational achievements of women and their decision to have a baby have been characterized by a rapid development of higher education in Poland.

All theories related to the demographic changes described by the Second Demographic Transition (Van de Kaa, 1994; 1996; Lesthaeghe, 1991; 1998), that is: (a) the theory of increased female economic autonomy (Becker, 1991), (b) the theory of relative economic deprivation (Easterlin, 1976; 1979), and (c) the theory of ideational shift (Lesthaeghe and Surkyn, 1988; Bumpass, 1990**)** may be useful for explaining the changes in cohort fertility in Poland. On the other hand, in Poland after 1989, the intensity of changes of cohort fertility patterns increased rapidly along with the socio-economic transformation processes, which directly follows from the distribution of stage probabilities of births for women aged 20-24 (Figure 2). Therefore, it is worth agreeing with the opinion of Espanding-Andersen and Billari (Espanding-Andersen and Billari, 2015), who state that because of the high intensity of changes in families in post-transitional societies, the explanation of the changes using the abovementioned three theories related to the Second Demographic transition is no longer sufficient. Similarly to many other transformation countries from Central and Eastern Europe, Poland is experiencing the process of both cohort and cross-section transformation, which is determined by numerous phenomena and processes.

The changes in actual cohorts translated into the changes of cross-sectional fertility and fertility rates, which is reflected in the low values of cross-sectional fertility rates. Our results (Figure 3) indicate that for women aged 25-29 a decline in the stage probability of second births takes place; at the same time, the stage probability of first birth is getting higher. This is consistent with other studies showing that there are significant changes in the nuclear family model in Poland (Frątczak, 2001; Frątczak and Kozłowski, 2005). During the transformation period in Poland the model of nuclear family changed from two-child model into one-child model, with a high percentage of childless families in the general structure. These changes were explained by some researchers using the effect of shifting, which in developed countries has been observed for cohorts of women born after World War II (Sobotka et al., 2012). However, more recent analysis of 15 Central and East European (CEE) countries, including Poland, confirms these tendencies (Frejka et al., 2016) and shows that despite the growth in fertility rates in the late 2000s, the fertility still remains at a low level.

REFERENCES

BECKER, G., (1991). A Treatise on the Family, Enlarged Edition, Cambridge: Harvard University Press.

BOLESŁAWSKI, L., (1975). Marriage and childbearing probability (cohort life tables), Warsaw: Central Statistical Office of Poland, (in Polish).

BOLESŁAWSKI, L., (1974). Generation-based fertility tables of women, Warsaw: Central Statistical Office of Poland, (in Polish).

BUMPASS, L., (1990). What's happening to the family? Interactions between demographic and institutional changes. Demography, 27 (4), pp. 483–498.

CENTRAL STATISTICAL OFFICE OF POLAND, (2014). Population forecast for 2014-2050 period, Warsaw, (in Polish).

CHIANG, C. L., (1985). A Staging Process with Applications in Biology and Medicine, Mathematics in Biology and Medicine, 57, pp. 374–385.

CHIANG, C. L., (1984). The Life Table and Its Applications, Florida: Krieger.

CHIANG, C. L., (1980). An Introduction to Stochastic Processes and Their Applications, New York: Krieger.

CHIANG, C. L., (1968). Introduction to Stochastic Processes in Biostatistics, New York: Wiley.

CHIANG, C. L., Van Den Berg B. J., (1982). A fertility table for the analysis of human reproduction, Mathematical Biosciences, 62 (2), pp. 237–251.

CHENG, P. C. R., LIN, E. S., (2010). Completing incomplete cohort fertility schedules, Demographic Research, 23 (9), pp. 223–256.

CIGNO, A., (1994). Consideration in the Timing of Births, Theory and Evidence, In: J., Ermisch, N., Ogaza ed, The Family, The Market and the State in the Ageing Societies. Oxford: Clarendon Press.

EASTERLIN, R. A., (1979). The Economics and Sociology of Fertility: A Synthesis, In: Ch. Tilly, ed. Historical Studies of Changing Fertility, Princeton.

EASTERLIN, R., (1976). The conflict between aspirations and resources, Population and Development Review, 2 (3), pp. 417–425.

FRĄTCZAK, E., (1996). Cohort analyses of fertility based on Polish retrospective Survey 1988 „Life Course – family, occupational and migratory biographies", Warsaw: Polish Demographic Society, (in Polish).

FRĄTCZAK, E., (2001). Family tables of life Poland 1988/1989, 1994/1995, Monografie i Opracowania (485), Warsaw: Warsaw School of Economics, (in Polish).

FRĄTCZAK, E., PTAK-CHMIELEWSKA, A. ed., (2011a). Fertility in Poland – cohort analysis: birth cohorts 1911–1986, Vol. I. Warsaw: WSE.

FRĄTCZAK, E., PTAK-CHMIELEWSKA, A. ed., (2011b). Fertility and Nuptiality in Poland – cohort analysis: birth cohorts 1911–1986, Vol. II. Warsaw: WSE.

FRĄTCZAK, E., GRZENDA, W. (2011). Fertility life tables based on stochastic process, In: The selected problems of socio-demographic development of Poland and research methods, Prace i Materiały WZ UG, 2/3, pp. 7–20, (in Polish).

FRĄTCZAK, E., KOZŁOWSKI, W., (2005). Family status life tables. Poland 1988/1989, 1994/1995, 2002, Warsaw: WSE, (in Polish).

FREJKA, T., GIETEL-BASTEN, S., ABOLINA, L., ABULADZE, L., AKSYONOVA, S., AKRAP, A., FOLDES, I., (2016). Fertility and family policies in Central and Eastern Europe after 1990, Comparative Population Studies, 41 (1), pp. 3–56.

FREJKA, T., SARDON, J-P., (2004). Childbearing trends and prospects in low-fertility countries. A cohort analysis, The Netherlands: Kluwer Academic Publishers.

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

GRAUNT, J., (1962). Natural and political observations mentioned in a following index, and made upon the bills of mortality. [online] Available at: <http://www.neonatology.org/pdf/graunt.pdf.> [Accessed 26 May 2018].

KOTOWSKA, I., JÓŹWIAK, J., MATYSIAK, A., BARANOWSKA, A., (2008). Poland: Fertility decline as a response to profound societal and labour market changes, Demographic Research, 19 (22), pp. 795–854.

LEE, R. D., (1974). Forecasting births in post-transition populations: Stochastic renewal with serially correlated fertility, Journal of the American Statistical Association, 69(347), pp. 607–617.

LESTHAEGHE, R., (1998). On Theory Development: Applications to the Study of Family Formation, Population and Development Review, 24 (1), pp. 1–14.

LESTHAEGHE, R., (1991). The Second Demographic Transition in Western Countries: An Interpretation IPD Working Paper, pp. 1991–2.

LESTHAEGHE, R., SURKYN, J., (1988). Cultural dynamics and economic theories of fertility change, Population and Development Review, 14 (1), pp. 1–45.

LIU, X., (2012). Survival analysis: models and applications, John Wiley & Sons.

MCDONALD, P., (2008). Very Low Fertility Consequences, Causes and Policy Approaches, The Japanese Journal of Population, 6 (1), pp. 19–23.

MCDONALD, P., (2006). Low Fertility and the State: The Efficacy of Policy, Population and Development Review, 32 (3), pp. 485–510.

NAMBOODIRI, K., (1991). Demographic Analysis, A Stochastic Approach, California: Academic Press.

NAMBOODIRI, K., SUCHINDRAN, C. M., (1987). Life Table Techniques and Their Applications, New York: Academic Press.

OLAH L. S., FRĄTCZAK E. (eds.), (2013). Childbearing Women's Employment and Work-Life Balance Policies in Temporary Europe, Hampshire: Palgrave and Macmillan.

PARADYSZ, J., (1992). Women's fertility in Poland, Materials and statistical studies, NC'88, Warsaw: Central Statistical Office of Poland, (in Polish).

SOBOTKA, T., ZEMAN, K., LESTHAEGHE, R., FREJKA, T., NEELS, K., (2012). Postponement and recuperation in cohort fertility: Austria, Germany and Switzerland in a European context, Comparative Population Studies, 36 (2–3).

SOBOTKA, T., (2003). Tempo-quantum and period-cohort interplay in fertility changes in Europe, Evidence from the Czech Republic, Italy, the Netherlands and Sweden. Demographic Research, 8 (6), pp. 151–214.

VAN DE KAA, D. J., (1996). Anchored Narratives: The story and Findings of Half a Century of Research into the Determinants of Fertility, Population Studies, 50, pp. 389–432.

VAN DE KAA, D. J., (1994). The Second Demographic Transition Revisited: Theories and Expectations, In: G. Beets et al. (eds.), Population and family in the Low Countries 1993: Late fertility and other current issues, Swets and Zeitlinger, Berwyn, Pennsylvania/Amsterdam: NIDI/CBGS Publication, 30, pp. 81–126.

VAN DE KAA, D. J., (1987). Europe's second demographic transition, Population Bulletin, 42 (1).

WILLEKENS, F., (1991) Life table analysis of staging processes, In: H. Becker (ed.), Life histories and generations, Utrecht: ISOR Press, pp. 477–518.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 331-350, DOI 10.21307/stattrans-2018-019

GENERALIZED EXPONENTIAL SMOOTHING IN PREDICTION OF HIERARCHICAL TIME SERIES

Daniel Kosiorowski[1], **Dominik Mielczarek**[2], **Jerzy P. Rydlewski**[3], **Małgorzata Snarska**[4]

ABSTRACT

Shang and Hyndman (2017) proposed a grouped functional time series forecasting approach as a combination of individual forecasts obtained using the generalized least squares method. We modify their methodology using a generalized exponential smoothing technique for the most disaggregated functional time series in order to obtain a more robust predictor. We discuss some properties of our proposals based on the results obtained via simulation studies and analysis of real data related to the prediction of demand for electricity in Australia in 2016.

Key words: functional time series, hierarchical time series, forecast reconciliation, depth for functional data.

1. Introduction

The problem of optimal reconciliation of forecasts of complex economic phenomena partitioned into certain groups and/or levels of hierarchy has been considered in the economic and econometric literature many times and is still present in a public economic debate (see Kohn (1982), Weale (1988), Hyndman et al. (2011)). National import/export quantities or Gross National Product balances are important examples here. Discrepancies between forecasts prepared at the global level and obtained by aggregating regional forecasts or forecasts prepared according to certain hierarchy levels are usually thought to be caused by different methodologies or different precision of measurements used at different hierarchy levels or "prediction clusters". The issue is also very important from a particular company's point of view in a context of a product or a service lines management, consumers portfolio optimization and consumers segmentation. Let us take the equipment for running grouped in levels of hierarchy with respect to age, sex, competitive or re-creative usage and season designation as an example of material product line management. Let us take demand and supply of electricity within day and night optimized with respect to forecasted day and night demand, customers grouped with respect to

[1]Department of Statistics, Cracow University of Economics, Kraków, Poland.
E-mail: daniel.kosiorowski@uek.krakow.pl.
[2]AGH University of Science and Technology, Faculty of Applied Mathematics, al. A. Mickiewicza 30, 30-059 Krakow, Poland. E-mail: dmielcza@wms.mat.agh.edu.pl.
[3]AGH University of Science and Technology, Faculty of Applied Mathematics, al. A. Mickiewicza 30, 30-059 Krakow, Poland. E-mail: ry@agh.edu.pl.
[4]Department of Financial Markets, Cracow University of Economics, Kraków, Poland.
E-mail: malgorzata.snarska@uek.krakow.pl.

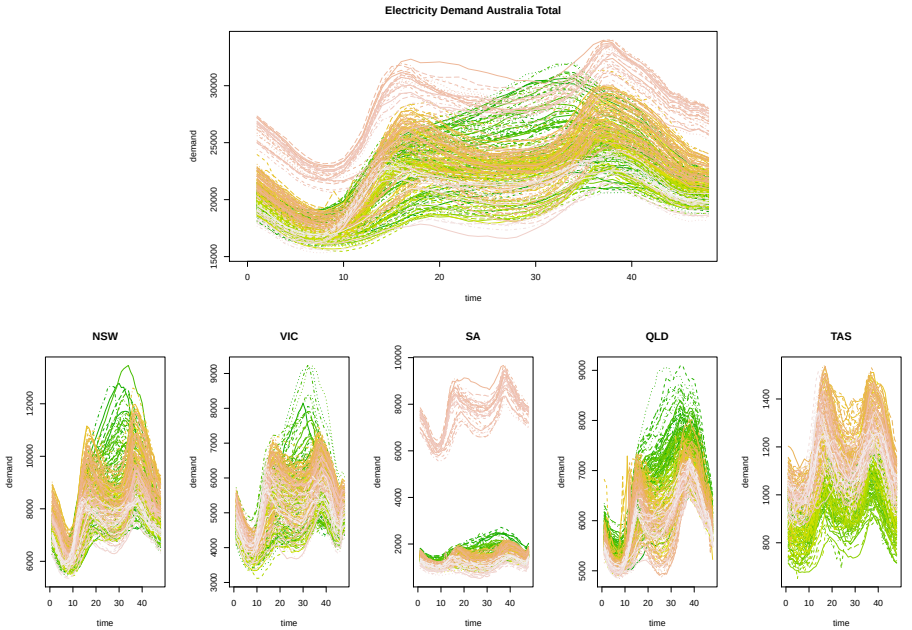


Figure 1: Electricity demand in regions of Australia in 2016 – hierarchical functional time series example

regions of living or residing and a degree of ”consumer priority” as examples of immaterial product sales optimization. Let us take the Internet holiday booking service divided into sub-services with respect to certain wealth or ”an inclination to adventures” criterion as an example of a service management. In recent years a very interesting statistical methodology named functional data analysis (FDA) for analyzing functional data has been developed (see Bosq (2000), Ramsay et al. (2009), Horvath and Kokoszka (2012), Krzyśko et al. (2013), Shang and Hyndman (2017)). For applications of FDA in economics see Kosiorowski (2014), Kosiorowski (2016), Kosiorowski, Rydlewski and Snarska (2017a). Economic usefulness of outliers detection procedures in the FDA setup has been recently described and discussed in Nagy et al. (2017), Kosiorowski, Rydlewski and Zawadzki (2018a), Kosiorowski, Mielczarek and Rydlewski (2018c).
In this context, it is worth stressing, that many others economic phenomena may effectively be described by means of functions or their series (i.a. utility curves, yield curves, development paths of companies or countries).

Following the above cited authors we consider a random curve $X = \{x(t), t \in [0,T]\}$, where T is fixed, as a random element of the separable Hilbert space $L^2([0,T])$ with the inner product $< x,y >= \int x(t)y(t)dt$. The space is equipped with the Borel $\sigma-$algebra. Furthermore, in Bosq's (2000) monograph it is proved that

probability distributions do exist for functional processes with values in Hilbert space. We denote this probability distribution by $\mathscr{F}$. Functional time series (FTS) is a series of functions indexed by time (e.g. see Fig. 1, colors indicate time succession of sequence of the functional objects according to the base R *terrain.colors* color palette). A hierarchical functional time series is a series of functions grouped at specified levels (household, town, region, whole country), (i.e. see Fig. 1). At each level a forecast can be made. A natural problem arises: how to use information obtained at different levels to obtain a reconciliated prediction for all levels?

The problem of hierarchical time series prediction is solved with various ways. Bottom-up method relies on forecasting each of the disaggregated series at the lowest level of the hierarchy, and then using simple aggregation to obtain forecasts at a higher level of the hierarchy (see Kahn (1998)). Top-down method involves forecasting of aggregated series at the top level of the hierarchy, and then using disaggregation to obtain forecasts at a lower level of the hierarchy based on historical proportions. Shang and Hyndman (2017), extending the method of Hyndman et al. (2011), considered grouped functional time series forecasting as an optimal combination of individual forecasts using generalized least squares regression with level forecasts treated as regressors. In the context of hierarchical FTS prediction a general problem arises: which method of forecasting should be chosen (see Bosq (2000), Besse et al. (2000), Hyndman and Ullah (2007), Hyndman and Shang (2009), Aue et al. (2015), Kosiorowski, Mielczarek and Rydlewski (2017b, 2018b)). Shang and Hyndman (2017) proposed a grouped functional time series forecasting approach as a combination of individual forecasts obtained by means of their smart predicting method, in which functional time series is reduced to a family of one dimensional time series of principal component scores representing original functional series (see Kosiorowski, 2014). As a result of conducted simulation studies, we decided to modify their methodology. Instead of using principal component scores forecast methods, we decided to propose a certain functional generalization of exponential smoothing technique (see Hyndman et al. (2008) for a theoretical background of the exponential smoothing), i.e. we used moving local medians and moving local functional trimmed averages (Febrero-Bande and de la Fuente, 2012) for the most disaggregated series in order to obtain more robust predictor than Shang and Hyndman (2017). The main aim of the paper is to modify Shang and Hyndman (2017) predictor so that it could cope with functional outliers and/or it would be elastic enough to adapt to changes in data generating mechanism. The remainder of the paper is as follows: in the second section elements of depth concept for functional data are sketched and in the third section our proposals are introduced. Fourth section presents results of simulation as well as empirical studies. The paper ends with conclusions, references and a short appendix containing R script showing how to calculate forecasts using our proposals with free *DepthProc* and *fda.usc* R package (see Kosiorowski and Zawadzki, 2018; Febrero-Bande and de la Fuente, 2012).

2. Depths for functional data

For obtaining robust hierarchical FTS predictor we focused our attention on the functional data depth concept (Nagy et al. (2016) and Nieto-Reyes and Battey (2016)). We have chosen, in our opinion the best depth for the considered functional data, namely the corrected generalized band depth (cGBD, see López-Pintado and Jörnsten (2007)), but for computational reasons we restrict our considerations to the case of the band consisting of two functions.
If X_1 and X_2 are independent functional random variables generated by the functional time series, and generating the observations (real functions in the $L^2([0,T])$ space), the cGBD of curve x with respect to $\mathscr{F}$ is defined as

$$cGBD(x|\mathscr{F}) = \mathscr{F}\left(G(x) \subset A^c(x;X_1,X_2)\right)$$

where $G(x) = \{(t,x(t)) : t \in [0,T])\}$ is a graph of function x, and $a_{1,2} = \{t \in [0,T] : X_2(t) - X_1(t) \geq 0\}$ and

$$A^c(x;X_1,X_2) = \{t \in a_{1,2} : X_1(t) \leq x(t) \leq X_2(t)\}, \text{ if } a_{1,2} \geq a_{2,1}$$

or

$$A^c(x;X_1,X_2) = \{t \in a_{2,1} : X_2(t) \leq x(t) \leq X_1(t)\}, \text{ if } a_{2,1} > a_{1,2}.$$

Now, let $\mathbf{X}^N = \{x_1,...,x_N\}$ be a sample of continuous curves defined on the compact interval $[0,T]$. Let λ denote the Lebesgue measure and let $a(i_1,i_2) = \{t \in [0,T] : x_{i_2}(t) - x_{i_1}(t) \geq 0\}$, where x_{i_1} and x_{i_2} are band delimiting objects. Let $L_{i_1,i_2} = \frac{\lambda(a(i_1,i_2))}{\lambda([0,T])}$. The empirical cGBD of a curve x with respect to the sample $\mathbf{X}^N$, which estimates cGBD for curve x in the considered functional space with respect to $\mathscr{F}$, is defined as (see López-Pintado and Jörnsten, 2007)

$$cGBD(x|\mathbf{X}^N) = \frac{2}{N(N-1)} \sum_{1 \leq i_1 < i_2 \leq N} \frac{\lambda\left(A^c(x;x_{i_1},x_{i_2})\right)}{\lambda([0,T])}$$

where

$$A^c(x;x_{i_1},x_{i_2}) = \{t \in a(i_1,i_2) : x_{i_1}(t) \leq x(t) \leq x_{i_2}(t)\}, \text{ if } L_{i_1,i_2} \geq \frac{1}{2}$$

or

$$A^c(x;x_{i_1},x_{i_2}) = \{t \in a(i_2,i_1) : x_{i_2}(t) \leq x(t) \leq x_{i_1}(t)\}, \text{ if } L_{i_2,i_1} > \frac{1}{2}.$$

Within this definition, the introduced earlier band depth (López-Pintado and Romo, 2009) is modified so that it only takes into account the proportion of the domain where the delimiting curves define a contiguous region which has non–zero width. Further in our proposals we use *the depth regions of order* α for the considered cGBD, i.e. $R_\alpha(\mathscr{F}) = \{x : cGBD(x,\mathscr{F}) \geq \alpha\}$. Note, that α-central regions $R_\alpha(\mathscr{F}) = \{x \in L^2([0,T]) : D(x,\mathscr{F}) \geq \alpha\}$ may be defined for any statistical depth function $D(x,\mathscr{F})$, where $\mathscr{F}$ denotes a probability distribution (Zuo and Serfling, 2000). Note also that various robust and nonparametric descriptive characteristics, like scatter, skew-

ness, kurtosis, may be expressed in terms of $\alpha-$ central regions. These regions are nested and inner regions contain less and less probability mass. Following Paindaveine and Van Bever (2013), when defining local depth it will be more appropriate to index the family $\{R_\alpha(\mathscr{F})\}$ by means of probability contents. Consequently, for any $\beta \in (0,1]$ we define the smallest depth region with $\mathscr{F}$-probability equal or larger than β as

$$R^\beta(\mathscr{F}) = \bigcap_{\alpha \in A(\beta)} R_\alpha(\mathscr{F}),$$

where $A(\beta) = \{\alpha \geq 0 : P(R_\alpha(\mathscr{F})) \geq \beta\}$. The depth regions $R_\alpha(\mathscr{F})$ and $R^\beta(\mathscr{F})$ provide only the deepest point neighborhood. In our considerations, we can replace probability distribution $\mathscr{F} = \mathscr{F}^{\mathbf{X}}$ by its symmetrized version for any function x, namely $\mathscr{F}_x = \frac{1}{2}\mathscr{F}^{\mathbf{X}} + \frac{1}{2}\mathscr{F}^{2x-\mathbf{X}}$. For any depth function $D(\cdot,\mathscr{F})$ the corresponding *sample local depth function at the locality level* $\beta \in (0,1]$ is $LD^\beta(x,\mathscr{F}^{(N)}) = D(x,\mathscr{F}_x{}^{\beta(N)})$, where $\mathscr{F}_x^{\beta(N)}$ denotes the empirical probability distribution related to those functional observations that belong to $R_x^\beta(\mathscr{F}^{(N)})$, where $\mathscr{F}^{(N)}$ denotes empirical probability distribution calculated from X^N. Thus $R_x^\beta(\mathscr{F}^{(N)})$ is the smallest sample depth region that contains at least a proportion β of the $2N$ random functions $x_1,...,x_N, 2x-x_1,...,2x-x_N$. Depth is always well defined – it is an affine invariant from original depth. For $\beta = 1$ we obtain global depth, while for $\beta \simeq 0$ we obtain extreme localization. As in the population case, our sample local depth will require considering, for any $x \in L^2([0,T])$, the symmetrized distribution $\mathscr{F}_x^{(N)}$, which is empirical distribution associated with $x_1,...,x_N, 2x-x_1,...,2x-x_N$. Sample properties of the (global) depths result from general findings presented in Zuo and Serfling (2000). Implementations of local versions of several depths including projection depth, Student, simplicial, L^p depth, regression depth and modified band depth can be found in free R package *DepthProc* (see Kosiorowski and Zawadzki, 2018). In order to choose the locality parameter β we recommend using expert knowledge related to the number of components or regimes in the considered data. Sample properties of the local versions of depths result from general findings presented in Paindaveine and Van Bever (2013). For other concepts of local depths see, e.g., Sguera et al. (2016).

3. Our proposals

We consider a sample of N functions $\mathbf{X}^N$=$\{x_i(t) : t \in [0,T], i = 1,...,N\}$. Let $FD^\beta(y|\mathbf{X}^N)$ denote sample functional depth of $y(t)$ with locality parameter β, e.g., the functional depth is equal to corrected generalized band depth: $FD = cGBD$. Sample β–local median is defined as

$$MED_{FD^\beta}(\mathbf{X}^N) = \underset{i=1,...,N}{\arg\max}\, FD^\beta(x_i|\mathbf{X}^N).$$

Assume now, that we observe a functional time series $x_i^{cluster}(t)$, $i = 1,2,...$ for each considered level of hierarchy. We put our proposals forward.

PROPOSAL 1: We independently apply on each considered level of hierarchy and in each cluster of that level a moving median predictor to obtain a forecast for the cluster, and then generalized exponential smoothing takes the following form

$$\hat{x}_{n+1}^{cluster}(t) = MED_{FD^{\beta}}(W_{n,k}^{cluster}),$$

where $W_{n,k}$ denotes a moving window of length k ending in a moment n, i.e. $W_{n,k}^{cluster} = \{x_{n-k+1}^{cluster}(t),...,x_n^{cluster}(t)\}$.

Then joint reconciliation of forecasts is conducted and our reconciled predictor takes the form:

$$\hat{\mathbf{X}}_{n+1}(t) = \mathbf{F}(\hat{x}_{n+1}^{cluster_1}(t),...,\hat{x}_{n+1}^{cluster_M}(t)),$$

where $\mathbf{F}$ denotes the reconciliation procedure. In Proposal 1, $\mathbf{F}$ is a generalized least squares procedure originally proposed by Shang and Hyndman (2017) and M is the number of moving median predictors.

PROPOSAL 2: Sample $\beta_{threshold}$–trimmed mean with locality parameter β is defined as

$$ave(\beta_{threshold},\beta)(\mathbf{X}^N) = ave(x_i : FD^{\beta}(x_i|\mathbf{X}^N) > \beta_{threshold}),$$

where ave denotes the sample functional average, and $\beta_{threshold}$ is a a pre-specified trimming threshold. In this setup a generalized exponential smoothing technique is applied independently on each considered level of hierarchy and in each cluster of that level as well. As a predictor for $(n+1)th$ moment we take

$$\hat{x}_{n+1}^{cluster}(t) = z \cdot ave(\beta_{threshold}^1,\beta_1)(W_{n_1,k_1}^{cluster}) + (1-z) \cdot ave(\beta_{threshold}^1,\beta_1)(W_{n_2,k_2}^{cluster}),$$

where $W_{n,k}^{cluster}$ denotes a moving window of length k ending in a moment n , i.e. $W_{n,k}^{cluster} = \{x_{n-k+1}^{cluster}(t),...,x_n^{cluster}(t)\}$, $z \in [0,1]$ is a forgetting parameter and $n_2 < n_1$. Thus we consider a closer past represented by $W_{n_1}^{cluster}$ and a further past represented by $W_{n_2}^{cluster}$.

Note that lengths of the moving windows k, k_1, k_2 used in Proposals 1 – 2 relate to the analogous forgetting parameters α in the classical exponential smoothing. Additionally, we have at our disposal the resolution parameter β, at which we predict a phenomenon. Such an approach allows us to accommodate expert knowledge and adjust forgetting and resolution parameters to the researcher's requirements. For comparison purpose, note that in original Shang and Hyndman (2017) paper, the authors made predictions using functional regression based on constant in time functional principal components scores modelled by means of the well known one–dimensional time series models (see Hyndman and Shang, 2009).

In the next step we consider the whole hierarchy structure of the phenomenon. Assume that a hierarchical structure is described by fixed hierarchy levels, which are also divided into sub-levels, which are divided into sub-levels, and so on – assume that we have M clusters in the whole hierarchical structure. Smart reconciliation of forecasts is conducted in this step and our reconciled predictor takes the form:

$$\hat{\mathbf{X}}_{n+1}(t) = \mathbf{F}(\hat{x}_{n+1}^{cluster_1}(t), ..., \hat{x}_{n+1}^{cluster_M}(t)),$$

where $\mathbf{F}$ is a Generalized Least Squares Estimator (see Shang and Hyndman, 2017).
Using Shang and Hyndman (2017) notation we can write our model in the form

$$R_n = S_n b_n,$$

where R_n is a vector of all series at all clusters, b_n is a vector of the most disaggregated data and S_n is a fixed matrix that shows a relation between them. In the considered example we have $R_n = [R_{Australia},\ R_{NSW},\ R_{QLD},\ R_{SA},\ R_{TAS},\ R_{VIC}]^T$, $b_n = [R_{NSW},\ R_{QLD},\ R_{SA},\ R_{TAS},\ R_{VIC}]^T$, where T denotes a vector transpose. The matrix S_n is an 6×5 matrix, where the only non-zero elements are $S_{1i} = 1$ for $i = 1, 2, ..., 5$ and $S_{jj-1} = 1$ for $j = 2, ..., 6$. We propose to do the base forecast:

$$\hat{R}_{n+1} = S_{n+1}\beta_{n+1} + \varepsilon_{n+1},$$

where $\hat{R}_{n+1}$ is a matrix of the base forecast for all series at all levels,
$\beta_{n+1} = E[b_{n+1}|R_1, ..., R_n]$ is the unknown mean of the forecast distribution of the most disaggregated series and ε_{n+1} is responsible for errors. We propose to use a generalized least square method as in Shang and Hyndman (2017)

$$\hat{\beta}_{n+1} = \left(S_{n+1}^T W^{-1} S_{n+1}\right)^{-1} S_{n+1}^T W^{-1} \hat{R}_{n+1}$$

modified so that we use a robust estimator of the dispersion matrix W, i.e. instead of diagonal matrix, which contains forecast variances of each time series, we can use a robust measure of joint forecast dispersion taking into account dependency structure between the level forecasts. Note that a dynamic updating of the dispersion matrix estimates should be considered in further studies. Let us define our dispersion matrix:

$$W = diag\{v^{total}, v^{cluster_1}, ..., v^{cluster_M}\}$$

where

$$v^{cluster} = V\left\{\int_0^T \left(x_{nk}^{cluster}(t) - \hat{x}_n^{cluster}(t)\right)^2 dt, k = 1, 2, ..., K_{cluster}, n = 1, 2, ..., N\right\},$$

$K_{cluster}$ is the number of observations in the considered cluster in time n, N is here the number of moments at which observations have been made and where V is some chosen robust measure of dispersion. *We propose to substitute* $V = c \cdot MAD^2$

instead of Shang and Hyndman's (2017) variance. If we consider a hierarchy as in the above Figure 1, our dispersion matrix takes a simple form:

$$W = diag\{v^{Australia}, v^{NSW}, v^{QLD}, v^{SA}, v^{TAS}, v^{VIC}\}.$$

We propose to use $c \cdot MAD^2$ instead of variance or take into account a dependency structure between level series using the well-known minimum covariance determinant (MCD) or recently proposed PCS robust matrix estimators of multivariate scatter (see Vakili and Schmitt (2014)).

4. Properties of our proposals

Thanks to the kindness of prof. Han Lin Shang, who made his R script available for us, we calculated the optimal combination of forecast predictor and we compared Shang and Hyndman (2017) proposal with our Proposals 1 and 2 and with independent moving functional mean (no reconciliation was conducted for this predictor).

We generated samples of trajectories from one dimensional SV, GARCH, Wiener, Brownian bridge processes, functional FAR(1) processes tuned as in Didericksen et al. (2012), and various mixtures of them. In the cases of the considered mixtures, we treated one of the component as "good" and the rest as "bad" - outlying. In the simulations we considered several locality parameters which differed within the levels of hierarchy and several moving window lengths. We considered samples with and without functional magnitude as well as shape outliers (see Kosiorowski, Mielczarek, Rydlewski, 2018b, 2018c, and references therein). The outliers were defined with respect to the functional boxplot induced by cGBD, i. e. we replaced 1%, 5%, 10% of curves in the samples by means of arbitrary curves being outside a band determined by the functional boxplot whiskers, and compared medians and medians of absolute deviations from the medians (MAD) of integrated forecasts errors in these two situations. Fig. 2 presents simulated hierarchical functional time series consisting of three functional autoregression models of order 1 (i.e. FAR(1)) with Gaussian kernels and sine–cosine errors design (see Didericksen et al. 2012). Fig. 3 presents corresponding level forecasts obtained by means of our Proposal 1 and local moving median calculated from 15-obs. windows and locality parameters equal to 0.45. Fig. 4 presents simulated hierarchical time series consisting of three processes, each being mixtures of two one-dimensional stochastic volatility processes (SV). Fig. 5 presents corresponding level forecasts obtained by means of our Proposal 1 and local moving median calculated from 15-obs. window and locality parameters equal to 0.45. In Figures 2, 4, 5 and 6 colours indicate time sequence of the functional objects according to basic R package *terrain.colors* colour palette. We indicated the order of appearance of observations using colors palette starting from yellow and ending in blue. In the appendix we placed a simple script depending on *DepthProc* R package illustrating a general idea of the performed simulations.

In order to check the statistical properties of our proposals we considered em-

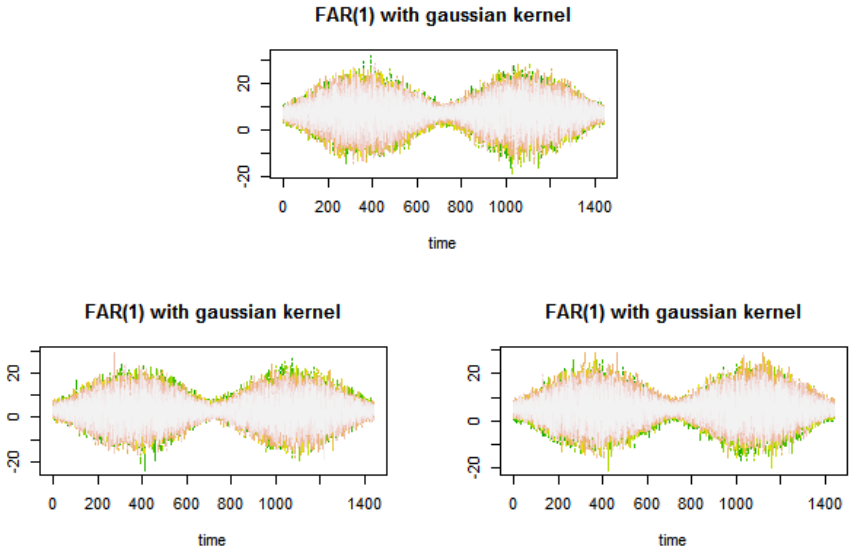


Figure 2: Simulated HFTS consisting of FAR(1) processes

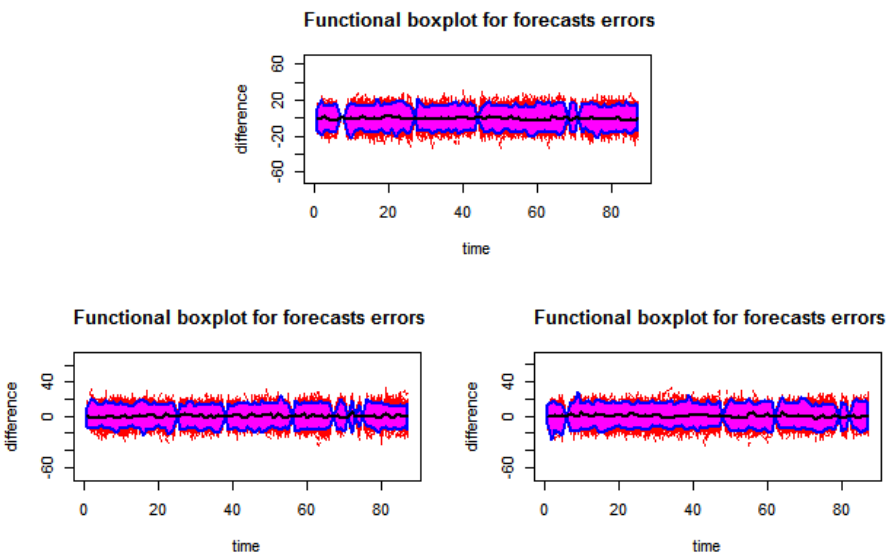


Figure 3: HFTS prediction using our Proposal 1

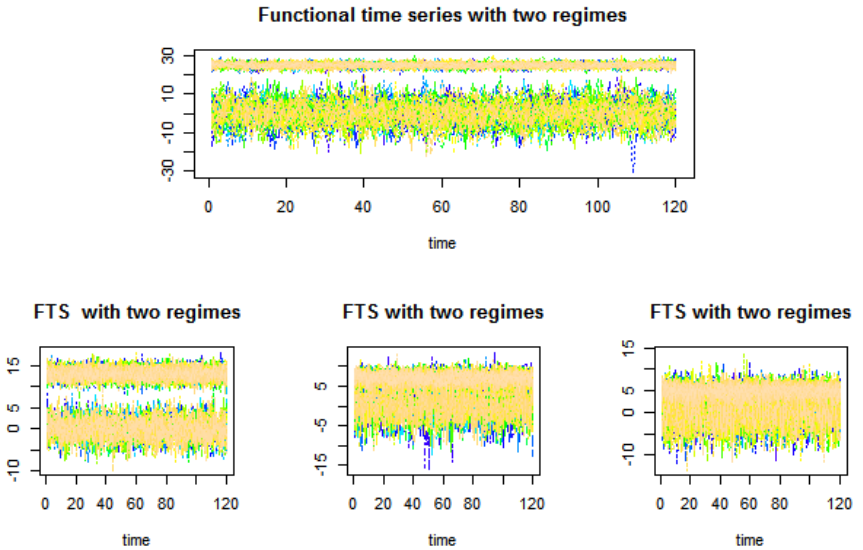


Figure 4: Simulated HFTS consisting of two regime FTS processes

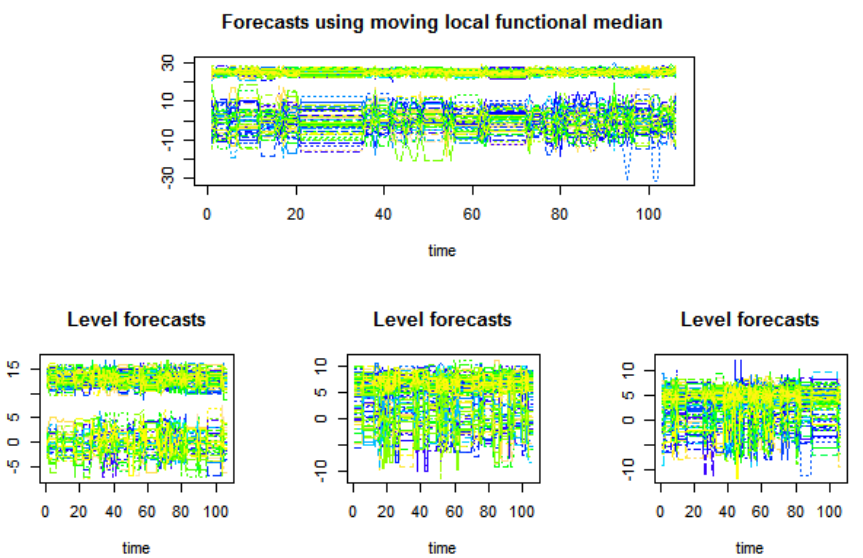


Figure 5: HFTS prediction using our Proposal 1

pirical data set related to an electricity demand in the period from 1 to 31 January 2016 in Australia. The data come from five regions of Australia, denoted with the following symbols: NSW, QLD, SA, TAS, VIC. All the considered data were taken from Australian Energy Market Operator https://www.aemo.com.au/. Fig. 6 presents 291 predicted electricity demand curves obtained by means of our Proposal 1 using moving local median to calculate level forecasts, window lengths $k = 10$ observations and locality parameter equal to $\beta = 0.2$. Fig. 7 presents boxplots for integrated prediction errors in each region and in the whole Australia. Fig. 8 presents functional boxplot for the prediction of errors for the whole Australia in 2016, where the predictions were obtained using Proposal 1. Fig. 9 presents functional boxplots for prediction of errors of electricity demand in NSW, QLD, VIC and TAS in 2016 obtained using Proposal 1. Note that the functional median of prediction errors is close to 0 for Australia and its regions (see Fig. 8 and 9). The boxplots in Fig. 7 show that a prediction error is small as well. The volumes of central regions in Fig. 8 and 9 may be treated as the predictor effectiveness measure. We could therefore deduce that our predictor is median-unbiased (for more details on median-unbiasedness for functional data see Kosiorowski, Mielczarek and Rydlewski, 2017b, and references therein).

The performance of our proposals was compared with Shang and Hyndman (2017) proposal and with the independent moving functional mean predictor (without the reconciliation of forecasts), in terms of the median of absolute deviation of integrated prediction errors in each region and in the whole Australia. Table 1 summarizes results of this comparison. In general, the obtained results lead us to the conclusion that our proposals seem to be more robust to functional outliers than Hyndman and Shang proposal. It is not surprising, as the authors made their forecasts on the basis of nonrobust generalized least squares method. Admittedly, Shang and Hyndman (2017) claimed that their proposal performed better in comparison with bottom-up approach based on moving medians, but note that they considered Fraiman and Muniz global depths only. Moreover, thanks to the locality parameter adjustment, our proposals are more appropriate for detecting the change of regimes in the HFTS setup. In the cases of data sets without outliers, simple functional moving means, where the reconciliation procedure is not conducted, seem to outperform all other proposals. In this "clean data" situation, the performance of our both proposals and of Shang and Hyndman (2017) proposal is comparable. Our second proposal is more computationally demanding, however.

Table 1. MAD of integrated forecasts errors

Predictor	*Australia*	*NSW*	*VIC*	*SA*	*QLD*	*TAS*
Proposal 1	1126	470	401	146	224	49
Proposal 2	1311	628	452	147	181	52
H & S Proposal	1275	627	1004	176	230	51

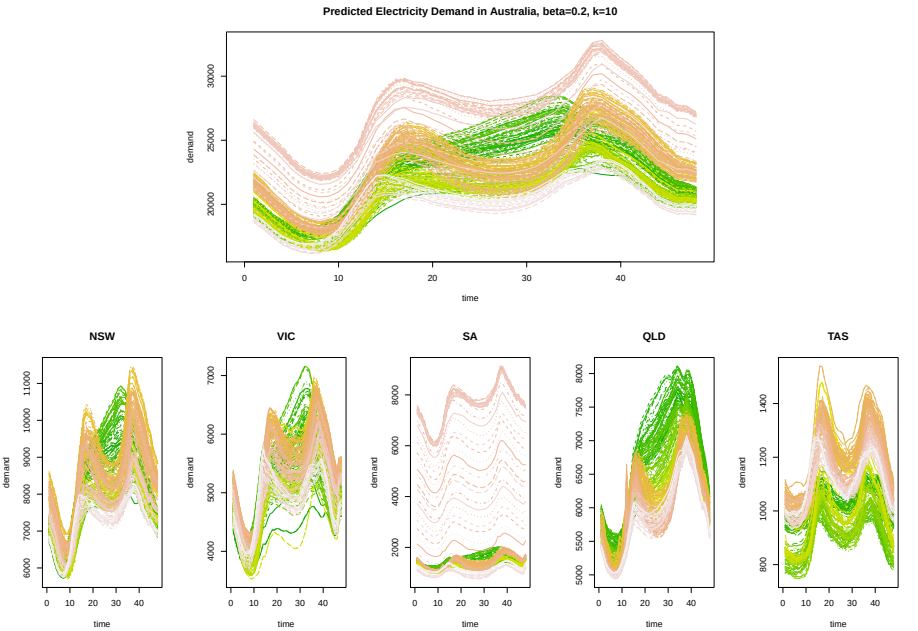


Figure 6: Predicted electricity demand in Australia in 2016

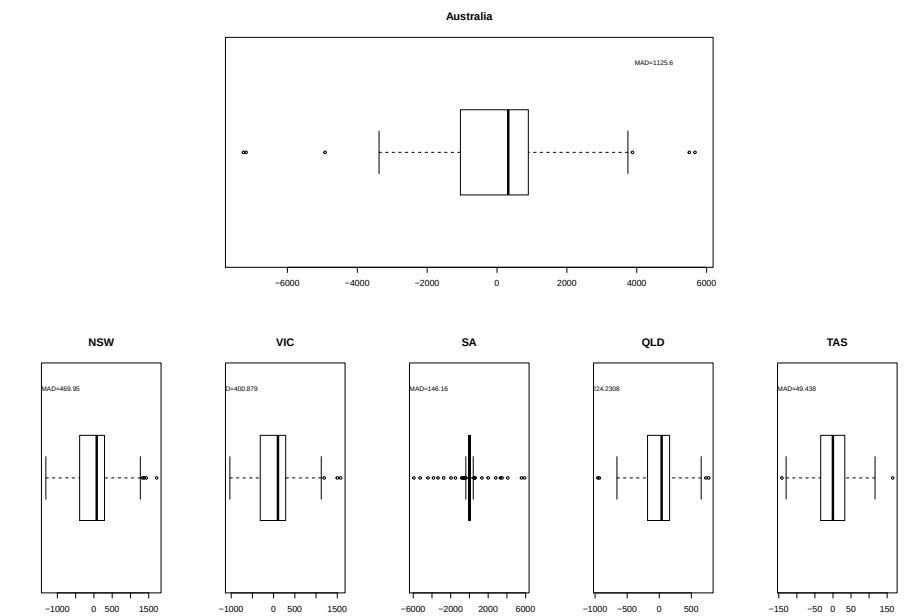


Figure 7: Boxplots for integrated forecasts errors for electricity demand in Australia and its regions in 2016, predictions obtained using Proposal 1

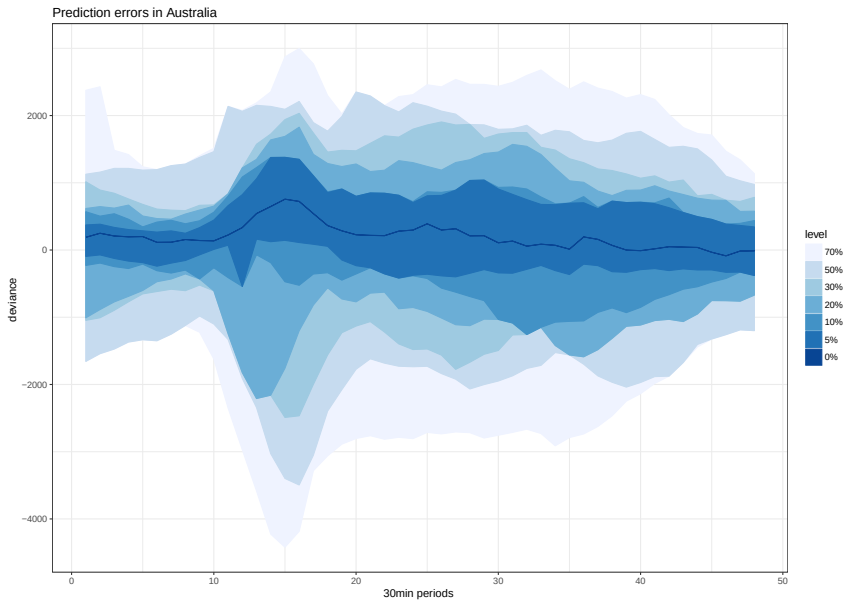


Figure 8: Functional boxplot for prediction of errors for the whole Australia in 2016, predictions obtained using Proposal 1

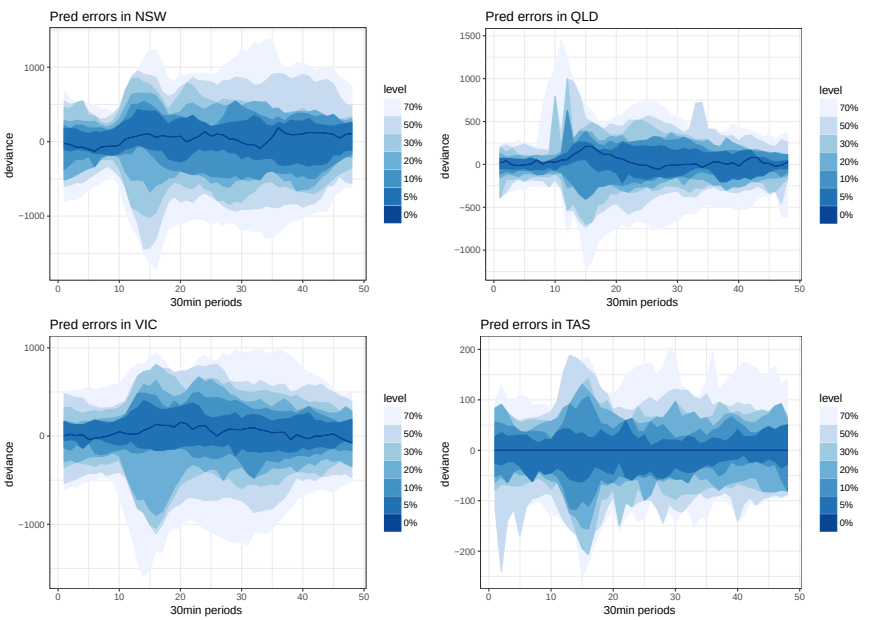


Figure 9: Functional boxplots for prediction errors of electricity demand in NSW, QLD, VIC and TAS in 2016 obtained using Proposal 1

4.1. Uncertainty evaluation

Series of functional principal component scores are considered as surrogates of original functional time series (see Aue et al. (2015), Hyndman and Shang (2009)). Several authors postulate using the dynamic functional principal components approach in order to take into account the time changing dependency structure of the described phenomenon (Aue et al., 2015). Note that such modification may drastically increase computational complexity of the HFTS procedure. In a context of uncertainty evaluation of our proposals, we suggest considering Vinod and López-de-Lacalle (2009) maximum entropy bootstrap for time series approach. Bootstrap methods for FTS were studied by Hörmann and Kokoszka (2012) and Shang (2018), among others. Similarly, as in Shang and Hyndman (2017) we propose to use maximum entropy bootstrap methodology to obtain confidence regions and to conduct statistical inference. The *meboot* and *DepthProc* R packages give the appropriate computational support for these aims.

5. Conclusions

The hierarchical functional time series methodology opens new areas of statistical as well as economic research. E–economy provides a great deal of HFTS data. Our HFTS predictor proposal, which is based on local moving functional median, performs surprisingly well in comparison with Shang and Hyndman (2017) proposal. The lengths of the moving windows used in our proposals relate to the forgetting parameters α's in the classical exponential smoothing. Moreover, we have at our disposal a "data resolution parameter" - β, at which we predict the phenomenon. When using the locality parameter, we can take into account different sensitivity to details, e.g. the number of different regimes of the considered phenomena. Further economic applications of the HFTS methodology may be found, for example, in Kosiorowski, Mielczarek, Rydlewski (2018b).

Acknowledgements

JPR research has been partially supported by the AGH UST local grant no. 15.11.420.038.
DM and JPR research has been partially supported by the Faculty of Applied Mathematics AGH UST statutory tasks within subsidy of the Ministry of Science and Higher Education, grant no. 11.11.420.004.
MS research was partially supported by NSC Grant no. OPUS.2015.17.B.HS4.02708,
DK research has been partially supported by the grant awarded to the Faculty of Management of CUE for preserving scientific resources for 2016, 2017 and 2018.

REFERENCES

AUE, A., DUBABART NORINHO, D., HÖRMANN, S., (2015). On the prediction of stationary functional time series, Journal of the American Statistical Association, 110 (509), pp. 378–392.

BESSE, P. C., CARDOT, H., STEPHENSON, D. B., (2000). Autoregressive forecasting of some functional climatic variations, Scandinavian Journal of Statistics, 27 (4), pp. 673–687.

BOSQ, D. (2000). Linear processes in function spaces. Springer.

DIDERICKSEN, D., KOKOSZKA, P., ZHANG, X. (2012). Empirical properties of forecasts with the functional autoregressive model, Computational Statistics, 27 (2), pp. 285–298.

FEBRERO-BANDE, M. O., DE LA FUENTE, M., (2012). Statistical computing in functional data analysis: the R package fda.usc, Journal of Statistical Software, 51 (4), pp. 1–28.

HORVATH, L., KOKOSZKA, P., (2012). Inference for functional data with applications, Springer-Verlag.

HÖRMANN S., KOKOSZKA, P., (2012). Functional Time Series, in Handbook of Statistics: Time Series Analysis – Methods and Applications, 30, pp. 157–186.

HYNDMAN, R. J., AHMED R. A., ATHANASOPOULOS, G., SHANG, H. L., (2011). Optimal combination forecasts for hierarchical time series, Computational Statistics & Data Analysis, 55 (9), pp. 2579–2589.

HYNDMAN, R.J., KOEHLER, A.B., ORD, J. K., SNYDER, R. D., (2008). Forecasting with exponential smoothing – the state space approach, Springer-Verlag.

HYNDMAN, R. J., SHANG, H., L., (2009). Forecasting functional time series, Journal of the Korean Statistical Society, 38 (3), pp. 199–221.

HYNDMAN, R. J., ULLAH, M., (2007). Robust forecasting of mortality and fertility rates: A functional data approach, Computational Statistics & Data Analysis, 51 (10), pp. 4942–4956.

HYNDMAN, R. J., KOEHLER, A. B., SNYDER, R.D., GROSE, S., (2002). A state space framework for automatic forecasting using exponential smoothing methods, International Journal of Forecasting, 18 (3), pp. 439–454.

KAHN, K. B., (1998). Revisiting top-down versus bottom-up forecasting, The Journal of Business Forecasting Methods & Systems, 17 (2), pp. 14–19.

KOHN, R., (1982). When is an aggregate of a time series efficiently forecast by its past, Journal of Econometrics, 18 (3), pp. 337–349.

KOSIOROWSKI, D., ZAWADZKI, Z., (2018). DepthProc: An R package for robust exploration of multidimensional economic phenomena, *arXiv: 1408.4542*.

KOSIOROWSKI, D., (2014). Functional regression in short term prediction of economic time series, Statistics in Transition, 15 (4), pp. 611–626.

KOSIOROWSKI, D. (2016). Dilemmas of robust analysis of economic data streams, Journal of Mathematical Sciences (Springer), 218 (2), pp. 167–181.

KOSIOROWSKI, D., RYDLEWSKI, J. P., SNARSKA, M., (2017a). Detecting a structural change in functional time series using local Wilcoxon statistic, Statistical Papers, pp. 1–22, URL http://dx.doi.org/10.1007/ s00362-017-0891-y.

KOSIOROWSKI, D., MIELCZAREK, D., RYDLEWSKI, J. P., (2017b). Double functional median in robust prediction of hierarchical functional time series – an application to forecasting of the Internet service users behaviour, available at: arXiv:1710.02669v1.

KOSIOROWSKI, D., RYDLEWSKI, J.P., ZAWADZKI Z., (2018a). Functional outliers detection by the example of air quality monitoring, Statistical Review (in Polish, forthcoming).

KOSIOROWSKI, D., MIELCZAREK, D., RYDLEWSKI, J. P., (2018b). Forecasting of a Hierarchical Functional Time Series on Example of Macromodel for the Day and Night Air Pollution in Silesia Region - A Critical Overview, Central European Journal of Economic Modelling and Econometrics, 10 (1), pp. 53–73.

KOSIOROWSKI, D., MIELCZAREK, D., RYDLEWSKI, J. P., (2018c). Outliers in Functional Time Series – Challenges for Theory and Applications of Robust Statistics, In M. Papież & S. Śmiech (eds.), The 12^{th} Professor Aleksander Zeliaś International Conference on Modelling and Forecasting of Socio-Economic Phenomena, Conference Proceedings, Cracow: Foundation of the Cracow University of Economics, pp. 209–218.

KRZYŚKO, M., DEREGOWSKI, K., GÓRECKI, T., WOŁYŃSKI, W., (2013). Kernel and functional principal component analysis, Multivariate Statistical Analysis 2013 Conference, plenary lecture.

LÓPEZ-PINTADO, S., ROMO, J., (2009). On the concept of depth for functional data, Journal of the American Statistical Association, 104, pp. 718–734.

LÓPEZ-PINTADO, S., JÖRNSTEN, R., (2007). Functional analysis via extensions of the band depth, IMS Lecture Notes–Monograph Series Complex Datasets and Inverse Problems: Tomography, Networks and Beyond, Vol. 54, pp. 103–120, Institute of Mathematical Statistics.

NAGY, S., GIJBELS, I., OMELKA, M., HLUBINKA, D., (2016). Integrated depth for functional data: Statistical properties and consistency, ESIAM Probability and Statistics, 20, pp. 95–130.

NAGY, S. GIJBELS, I., HLUBINKA, D., (2017). Depth-Based Recognition of Shape Outlying Functions, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2017.1336445.

NIETO-REYES, A., BATTEY, H., (2016). A topologically valid definition of depth for functional data, Statistical Science 31 (1), pp. 61–79.

PAINDAVEINE, D., G. VAN BEVER, G., (2013). From depth to local depth: a focus on centrality, Journal of the American Statistical Asssociation, Vol. 108, No. 503, Theory and Methods, pp. 1105–1119.

RAMSAY, J.O., G. HOOKER, G., GRAVES, S., (2009). Functional data analysis with R and Matlab, Springer-Verlag.

SGUERA, C., GALEANO, P., LILLO, R. E., (2016). Global and local functional depths, arXiv 1607.05042v1.

SHANG, H., L., HYNDMAN, R. J., (2017). Grouped functional time series forecasting: an application to age-specific mortality rates, Journal of Computational and Graphical Statistics, 26(2), pp. 330–343.

SHANG, H., L., (2018). Bootstrap methods for stationary functional time series, Statistics and Computing, 28(1), pp. 1–10.

WEALE, M., (1988). The reconciliation of values, volumes and prices in the national accounts, Journal of the Royal Statistical Society A, 151(1),pp. 211–221.

VAKILI, K., SCHMITT, E., (2014). Finding multivariate outliers with FastPCS, Computational Statistics & Data Analysis, 69, pp. 54–66.

VINOD, H.D., LÓPEZ-DE-LACALLE, J. L., (2009). Maximum entropy bootstrap for time series: the meboot R package, Journal of Statistical Software, 29 (5).

ZUO, Y., SERFLING, R., (2000). Structural properties and convergence results for contours of sample statistical depth functions, Annals of Statistics, 28 (2), pp. 483–499.

Australian Energy Market Operator, https://www.aemo.com.au/

APPENDIX

```
#Simple R script, example showing how to calculate base forecasts for three hierarchy levels
#using moving functional median implemented within the DepthProc R package.
require(DepthProc)
require(fda.usc)
require(RColorBrewer)
require(zoo)

wrapMBD = function(x){ depthMedian(x, depth_params = list(method="Local", beta=0.45,
depth_params1 = list(method = "MBD"))) }
#Simple stochastic volatility 1D process simulator
SV ¡- function(n, gamma, fi, sigma, delta) {
epsilon ¡- rnorm(n)
eta ¡- rnorm(2*n, 0, delta)
h ¡- c()
h[1] ¡- rnorm(1)
for (t in 2:(2*n)) {
h[t] ¡- exp(gamma+fi*(h[t-1]-gamma)+sigma*eta[t]) }
Z ¡- sqrt(tail(h,n)) * epsilon
return(Z)}
example ¡- SV(100, 0, 0.2, 0.5, 0.1)
plot(ts(example))
#functional time series simulator
m.data1¡-function(n,a,b) {
M¡-matrix(nrow=n,ncol=120)
for (i in 1:n) M[i,]¡- a*SV(120,0,0.3,0.5,0.1)+b
M }
m.data.out1¡-function(eps,m,n,a,b,c,d){
H¡-rbind(m.data1(m,a,b),m.data1(n,c,d))
ind=sample((m+n),eps)
H1=H[ind,]
H1 }
m ¡- matrix(c(1, 0, 1, 3, 2, 3, 2, 0), nrow = 2, ncol = 4) m[2,]=c(2,2,3,3) m[1,]=c(0,1,1,0)
#below three functional time series
M2A= m.data.out1(150,3000,7000,5,0,1,25)
M2B= m.data.out1(150,3000,7000,2,0,1,15)
M2C= m.data.out1(150,3000,7000,3,0,1,10)
matplot(t(M2A),type="l",col=topo.colors(151),xlab="time", main="Functional time series with
two regimes")
matplot(t(M2B),type="l",col=topo.colors(151),xlab="time", main="FTS with two regimes")
matplot(t(M2C),type="l",col=topo.colors(151), xlab="time", main="FTS with two regimes")
#below moving local medians applied to the above series, window lengths = 15 obs.,
#locality parameters betas = 0.45
result4A = rollapply(t(M2A),width = 15, wrapMBD, by.column = FALSE)
result4B = rollapply(t(M2B),width = 15,wrapMBD, by.column = FALSE)
result4C = rollapply(t(M2C),width = 15, wrapMBD, by.column = FALSE)
```

matplot(result4A,type="l",col=topo.colors(87), xlab="time",main="local 15-obs moving functional median, beta=0.45")
matplot(result4B,type="l",col=topo.colors(87), xlab="time",main="local 15-obs moving functional median, beta=0.45")
matplot(result4C,type="l",col=topo.colors(87), xlab="time",main="local 15-obs moving functional median,
beta=0.45")

#basic function for calculating $\beta\_treshold-$ trimmed $\beta-$ local MBD functional mean

beta_tresh_mean¡-function(x,beta_tresh,beta)
depths= depth(x, depth_params = list(method="Local", beta=beta, depth_params1 = list(method = "MBD")))
ind=which(depths$_beta_tresh$)
$wyn = func.mean(x[ind,])$
wyndata$

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 351–357, DOI 10.21307/stattrans-2018-020

ON A SURPRISING RESULT OF TWO-CANDIDATE ELECTION FORECAST BASED ON THE FIRST LEADERSHIP TIME

Czeslaw Stępniak[1]

ABSTRACT

This is a simple but provocative note. Consider an election with two candidates and suppose that candidate A was the leader until counting n votes. How to use this information in predicting the final results of the election? According to the common belief the final number of votes for the leader should be a strictly increasing function of n. Assuming the votes are counted in random order we derive the Maximum Likelihood predictor of the final number of votes for the future winner and loser based on the first leadership time. It appears that this time has little effect on the predicting.

Key words: two-candidate election, winner, leader, leadership time, predicting number of votes for winner, Maximum Likelihood.

1 Introduction

Two-candidate election such as the last round of presidential election always attracts a great attention. Suppose that candidate A was the leader until counting n votes. We write $T = n$ for the first leadership time T. The problem is how to use this information in predicting the final results of the election. According to the common belief the final number of votes for the leader should be a strictly increasing function of n.

Assume the votes are counted in random order. Combinatorial tools for the process of counting of votes in this situation may be found in many books and articles (Brémaud (1994), Feller (1968), Goulden and Serrano (2003), Lengyel (2011), Renault (2007) and Takacs (1997), among others) under the name of the ballot problem. The results are usually formulated in probabilistic terms.

Statistical inference is often based on the notion of likelihood (cf. Azzalini (1996)) and the Maximum Likelihood principle plays the fundamental role in the process. In the present note we derive the Maximum Likelihood predictor of the final number of votes for the future winner.

Presentation of this note is accessible not only for specialists.

[1] Faculty of Mathematics and Natural Sciences, University of Rzesów. Poland. E-mail: stepniak@umcs.lublin.pl.

2 Initial formalization and predicting the future winner

Consider election with two candidates A and B. In this note the number of all votes is known and is denoted by N. So that all potential results of the election were conclusive we assume that N is odd and each vote indicates exactly one candidate. The votes are counted in random order, that is all permutations are equally probable.

The results of the election are usually announced by the winner (W) and by the final number of votes for W. For some technical reasons, instead of this number, it will be convenient to handle the final number of votes for loser. Denote the last number by M. In this situation the pair (W,M), where $W \in \{A,B\}$ and $M \in \{0,1,...,\frac{N-1}{2}\}$ plays the role of unknown parameter.

One can consider two problems: predicting W under assumption that M is the nuisance parameter, and predicting M. The both predictors are based on the observation (L,T), where $L \in \{A,B\}$ is the first leader and T is the first leadership time.

As regards the first problem we may choose between two predictors: $W = L$ and $W \neq L$. Intuitively, the first one is better. We shall formally confirm this intuition. To this aim we only need to observe that a candidate will be the first leader if and only if the first vote is for him.

In consequence

$$P_M(W = L) = P_M(L = W) = \frac{\text{final number of votes for winner}}{\text{number of all votes}} = \frac{N-M}{N} > \frac{1}{2}$$

while

$$P_M(W \neq L) = P_M(L \neq W) = \frac{\text{final number of votes for loser}}{\text{number of all votes}} = \frac{M}{N} < \frac{1}{2}$$

for all $M \leq \frac{N-1}{2}$. Therefore, the predictor $W = L$ is better.

For predicting the final number of votes for winner and loser we shall use distribution $P_M(T = n)$ of the first leadership time T.

3 Towards distribution of the first leadership time

Some results on this distribution were derived in Stępniak (2015) under silent assumption that $M > 0$. We shall prove the following

Theorem 1 *For all $M = 0,1,...,\frac{N-1}{2}$ distribution $P_M(T = n)$ of the first leadership time T is given by*

$$P_M(T=n)=\begin{cases}\frac{2}{n}\binom{n}{\frac{n+1}{2}}\binom{N-n-1}{M-\frac{n+1}{2}}\binom{N}{M}^{-1}, & \text{if } n \text{ is a positive odd integer such that } \frac{n+1}{2}\le M,\\ \frac{N-2M}{N}, & \text{if } n=N,\\ 0, & \text{otherwise.}\end{cases} \tag{1}$$

Proof. Let us consider three cases:

I. $M=0$ (n arbitrary),

II. $n=N$ (M arbitrary),

III. $M>0$ and $n<N$.

We mention that the classical ballot problem refers only to the probability $P_M(T=N,L=W)$. In our notation the well-known Ballot Theorem (see Brémaud(1994), Feller (1968), Goulden and Serrano (2003), Lengyel (2011), Renault (2007) and Takacs (1997)) may be expressed in the form

$$P_M(T=N,L=W)=\frac{(N-M)-M}{N}=\frac{N-2M}{N} \quad \text{for all } M. \tag{2}$$

In the case II $P_M(T=N,L\neq W)=0$ and, therefore,

$$P_M(T=N)=\frac{N-2M}{N} \quad \text{for all } M=0,1,...,\frac{N-1}{2}.$$

The case I is trivial and it leads to

$$P_0(T=n)=\begin{cases}1, & \text{if } n=N,\\ 0, & \text{otherwise.}\end{cases} \tag{3}$$

In this situation the set of all positive integers n such that $\frac{n+1}{2}\le M$ is empty and hence the formula (3) coincides with (1).

Now let us consider the case III.

Any record of the counting of votes may be represented as a lattice path from the origin to $(N,N-2M)$ with steps of type $(1,1)$ and $(1,-1)$ indicating that a successive voice is for the future winner or loser. In particular, the first leader is the future winner with the leadership time n, if and only if the path is touching the x-axis for $x=n+1$ and lying above the axis for all positive integers $x\le n$. Similarly, the first leader is the future loser with the leadership time n, if the corresponding segment of the path is lying below the x-axis. In consequence, $P_M(T=n,L=W)$ is positive only for n odd and less than $2M$. Moreover

$$P_M(T=n,L\neq W)=P_M(T=n,L=W). \tag{4}$$

This fact is known as the Reflection Principle (see, Brémaud(1994) or Feller (1968), for instance). Thus it remains to find $P_M(T=n,L\neq W)$ for $n=1,3,...,2M-1$. The

desired probability may be expressed in the form

$$P_M(T=n, L\neq W) = P(A)P(B/A)P(C/A\cap B),$$

where

• A is the event that $\frac{n+1}{2}$ among the first n votes will be for loser (and $\frac{n-1}{2}$ for winner),

• B is the event that $(n+1)$-th vote will be for winner,

• C is the event that during counting the first n votes the loser will be always the leader.

By the well-known formula for the hypergeometric distribution we get

$$P(A) = \frac{\binom{M}{\frac{n+1}{2}}\binom{N-M}{\frac{n-1}{2}}}{\binom{N}{n}}.$$

Moreover

$$P(B/A) = \frac{N-M-\frac{n-1}{2}}{N-n}.$$

On the other hand, by the Ballot Theorem (2) for $N=n$ and $M=\frac{n-1}{2}$

$$P(C/A\cap B) = P(C/A) = \frac{n-(n-1)}{n} = \frac{1}{n}.$$

In consequence for $n = 1,3,...,2M-1$

$$P_M(T=n, L\neq W) = \frac{1}{n}\frac{N-M-\frac{n-1}{2}}{N-n}\frac{\binom{M}{\frac{n+1}{2}}\binom{N-M}{\frac{n-1}{2}}}{\binom{N}{n}}.$$

By some elementary operations on the binomial coefficients the last one reduces to

$$P_M(T=n, L\neq W) = \frac{1}{n}\binom{n}{\frac{n+1}{2}}\binom{N-n-1}{M-\frac{n+1}{2}}\binom{N}{M}^{-1} \tag{5}$$

for all $M>0$ and for all $n = 1,3,...,2M-1$.

Finally, by collecting the formulae (3), (4) and (5) we get the desired result (1).

■

In the next section we will predict the final number M of votes for loser by the Maximum Likelihood.

4 Predicting the final number of votes for winner and loser

Let us recall that under given leadership time n the Likelihood Function is a function of the unknown parameter M defined by the formula

$$\mathscr{L}_n(M) = P_M(T = n)$$

and the Maximum Likelihood predictor $\widehat{M}(n)$ of M is defined by the argument of $\mathscr{L}_n$ realizing its maximum. This may be expressed more precisely in the form

$$\widehat{M}(n) = \arg \max_{M \in \{0,1,\ldots,\frac{N-1}{2}\}} \mathscr{L}_n(M).$$

We shall prove

Theorem 2 *The Maximum Likelihood predictor $\widehat{M}(n)$ of the final number of votes for loser based on the leadership time n is given by*

$$\widehat{M}(n) = \begin{cases} 0, & \text{if } n = N, \\ \frac{N-1}{2}, & \text{if } n < N \end{cases}$$

while the ML predictor of the final number of votes for winner is given by

$$\widehat{N-M}(n) = \begin{cases} N, & \text{if } n = N, \\ \frac{N+1}{2}, & \text{if } n < N. \end{cases}$$

Proof. For $n = N$ the probability P_M attains its maximum for $M = 0$ and hence $\widehat{M}(N) = 0$.

For all odd positive integers $n < N$ the Likelihood Function is defined by the formula

$$\mathscr{L}_n(M) = \begin{cases} \frac{2}{n} \binom{n}{\frac{n+1}{2}} \binom{N-n-1}{M-\frac{n+1}{2}} \binom{N}{M}^{-1}, & \text{for } M \in \{\frac{n+1}{2}, \ldots, \frac{N-1}{2}\}, \\ 0, & \text{otherwise} \end{cases}$$

and the ML predictor $\widehat{M}(n)$ of M is given by

$$\widehat{M}(n) = \arg \max_{M \in \left\{\frac{n+1}{2}, \ldots, \frac{N-1}{2}\right\}} \mathscr{L}_n(M).$$

We will show that in this case

$$\arg \max_{M \in \left\{\frac{n+1}{2}, \ldots, \frac{N-1}{2}\right\}} \mathscr{L}_n(M) = \frac{N-1}{2}.$$

To this aim we only need to verify that

$$\frac{\mathfrak{L}_n(M+1)}{\mathfrak{L}_n(M)} > 1 \tag{6}$$

for all integers M belonging to the interval $\left[\frac{n+1}{2}, \frac{N-1}{2}\right)$.

Indeed, for such M

$$\frac{\mathfrak{L}_n(M+1)}{\mathfrak{L}_n(M)} = \frac{(M+1)(N-M-k)}{(M-k+1)(N-M)}, \text{ where } k = \frac{n+1}{2}.$$

Thus the condition

$$\frac{\mathfrak{L}_n(M+1)}{\mathfrak{L}_n(M)} > 1$$

may be presented in the form

$$(M+1)(N-M-k) > (M-k+1)(N-M).$$

Since the last inequality holds for all integers $M \in \left[\frac{n+1}{2}, \frac{N-1}{2}\right)$, the desired condition (6) was verified.

Reassuming, the ML predictor of the final number of votes for loser based on the leadership time n is given by

$$\widehat{M}(n) = \begin{cases} 0, & \text{if } n = N, \\ \frac{N-1}{2}, & \text{if } n < N. \end{cases}$$

Finally, by the well-known fact that the results of the ML estimation do not depend on the parametrization (see, for instance, Schervish (1995, Th. 5.28, p. 308)) we get the predictor of the final number of votes for winner in the form

$$\widehat{N-M}(n) = \begin{cases} N, & \text{if } n = N, \\ \frac{N+1}{2}, & \text{if } n < N. \end{cases}$$

■

Therefore, the ML predictor of the final number of votes for winner does not depend on the first leadership time n unless $n = N$. This leads to the following conclusion.

5 Conclusion

The first leadership time is informative for the final results of the election only in the trivial case.

Acknowledgements

The author thanks the reviewers of the manuscript for their valuable suggestions.

REFERENCES

AZZALINI, A., (1996). Statistical Inference based on the Likelihood, Chapman & Hall, London, UK.

BRÉMAUD, P., (1994). An Introduction to Probabilistic Modelling, 2nd ed., Springer-Verlag, New York.

FELLER, W., (1968). An Introduction to Probability Theory and its Applications, Vol. 1, 3rd ed., Wiley, New York.

GOULDEN, I. P., SERRANO, L. G., (2003). Maintaining the spirit of the reflection principle when the boundary has arbitrary integer slope, J. Comb. Theory Ser. A, 104, pp. 317–326,

LENGYEL, T., (2011). Direct consequences of the basic Ballot Theorem, Statist. Probab. Lett., 81, pp. 1476–1481.

RENAULT, M., (2007). Four proofs of the Ballot Theorem, Math. Mag., 80, pp. 345–352.

SCHERVISH, M. J., (1995). Theory of Statistics, Springer-Verlag, New York.

STĘPNIAK, C., (2015). On distribution of leadership time in counting votes and predicting winners, Statist. Probab. Lett., 106, pp. 109–112.

TAKACS, L., (1997). On the ballot theorem, In: Advances in Combinatorial Methods and Applications to Probability and Statistics, Birkhauser, pp. 97–114.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 359–376, DOI 10.21307/stattrans-2018-021

THE WELLBEING EFFECT OF COMMUNITY DEVELOPMENT. SOME MEASUREMENT AND MODELING ISSUES[1]

Włodzimierz Okrasa[2], **Dominik Rozkrut**[3]

ABSTRACT

The two interconnected methodological tasks – measurement and modeling – become especially challenging in the context of exploration of the interaction between the local community development and individual wellbeing. In this paper, the preliminary results illustrate usefulness of an analytical framework aimed to assess an impact of the local development on individual wellbeing through multilevel modeling, accounting for spatial effects is. To this aim, a dual measurement system is employed with data from two independent sources: (i) the Local Data Bank (LDB) for calculating a multidimensional index of local deprivation (MILD), and to capture variations in geographically embedded administrative units, communes (the country's finest division), and (ii) the Time Use Survey data to construct the U-index ('unpleasant'), considered as a measure of individual wellbeing. Since one of the implications of the main hypothesis on the interaction between community development and individual wellbeing was the importance of 'place' and 'space' (effect of neighborhood and proximity), a special emphasize has been put on spatial effects, i.e. geographic clusters and spatial associations (autocorrelation, dependence The evidence that place and space matter for this relationship provides support for validity of both multilevel and spatial approaches (ideally, combined) to this type of problems.

I. Introduction

Background and problem

Although the view that *place* and *space* matter for both the community and individual wellbeing is widely shared among the analysts and experts interested in their improvement (separately or jointly), a little effort has been done so far to determine what type of functional form describes the relationships or mutual influence between the two kinds of wellbeing, including their spatial patterns and factors of dynamics. This research was motivated as much by the knowledge gap in the literature concerning this methodological issue, including the ways of

[1] This article is based on the presentation "The Time Use Data-based Measures of the Wellbeing Effect of Community Development" at the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference, Washington DC, March 7-9, 2018.

[2] Cardinal Stefan Wyszyński University in Warsaw and Statistics Poland. Poland.
[3] Statistics Poland. Poland.

parameterization of these relationships, as also by the policy practitioners' demand (addressed to statisticians) for tooling devices to better allocate the scare resources to local communities while accounting for individual wellbeing. It embraces exploration of the relationships between subjective and objective measures of wellbeing at micro- (individual) and macro- (community) level while accounting for their cross-level operating factors in the presence of spatial effects, including quality of the place and spatial dependency.

Since the relationship between community wellbeing (CWB) and individual wellbeing (IWB) is of particular interest here, the three kinds of intertwined issues must be addressed concurrently: measurement – data – models. The measurement problem is complicated by the fact that, as noted by Gibson (2016), there is no theoretical justification for maximizing either happiness or life satisfaction, because neither corresponds to *utility* (p. 439). *'Happiness* is not all that matters, but first of all, it does matter (…), and second, it can often provide useful evidence on whether or not we are achieving our objectives in general' (Sen, 2008). However, an alternative approach, *Sen's capability approach,* which stresses priority of *functionings* and *capabilities* instead of resources or utility is becoming more useful also for policy purposes (Alkire, 2015): "The need for identification and valuation of the important functionings cannot be avoided by looking at something else, such as happiness, desire fulfillment, opulence, or command over primary goods' (Sen 1985 – in Alkire, op cit., p.1).Therefore, different information sources, including subjective data, can provide better insights on values and perceptions of people.

Within such a type of analytical framework, an ideal strategy seems to be the multilevel spatial modeling. However, some restrictions related to availability of data – which are here *ad hoc* combined from different sources instead to be generated by design to have the appropriate nested (hierarchical) structure - the cross-level modeling methodology will be illustrated below in a simplified version. Both types of possible strategies are explored and will be demonstrated as complementary to each other, 'interactive' and 'structural'. The former being focused on assessing the effect of interaction in searching for sources of variability at both individual and community levels. The latter is aimed at identifying causal *mediator* in searching for sources of influence (direct and indirect impact).

Analyses conducted in this paper use the multi-source database constructed through 'integrating' data – i.e. matching them on the ground of commune (*gmina*) – from three different sources: the Local Data Bank (public data file), the Time Use Survey (TUS, carried out by the Central Statistical Office in 2013), and Social Diagnosis (representative survey conducted this same year by an independent academic consortium). The measures derived from these data sets made it possible to explore spatial patterns of associations, autocorrelations and the dependency between measures of local deprivation (gmina-level) and the TUS-based indicators of wellbeing (the so-called index of 'unpleasant state', U-index). Although the results are preliminary and hardly robust - given incidental rather than natural hierarchically structured spatially distributed data, used in this study - they firmly support the adopted approach, i.e. employing spatially integrated social research framework for both analytical and policy purposes as a 'good practice' (methodologically) whenever place and space matter.

The paper is structured as follows. In the next section presented are some conceptual and measurement issues of key variables. It is followed by discussion (in section 3) of cross-level interrelation models, along with empirical results of their application. The explicitly included spatial aspects and spatial analysis of geo-referenced data, along with preliminary results are discussed in section 4. The concluding section closes up the paper, with some suggestions on prospective directions of further investigations.

II. The conceptualization, measurement and modeling of wellbeing

Conceptualization, *operationalization* and the measurement of wellbeing typically start with questions *what?, how?* and *why?* Consequently, the three types of issues – *measurement, data* and *modeling*- need to be considered concomitantly. Such an approach is adopted in the form of a perspective of spatially integrated social research within a multilevel spatial analytical framework capable of guiding methodological choices for selecting and integrating the needed data from different sources

While focusing on functionings as things that people actually value, one may consider using data from time use survey in which respondent is asked to report what s/he did in the previous day - Day Reconstruction Method (Kahneman et al., 2004) [4]. Respondent makes also an assessment of the time spent on performing particular activity as pleasant or unpleasant (the so-called 'time of unpleasant state', Krueger et al., 2009). This approach is the key for constructing individual wellbeing measure here. It converges with conceptualizations of subjective wellbeing that take into account both positive and negative affectivity (Bradburn 1969) associated with the performed activity, and is now common in empirical research following international recommendations for measuring subjective wellbeing in public statistics (OECD 2013; NRS 2013; Kalton, Mackie, Okrasa 2015; Maggino 2017).

The key importance of community wellbeing in both research and policy considerations of the individual wellbeing determinants, especially in the development context (with clear distinction between local and regional development, e.g. Capello, 2009) is due to several reasons. Many of them have been recognized and discussed thoroughly in the literature, either as a part of the process or of outcome of such development, challenging the traditional use of GDP and other economic indicators as measures of social progress (Stiglitz et al., 2010, OECD, 2013, Kim and Ludwigs, 2017; Lee et al., 2015). Methods of community wellbeing assessment, including subjective aspects of wellbeing, are becoming standard tools for policy purposes in several countries (notably in Australia, Canada, USA and UK). They all have one feature in common, namely, they are based on self-reported feeling about selected aspects of wellbeing in connection with community, and community itself is among the components of the wellbeing measures.

One special feature of local community that affects its wellbeing in the development context is community cohesion. It is interpreted here in a broader

[4] **"**Functionings is a broad term used to refer to the activities and situations that people spontaneously recognize to be important" (Alkire, op cit, p. 3-4).

sense than the latter – hence termed *spatial cohesion* - due to embracing all other types of cohesion: social, economic or territorial cohesion, which are typically considered among the goals of the European Union's development policies and studies (focused often on so-called *β-convergence and σ-convergence,* respectively). Usually, it is meant consistently with classical interpretation of the term, e.g. following Forrest's and Kearns' (2001) specification of the component topic areas: (I) *common values and a civic culture*, (ii) *social order and social control*, (iii) *social solidarity and reduction in wealth disparities,* (iv) *social networks and social capital*, and (vi) *place attachment and identity* (p. 2129). The last one is of special interest here due to focusing on „…creating relationships between individuals, about empowering the individual as well as local communities" (Kearn and Forrest, 2000), and is assumed here as being covered by the measures of subjective community wellbeing. This aspect will be briefly explored with data from Social Diagnosis, a biannual survey of attitude and wellbeing on a large nation-wide representative sample.

As regards modeling of multilevel *relationships* **–** between ***individual*** and community wellbeing – two approaches are employed here (Okrasa, 2017). One is a between level *interaction***-**focused approach concentrated on decomposition of variance into *within* group (differences among individuals in community, level – 1) and *between groups* (communes, level-2), *reflecting* differences across communities. To this aim, models for hierarchically structured data seem appropriate, which however are not free of a risk of 'ecological fallacy' (Goldstein, 2003(2010); Subramanian, 2009). In a parallel way, there is a 'causal' type of modeling checked as well. Specifically, we employ *structural* modeling of *(causal) mediation mechanism,* which consists of decomposition of total effect of the independent ('treatment') variable into the natural *direct* and *indirect* effects (Hong, 2015). Within this approach, community wellbeing can be hypothesized as a mediating factor between an objective (material) status of a person and her subjective wellbeing.

It was also hypothesized that in addition to the characteristics of a locality (place/commune) itself, spatial relations, proximity (distance) have impact not only on both the level of relevant measures – i.e., on both individual and community wellbeing – but also on the character of the relationship between them. Consequently, spatial (dependence) analysis is explicitly applied too. However, given the nature of the problem involving estimation of the impact of space on relation between variables rather than of their parameters we do not employ *spatial statistics* in version of *model-based* strategies, i.e. SAE/*Small Area Estimation* (Rao and Molina, 2015). Given also character of available data, a spatial *econometric* version of *data-exploration* was applied using *data-driven* strategies for analyzing patterns in geo-referenced data. Specifically, GeoDa (*Spatial Data Analysis* for non-GIS data, Anselin, 1995, 2005), and ESDA (*Exploratory Spatial Data Analysis*, Fischer and Getis, 2009) were used to this aim.

Data and measures of wellbeing

In order to analyze individual (subjective) wellbeing and quality of the living environment, community wellbeing, a multi-source analytical database was constructed which contains data from *Time Use Survey* 2013 (TUS) and the *Local*

Data Bank (LDB), and data from *Social Diagnosis* 2013 (SD). The procedure for integrating the data sets in the multisource analytical database (MADb) was based on the geographical information, i.e. X,Y coordinates of the locations (gminas) from which the respondents were drawn to the respective surveys (TUS and Social Diagnosis). Initially, it was arbitrary decided that 20 persons is the minimum number of respondents needed to be identified in a gmina to have it included to the MADb[5]. But for some calculations 10 persons were also used.

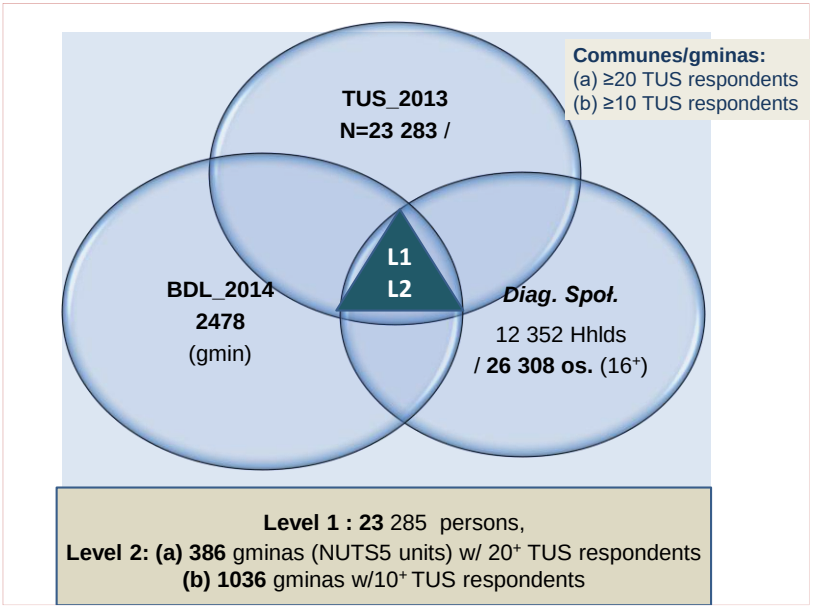


Figure 1. Multi-source analytical database: BDL & TUS (& DS). The NUTS5 unit's (commune's/gmina's) territorial KODTERYT was used as key merging code (X, Y– *coordinates)*

Community wellbeing – *Multidimensional Index of Local Deprivation* (MILD)

An objective measure of community wellbeing was applied to calculate the level of local deprivation of each of 2478 communes (gminas) using data from public file, Local Data Bank. The index – Multidimensional Index of Local Deprivation (MILD) – is composed of 11 domain-specific scales constructed by confirmatory Factor Analysis (each domain was pre-defined in a single-factor version of the FA, Okrasa 2013b[6]). The following domains of deprivation are

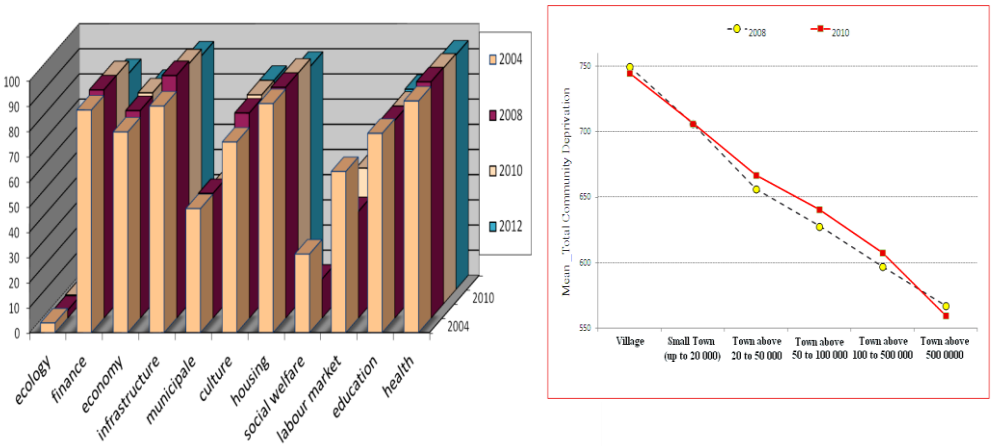
[5] The 20 persons cluster drops into interval 15-30 persons which is most often used in multilevel analysis under the rule of thumb, a rationale for which is that it satisfies the requirement of sufficient sparseness in defining a 'synthetic neighborhood' (Clarke and Wheaton 2007).

[6] The selection procedure consisted of: selection of domains – selection of indicators within each of the areas on the basis of factor analysis (principal component analysis) – standardization in the indicators – aggregation in the index for a given area – normalization of indicators for each area – composite aggregation in the global index (Okrasa 2013).

included: ecology, finance, economy, infrastructure, municipal utilities, culture, housing, social welfare, labor market, education, and health. Since Cronbach's alpha exceeded. 75, they were combined into as synthetic measure, MILD. As suggested by the term 'deprivation' all the component scales and composed of them MILD are of negative type measure (destimulants): the higher the index (scale) value the worse the community situation with respect to a given domain or to the total local deprivation (MILD). The values of MILD are strongly place dependent, decreasing sharply as moving from rural to urban areas, and along with the growing size of town – see Figure 2 a and b.

Figure 2. Multidimensional Index of Local Deprivation (MILD) by

(a) size of place of living (b) component scales over years 2004-2012



There was also a subjective community well being measure calculated using data from Social Diagnosis – on the basis of answers to questions about satisfaction from selected aspects of quality of life in a community: (1) locality (place), housing, and security (LHS); (2 social relations in family and in neighborhood, life achievements, and self-esteem (FSE); (3) life perspective while living 'in here', financial situation, and work possibilities (LPH). While regressed on the local deprivation (MILD), all these measures showed to negatively associated. It should be noted, however, that some items expressing community subjective wellbeing, such as 'sense of belongingness' or 'place attachment and identity" and so on, are also present among the items constituting scales of community cohesion.

Individual (subjective) wellbeing – the Time Use Survey data-based U-index

Following various definitions of individual (subjective) wellbeing there is variety of well advanced measures proposed in the literature. Nevertheless, there are still some doubts raised by psychometricians concerning validity of particular scales, while some statisticians and econometricians express reservations toward employing strong analytical tools to ordinal-level measurement data, as most of

the scales is built up of the Likert-type items. Therefore, an alternative approach consisting of use of the time use survey data (usually collected with the day reconstruction method – Kahneman et al., op cit. 2004) met recently with growing interest. Especially, the TUS-based methodology developed (notably, due to Krueger et al., 2008) which combines objective information about the time respondent spent on performing activities with subjective rating of feeling associated with this performance. In the TUS conducted by the Central Statistical Office in 2013 three-point scale was used: 'positive' – 'neutral' – 'negative'. In accordance with the above methodology, U-index is defined as follows::

$$U_i = \Sigma_j I_{ij} h_{ij} / \Sigma_j h_{ij} \quad \text{(where I = -1 or 0 or +1)}$$

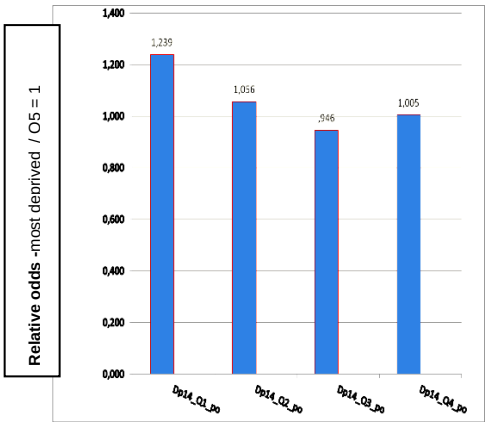
and $$U = \Sigma_i(\Sigma_j I_{ij} h_{ij} / \Sigma_j h_{ij}) / N \quad \text{for N-persons (in population)}$$

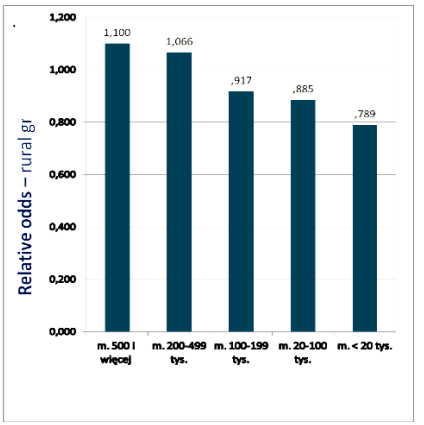
In calculation, it shown that the share of time spent for performing negatively rated activities was relatively low for most of the performed activities, hence to the U-index included were also 'neutral' cases – so, its interpretation should be rather as reflecting 'non-positive' than 'unpleasant state'.

At a glance, the relationship between objective CWB, as measured by MILD, and the U-index for all activities (excluding sleep) can be characterized by the relative odds of the U-index for MILD-quintiles of communes (gminas) – see Fig. 3, where the highest quintile, i.e. the 'most deprived' communes is set for the reference category. It suggests a tendency to generally bigger chance of being discontent due to spending relatively more time in 'non-positive state' among people living, on average, in more affluent communes (though the tendency is not strictly linear).

Figure 3. Odds of experiencing 'non-positive' feeling associated with activities, U - index, depending on

(a) the level of *local deprivation*/MILD (b) the size of the living place

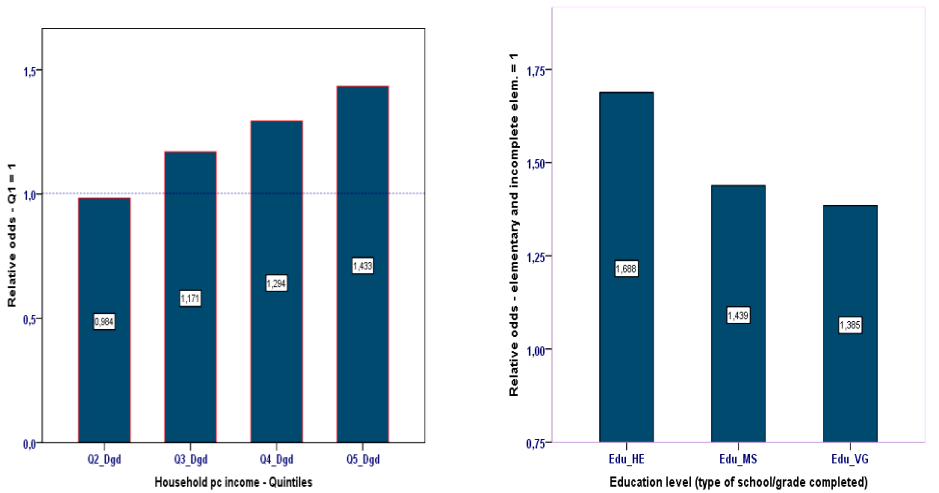




The negative pattern of tendencies – i.e. residents in more affluent urban environment (commune) are on average less satisfied with their life (in term of the U-index) than in less developed rural areas and small towns – should be interpreted with some caution due to the fact that they may perform different type of activities in different environments. For instance, shares of highly disliked activities associated with work or learning or house maintenance may be higher among city and big town dwellers, while shares of such activities like leisure time or social life or physical exercise or hobby and other performed on non-obligatory basis can be proportionally bigger among residents in small towns and rural areas. Validity of such observations can be supported indirectly by looking at some personal level characteristics which also are strongly related to the kind and size of the living place like income and education – Fig. 4, below.

Figure 4. Odds of experiencing 'non-positive' feeling associated with performed activities, U - index, depending on

(a) level of household pc income (b) level of education



The emerging pattern of tendencies presented by the above figures suggests that behind a given level of local deprivation (i.e. MILD score, used here as the key indicator of CWB) operates a pretty consistent configuration of place-related factors: Urban (rather big) areas populated by on average better educated and wealthier people, who also seem to function in qualitatively different way (e.g. they are more likely to engage in activities which are generally less valued than those performed by dwellers in apparently less displeased rural areas and small town).

Community deprivation, community cohesion, and individual wellbeing. The following questions were asked in the analysis prior to multilevel modeling:

- Does the level of community deprivation /MILD affect the measures of community cohesion?

- Does the level of community cohesion influence the level of individual wellbeing (U-index)?
- How community deprivation and community cohesion affect jointly the individual wellbeing (U-index)?

Community cohesion – meant here synonymously with the Community Subjective Wellbeing (CSWB) – entails 3 scales calculated from the Social Diagnosis data and concerning satisfaction from three aspects of life in the community: 1. Locality, housing, security (LHS); (2. Social relations in family and in neighborhood, and life achievements (FSE) 3. Life perspective while living where s/he lives ('in here', LPH). Regressed on the local deprivation (MILD) all these measures remain with it, as it could be expected, in negative relation – Table1.

Table 1. Influence of the community level of (overall) deprivation (MILD_2014) on the measures of subjective of community wellbeing/CSWB

Predictor:	1. Locality etc/LHS	2. Social relations /FSE	3. Life perspective 'here'	4. IWB/U-index (all activities)
Community deprivation/MILD_2014	- 0.027**	-0.120**	-0.237**	-0.034**

**) significant at $p < 0.01$

The relationships between the three datasets-based measures – CWB (in terms of local deprivation/MILD-2014), community cohesion (SD-based scales, used in Table 1 as measure of subjective-CWB), and individual wellbeing (U-index) were preliminary explored to determine the influence of the former two variables on the latter – results are in Table 2.

Each of the three measures of community cohesion (or subjective community wellbeing/S-CWB) – that was negatively associated with local deprivation (MILD-2014, in Table 2) – remains in also inverse relations with individual wellbeing. The U-index is consistently negatively affected by locality (place), housing, and security (LHS); by social relations in family and in neighborhood, (FSE), and by life perspective (LPH). However, the interaction effect, i.e. joint influence of such combination like, say, high gmina's deprivation and high level of satisfaction from own locality (gmina) is generally positive: higher (lower) satisfaction from their localities of the residents in better-off (worse-off) gminas reinforces the impact of the latter on lowering (increasing) the level of their displeasure (U-index).

Table 2. Regression of individual well-being – U-index – on *community deprivation* (MILD) *and community cohesion* (CSW-B)

Model	St. coeff	t	Signific			
Predictors:	**Beta**			**df**	**F**	**Signif.**
I						
(Constant)		-1,530	,126	3 22368	16,180	,000
Comm. deprivation/MILD_2014	,188	3,239	,001			
FSE – Soc. relations, family and neighborhood, self-esteem	-,170	-3,653	,000			
Interaction MILD*FSE	,304	3,893	,000			
II						
(Constant)		,379	,704	3 22368	15,972	,000
Comm. deprivation/MILD_2014	,105	1,787	,074			
PPH – Life perspective – 'in here'	-,118	-2,694	,007			
Interaction MILD*PPH	,181	2,444	,015			
III						
(Constant)		-1,530	,126	3 22368	13,607	,000
Comm. deprivation/MILD_2014	,188	3,239	,001			
LHS – Locality, housing, security	-,170	-3,653	,000			
Interaction MILD*LHS	,304	3,893	,000			

III. The cross-level interplay – issues in modeling

Effect of interaction

Multilevel modeling of individual wellbeing and community wellbeing starts with **a b**asic structure of a model to deal with cross-level relationships, which should have the following elements and features (following Goldstein 2003, Subramanian 2010, and Okrasa 2017):

- y_{ij}; wellbeing of *i* individual in *j* commune/gmina;
- x_{1ij} predictor of individual (level-1) – such as: income, age, education, or satisfaction (e.g. from life in a community, family life, etc.
- predictor of level-2/(macro-level) – CWB, here Multidimensional Index of Local Deprivation for j-gmina; $MILD_j$

Model for level-1: $y_{ij} = \beta_{0j} + \beta_{I}\, x_{1ij} + e_{0ij}$ (1)

where: β_{0j} – refers to x_{0ij} average score on a wellbeing scale in j-th commune /gmina; (e.g. 'less affluent' or 'more disadvantageous', etc.', < Me, x_{0ij} =1);

β_{I} – average differentiation of individual wellbeing associated with individual material status, (x_{1ij}), across all gminas;

e_{0ij} – residual term for the level-1.

Treating β_{0j} as random variable: ($\beta_{0j} - \beta_0$) + $u_{0j,}$ where u_{0j} is locally-specific associated with average value of β_0) for a specified group (e.g. less satisfied from a community) and grouping them into fixed and random part components ($e_{0ij} + u_{0j}$) we obtain *variance component model,* or *random-intercept* model:

$$y_{ij} = \beta_0 + \beta_{I}\, x_{1ij} + (e_{0ij} + u_{0j}) \quad (2)$$

Modeling *fixed-effect* we include a level-2 predictor – MILD – (index of local deprivation) along with individual characteristics, including *interaction* term between the two levels *characteristics*

$$\beta_{0j} = \beta_0 + \alpha_1 MILD_{1j} + u_{0j} \quad (3)$$

$$\beta_{1j} = \beta_1 + \alpha_2 MILD_{1j} + u_{1j} \quad (4)$$

where $MILD_{1j}$ – context variable, predictor of differences between gminas.

Two-level model can be specified as below (following Subramanian, op cit., p. 520-21):

$$y_{ij} = \beta_0 + \beta_{I}\, x_{1ij} + \alpha_1 w_{1j} + \alpha_2 w_{1j}\, x_{1ij} + (u_{0j} + u_{1j}\, x_{1ij} + e_{1ij}\, x_{1ij} + e_{2ij}\, x_{2ij}) \quad (5)$$

where w_{1j} is a 2-level predictor, i.e. the index of local deprivation, $MILD_{1j}$.

According to the above structure, $\boldsymbol{\alpha_1}$ provides an estimate of the (marginal) change in individual wellbeing (U-index) for a unit change in the level of gmina's deprivation for those below the median, or not in the 'unpleasant state'; while $\boldsymbol{\alpha_2}$ estimates the extent to which the marginal change in subjective wellbeing (U-index) for unit change in the gmina deprivation index (MILD) differs from that for those in the 'unpleasant state'.

Formally, such a specification of cross-level (between individual and community/gmina measures of wellbeing) *modification* or *interaction* effect should ensure robust estimation (e.g. Subramanian, op cit., p. 521, Hox et al., 2018). However, as already noted, the available data related limitations impose some restriction on the exactness of the employed calculation strategy. Therefore, the following model was calculated using data from Time Use Survey:

$$IWB(U\text{-}index)_{ij} = \beta_{00} + \beta_{10} education_{ij} + \beta_{20} age_{ij} + \alpha_1 MILD_j + \alpha_{11} education_{ij} * MILD_j$$

$$+ \alpha_{21} age_{ij} * MILD_j + u_{1j} education_{ij} + u_{2j} age + u_{0j} + e_{ij} \quad (6)$$

Preliminary results are in Table 3.

Table 3. Multilevel regression of individual wellbeing – *U-index* for all activities – on individual and commune/*gmina* level variables with cross-level interaction terms.

Model: predictors	Weekdays		Weekend /holiday	
	Beta	**t**	**Beta**	**t**
Constant	(.726)**	(6.316)	(.333)**	(3.515)
Education	-.085	-1.136	-.089	-1.209
Age	-.299**	-4.015	-.008	-.105
Multidimensional Index of Local Deprivation /MILD_2014	-.098*	-2.556	-.046	-1.209
Education * MILD_2014	.142*	1.900	.145*	1.97
Age * MILD_2014	.115	1.497	-.029	-.383
Urban (rural omitted)	.011	1.280	.016*	1.966
	F (6.22698) = 174.860**		F (6.24068) = 23.515**	

Since the additional but crucially important focus was on spatial aspects of the relationships (interactions), the working strategy shown to be in practice spatial regression with both level variables included in the respective equations, as an explicitly interaction term. Neglecting for the time being this path of analysis, the next modeling issue concerns searching for a causal mediating mechanism.

IV. Bringing *space* into the question

Estimation of the spatial regression model parameters (notation for individual observation *i*):

$$y_i = \rho \sum^{n}_{j=1} W_{ij}\, y_j + \sum^{k}_{r=1} X_{ir}\, \beta_r + \varepsilon_i \qquad (7)$$

where: y_i – the dependent variable for observation *i*; X_{ir} *k* – explanatory variables, r = 1, …, k with associated coefficient β_r; ε_i is the disturbance term; ρ is parameter of the strength of the average association between the dependent variable values for region/observations and the average of them for their neighbors (e.g. LeSage and Pace, 2010, p. 357)

The above specification of the spatial regression model assumes that ε_i is meant as the *spatially lagged* term – versus *spatial error* formulation – for the dependent variable (which is correlated with the dependent variable), that is:

$$\varepsilon_i = \rho\, W_i . y_i + X_i . \beta + \varepsilon_i \qquad (8)$$

These two types of models allow us to examine the impact that one observation has on other, proximate observations. The results in Table 4, below, are for the spatial error model.

Table 4. Spatial dependence/spatial regression of SW-B (U-index) on commune's attributes and compositional characteristics

***SPATIAL ERROR MODEL* - MAXIMUM LIKELIHOOD ESTIMATION**
Dependent Variable : *U –index* Number of Observations: 937; Mean dep var : 0.361195 Number of Variables : 8;
Degrees of Freedom : 929
Lag coeff. (Lambda) : 0.431769 ; R-squared : 0.123681

Variable	Coefficient	Std.Error	z-value	Probability
CONSTANT	0.523731	0.042847	12.2233	0.00000
MONTHLY INCOME	-0.002730	0.001960	-1.40359	0.16044
AGE_avg (%)	-0.014313	0.005653	-2.53177	0.01135 *
EDUCATION_HS+ (%)	0.000381	0.000222	1.71849	0.08571 *
NOT WORKING POP. (%)	-0.001304	0.000273	-4.77623	0.00000 *
ILD_ECOLOGY	0.000560	0.000462	1.21309	0.22510
ILD_SOCIAL POLICY	-0.000415	0.000312	-1.32693	0.18453
SUBSIDIES_pc	1.2323e-005	1.1588e-05	1.06344	0.28758
LAMBDA	0.431769	0.0677941	6.36883	0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY -- RANDOM COEFFICIENTS

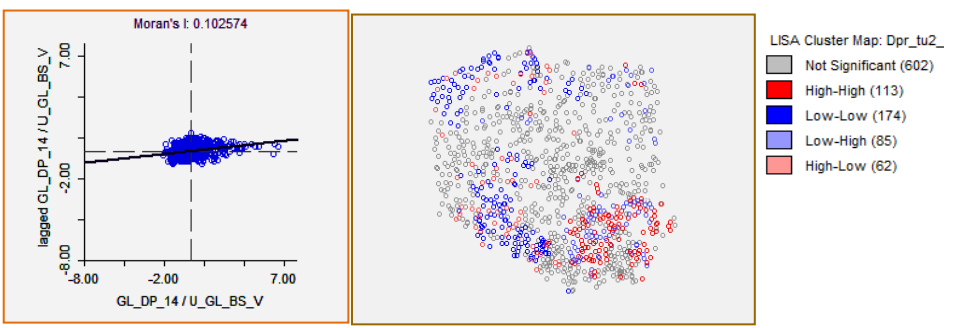
TEST	DF	VALUE	PROB
Breusch-Pagan test	7	54.7759	0.00000

DIAGNOSTICS FOR SPATIAL DEPENDENCE -- SPATIAL ERROR DEPENDENCE

TEST	DF	VALUE	PROB
Likelihood Ratio Test	1	36.1346	0.00000

Few variables that represent commune's compositional characteristics (average percentage) influence significantly the individual (subjective) wellbeing in a negative way: age, education, population not in work is the factors operating in space dependent manner. Other two – monthly income and deprivation in the domain of social policy – which also affect residents' wellbeing negatively (though not in statistically significant way) indicate important direction of further exploration and of clarification from the development policy standpoint accounting for spatial aspects. Some illustration is given below, in Fig. 6 and 7, following presentation of the scatter plot and map jointly for subjective wellbeing (U-index) and local community deprivation (MILD), Fig. 5.

Figure 5. Individual SWB/*U-index* (all activities) and the level of commune (gmina) local deprivation/$MILD_{2014}$. Moran I = 0.103



Compared to earlier results concerning the relationship between community (deprivation) and subjective wellbeing (according to U-index) addition of spatial aspects to its exploration brings clarification with respect to the question of where there are low-low or high-high levels of its occurrence.

While checked in a separate way, the spatial association between some of the above variables and subjective wellbeing in a particular type of activity (U-index) indicate different direction. For instance, local deprivation in social policy and U-index for 'caring for children', below Fig. 6, or deprivation in local labor market and U-index for commuting (work and other 'target places'), Fig. 7.

Figure 6. U-index for caring for children by the level of local deprivation in social policy. Moran I= 0.18

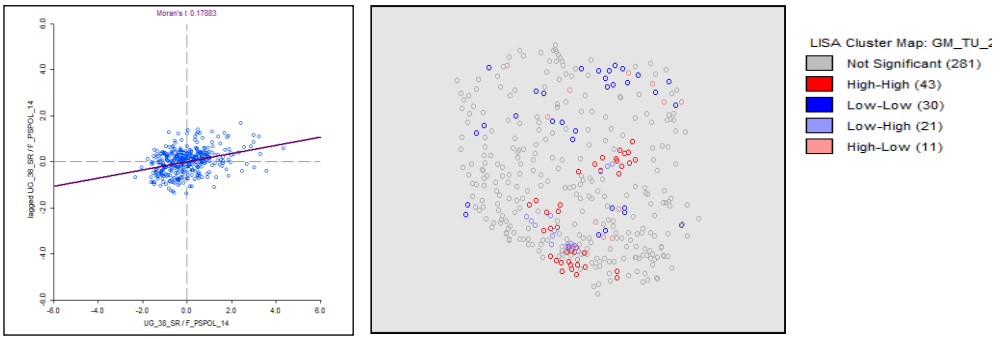
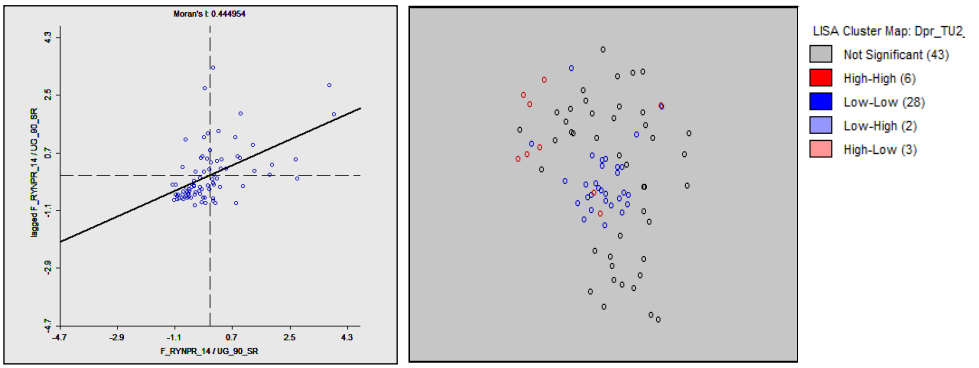


Figure 7. Individual wellbeing/U-index for commuting, i.e. associated with traveling to work and similar (target places or commuting), and gmina's (local) deprivation in the domain of local labor market (Masovian) Moran-I = 0.44)



Several conclusions can be drawn from the above patterns of spatially related association between quality of local environment (community constituting household's immediate surroundings) and subjective wellbeing, two of them are

worthwhile to note here. First, more specifically defined relations – for concrete type of activities and of domains of local deprivation – can be analyzed in the multilevel spatial analytical perspective more effectively than using synthetic measures. Also, lower level of territorial cross-section values rather than countrywide global values is more appropriate in search for identification of spatially dependent phenomena and their interconnections.

Conclusion

In view of the doubts and critique coming from experts of different disciplines - including psychometricians and econometricians – concerning the measurement of individual (subjective) wellbeing, the Time Use Survey data seem to provide a unique opportunity to explore relationships between individual and community wellbeing, using at the same time public statistics files created for other purposes.

In general, the level of dissatisfaction accompanying the performance of everyday activities – experiencing 'unpleasant state' and lower subjective wellbeing – increases along with greater household income. Paradoxically enough, individual wellbeing is diminishing (U-index grows) along with the lower level of commune's local deprivation. [In other words, overall conditions in less developed gminas constitute in general more favorable environment for individual (subjective) wellbeing – such aspects like social interaction, interpersonal relations might be of importance].

Community wellbeing reinforces significantly the subjective wellbeing effect of individual income. Since the influence of CWB on individual wellbeing is on average quite visible also in spatial terms – due to a tendency to cluster amongst gminas which are high-high or low-low on both dimensions – there is a need to analyze further such relationships, assuming availability of the appropriate data.

REFERENCES

ALDSTADT, J., (2010). Spatial Clustering. Chapt B.4. [in] Fischer M. M., Getis A., Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications, Springer.

ALKIRE, S., (2015). The Capability Approach and Well-Being Measurement for Public Policy, Oxford Poverty & Human Development Initiative (OPHI) Working Paper, No. 94.

ANSELIN, L., SYABRI, I., KHO, Y., (2010). GeoDa: An Introduction to Spatial Data Analysis. Chap. A.4 , [in] Fischer M. M., Getis A. Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications, Springer.

ARCAYA M., BREWSTER, M, ZIGLER, C M., SUBREMANIAN, S, V., Area Variations in Health, A Spatial Multilevel Modelling Approach. Health Place. 18 (4), pp. 824–831

BERNINI, C., GUIZZARDI, A., ANGELINI, G., (2013). DEA-Like Model and Common Weights Approach for the Construction of a Subjective Community Well-Being Indicator, Soc Indic Res 114.

BRADBURN, N., (1969). The Structure of Psychological Wellbeing. Aldine. Chicago.

CHAVIS, D. M., LEE, K. S., ACOSTA, J. D., (2008). Sense of Community (SCI), Lisboa, Portugal.

CAPELLO, R., Space, growth and development. Capello, R., Nijkamp, P., (eds) Handbook of Regional Growth and Development Theories. Edward Elgar, Cheltenham, UK. 2009 Clarke, Ph., Wheaton, B., 2007. Addressing Data Sparseness in Contextual Population Research Using Cluster Analysis to Create Synthetic Neighborhoods. Sociological Methods & Research, Vol. 35, No. 3, February 2007, 2007 Sage, pp. 311–351,

CORRADO, L., FINGLETON, B., (2011). Multilevel Modelling with Spatial Effects. Strathclyde Discussion Papers in Economics, No. 11–05, Department of Economics University of Strathclyde, Glasgow.

FISCHER, M. M., GETIS, A., (2010). Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications. Springer.

FORREST, R., KEARNS, A, (2001). Social Cohesion, Social Capital and the Neighbourhood, Urban Stud 2001, 38, 2125.

GIBSON, J., (2016). Poverty Measurement: We Know Less than Policy Makers Realize. Asia & the Pacific Policy Studies, Vol. 3, No. 3, pp. 430–442. DOI: 10.1002/app5.141.

GOLDSTEIN, H., (2003). Multilevel Statistical Models. Edward Arnold, Loondon.

GUS (Central Statistical Office), (2015). Time Use Survey, 2013, GUS, Warszawa.

HEWES, S., BUONFI, A., ALI, R., MULGAN, G., (2010). Cohesive communities – the benefits of effective partnership working between local government and the voluntary and community sector, The Young Foundation IDeA.

HONG, G., (2015). Causality in a Social World: Moderation, Mediation and Spill-over. Wiley.

HOX J. J., MOERBECK, M., SCHOOT, R. van de, (2018). Multilevel Analysis: Techniques and Applications, 3rd ed., N. Y. Routledge.

KAHNEMAN, D., KRUEGER, A. B., (2006). Developments in the measurement of subjective well-being, Journal of Economic Perspectives, 20, pp. 3–24.

KALTON, G., MACKIE, CH., OKRASA, W.,(eds.), (2015). The Measurement of Subjective Well-Being in Survey Research, Statistics in Transition new series. Vol. 16, No. 3.

KIM, Y., LUDWIGS, K., (2017). Measuring Community Well-Being and Individual Well-Being for Public Policy: The Case of thr Community Well-Being Atlas" [in] Phillips, R., Wang, C., (eds.), Handbook of Community Well-Being Reseach. Springer.

KRUEGER, A. B., KAHNEMAN, D., SCHKADE, D. A., SCHWARTZ, N., STONE, A., (2009). National Time Accounting: The Currency of Life, [in:] A. B. Krueger (ed.) Measuring Subjective Well-Being of Nations: National Account of Time Use and Well-Being. University of Chicago Press. Lee, S. J., Kim, Y., Phillips, R., (eds.), Community Well-Being and Community Development. Conceptions and Applications, 2015, Springer, http://www.springer.com/978-3-319-12420-9.

LESAGE, J. P., PACE, R. K., (2010). Spatial Econometric Models, [in] Fischer M. M., Getis A., op cit.

NATIONAL RESEARCH COUNCIL, (2013). Subjective Well-Being: Measuring Happiness, Suffering, and Other Dimensions of Experience. Panel on Measuring Subjective Well-Being in a Policy-Relevant Framework. A. A. Stone and C. Mackie (eds.), Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Washington, DC: The National Academies Press.

OECD, (2015). Better Life Index, OECD Publishing. Paris.

OECD, (2013). OECD Guidelines on Measuring Subjective Well-being, OECD Publishing.

OKRASA, W., (2017). Community Wellbeing, Spatial Cohesion and Individual Wellbeing – towards a multilevel spatially integrated framework, [in] W. Okrasa (ed.) Quality of Life and Spatial Cohesion: Development and Wellbeing in the Local Context. Cardinal Stefan Wyszynski University Press. Warsaw.

OKRASA, W., (2013). Spatial Aspects of Community Wellbeing. Analyzing Contextual and Individual Sources of Variation using Multilevel Modeling. Paper at th 59th World Statistics Congress, Hong Kong, August, pp. 25–30.

RAO, J. N. K., MOLINA, I., (2015). Small Area Estimation, 2nd ed., Wiley, Hoboken, New Jersey.

ROZKRUT, D., ROZKRUT, M., (2006). Analysis of the economic development of districts in Poland as a basis for the framing of regional policies. [in]: Spiliopoulou, M.; Kruse, R.; Borgelt, C.; Nürnberger, A.; Gaul, W. (eds.), From Data and Information Analysis to Knowledge Engineering – Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin 2006, pp. 518–525.

STIGLITZ, J., SEN, A., FITTOUSSI, J-P., (2010). Measuring our lives. New York.: The New Press.

SUBRAMANIAN, S. V., (2010). Multilevel Modeling [in] Fischer M. M., Getis A., Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications, Springer.

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 377

On behalf of the co-organizers of the Q2018 – the Statistics Poland (CSO) and Eurostat – we are pleased to inform that the 2018 European Conference on Quality in Official Statistics is being held on 26-29 June in Kraków, Poland.

"The Q2018 Conference will be one of the series of scientific gatherings covering methodological and quality-related issues that are relevant to the development of the European Statistical System".

http://ec.europa.eu/eurostat/web/ess/-/q2018-conference

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 379

The 2nd Congress of Polish Statistics organised on the occasion of the 100th anniversary of the establishment of the Statistics Poland will be held on July 10-12, 2018 in Warsaw.

"The Congress will last three days. The framework program of the event contains a series of thematic sessions, including a jubilee panel on the history of Polish statistics, as well as sessions devoted to Polish statistics on the international arena, methodology of statistical surveys, mathematical statistics, regional statistics, population statistics, social and economic statistics, statistical data issues and, also sports and tourism statistics.

In the Congress, which will emphasise the contribution of Poles to the global treasury of statistical knowledge, representatives of foreign institutions and scientific units will participate.
We are convinced that the Congress will constitute a unique opportunity for the representatives of official statistics, research centres and other partners involved in the study of social, economic and environmental processes to meet and exchange their knowledge, views and experiences."

https://kongres.stat.gov.pl/en/

STATISTICS IN TRANSITION new series, June 2018
Vol. 19, No. 2, pp. 381–385

ABOUT THE AUTHORS

Adepoju Abosede Adedayo is a Senior Lecturer and a former Acting Head of Department in the Department of Statistics, Faculty of Science, University of Ibadan. She has twenty years of teaching and research experience. Her research interests include econometric modelling, applied Bayesian analysis and economic/financial statistics. She is a member of many professional bodies like International Biometric Society (IBS) Group Nigeria, International Statistical Institute (ISI), Caucus of Women in Statistics of the American Statistical Association (ASA) and Nigerian Statistical Association (NSA).

Ayhan Öztaş H. is a Professor Emeritus at the Department of Statistics in Middle East Technical University, Ankara, Turkey. His research interests are survey sampling techniques, survey methodology research, and web survey methodology. Professor Ayhan has published more than 100 research papers in international/national journals and conferences. He has also published twelve books/monographs. Professor Ayhan is an active member of many scientific professional bodies.

Das Ujjwal is an Assistant Professor at Operation Management, Quantitative Methods and Information System Area of Indian Institute of Management, Udaipur. His research area includes high dimensional data analysis, analysis of missing data, time series modelling, and analyses of discrete and time to event data.

Ebrahimi Nader is a Professor at the Division of Statistics, Northern Illinois University. His research area includes life testing and reliability, survival analysis and actuarial science. He has published more than 120 papers since 1982 in renowned statistics journals and has served in the editorial board of some of them. As recognition of his research, he has been elected as a Fellow of the American Statistical Association. He is also an elected member of the International Statistical Institute.

Frątczak Ewa, dr hab. Professor of the Warsaw School of Economics – SGH. Head of Event History and Multilevel Analysis Unit. Research interest: statistics, business analytics and demography; longitudinal analysis (survival analysis, panel analysis, joint models for survival and panel data, survival predictive models, survival data mining); advanced methods of statistical analysis (multilevel models, mixed models, advanced models of logistic regression, advanced business models); theory and practice retrospective, panel and prospective survey. Author or co-author over 20 books in the field of statistics, demography and advanced methods of statistical analysis.

Grzenda Wioletta, PhD in Mathematics (2006). She is an Assistant Professor at the Institute of Statistics and Demography at Warsaw School of Economics since 2007. Her current research interests focus on modelling of socio-economic and demographic processes. She is an author of papers on the applications of

Bayesian and classical statistical methods in the analysis of fertility and labour market. She is also the author and co-author of books on Bayesian statistics, advanced methods of statistical analysis and SAS programming.

Khalil Alamgir is working as the Associate Professor of Statistics at University of Peshawar, Pakistan. His main areas of interest include robust statistics, regression analysis, survey sampling and time series data analysis. He has published 21 research papers in national/international reputed journals. He produced 13 M.Phil students in the field of survey sampling. He is an external reviewer of many journals.

Kosiorowski Daniel is an Associate Professor at the Department of Statistics Faculty of Management, Cracow University of Economics. His area of interest involves theory and applications of nonparametric and robust multivariate statistical analysis and functional data analysis. He is the author more than 70 research papers in international/national journals and conferences. He has also published three books/monographs.

Lumiste Kaur obtained his PhD in mathematical statistics from the University of Tartu in 2018. Presently he works as a Data scientist at Questro Analytics Ltd. in Estonia. During his PhD studies he worked on the European Social Survey and in the Estonian-Swedish Mental Health and Suicidology Institute. His primary research interest is in survey statistics, more specifically the use of auxiliary information in data collection and estimation stages.

Maqbool S. has been an Assistant Professor (Statistics) at Sher-e-Kashmir University of Agricultural Sciences and Technology-Kashmir (J & K), India, since 2007. He has published more than 120 research papers in reputed international and national journals. The area of specialization is sampling theory and mathematical programming. He has attended more than 30 international conferences.

Mielczarek Dominik is an Assistant Professor at the Department of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland. He received his PhD degree in mathematics from the Faculty of Mathematics and Computer Science of Jagiellonian University, Poland. His main areas of interest include functional analysis, functional data analysis and economic applications of robust and nonparametric statistics.

Muneer Siraj is working as the Lecturer of Statistics at Shaheed Benazir Bhutto University Sheringal Dir Upper, Pakistan. His main areas of interest include survey sampling, randomized response, non-response, design-based estimation and ranked set sampling. He has published 3 research papers in internationally reputed journals. He has produced 3 M.Phil students in the field of survey sampling. He is an external reviewer of many journals in the field of survey sampling.

Mussini Mauro is an Assistant Professor of Economic Statistics at the Department of Economics, University of Verona. He received his PhD degree in Statistics from the University of Milan Bicocca in 2008. His research interests include inequality and poverty measurement, statistical models for energy and environmental data analysis, statistical methods for the integration of data from different sources.

Okrasa Włodzimierz is a Professor and a Head of the Research Methods and Evaluation Unit at the Institute of Sociology, Cardinal Stefan Wyszynski University (UKSW). He serves as an Advisor to the President of the Statistics Poland and an Editor in Chief of the Statistics in Transition new series. He was teaching and researching in Polish and American universities and was an ASA Senior Research Fellow at the US Bureau of Labor Statistics (1990-1991), a Program Director in the Social Science Research Council, N. Y. (1991-1993), and then worked for the World Bank in Washington, D. C. (1994-2000), analyzing poverty and implementing new household surveys in several 'countries in transition'. He next headed the Social Sciences Unit at the European Science Foundation (2000-2003, in Strasbourg). Elected member of the International Statistical Institute. Author of numerous publications, including books and articles in reputed international journals

Pal Surya Kant is a Research Scholar in School of Statistics, Vikram University, India. His area of research are sampling theory and estimation procedures. He has published more than 50 research papers in international/national journals. Mr. Pal is also serving the journals as a reviewer.

Raja Tariq A. is working as a Senior Associate Professor (Statistics) at Sher-e-Kashmir University of Agricultural Sciences and Technology-Kashmir (J & K), India. He has published one hundred three research papers in various peer-reviewed international and national journals and has two manuals and a book to his credit. Besides being in the advisory committee of more than two hundred PG/PhD Scholars, one student has been awarded PhD and two are at the verge of completion of their PhD under his guidance. He has chaired six technical sessions and presented twenty-eight research papers in national and international conferences. He has completed two research projects and has organized three training programmes at national level. He has received three awards and is an executive member of various national and international societies.

Rozkrut Dominik is President of Statistics Poland. Graduated from School of Economics and Management at University of Szczecin (Poland). PhD in economics, statistics, econometrics time series analysis since 2003. Researcher at University of Szczecin. In 2008–2016 held a position of Director of Regional Statistical Office in Szczecin where conducted surveys on science, technology and innovations, information society as well as transport and communication. Considerable academic achievements and approximately one hundred scientific publications. Author/co-author of numerous scientific elaborations. Member of several intergovernmental working groups. He manages statistical cooperation with international organizations in Europe and world-wide and participates in the high-level meetings of United Nations, OECD and the European Union statistical systems.

Rydlewski Jerzy P. is a Faculty Member at the Department of Applied Mathematics, AGH University of Science and Technology, Kraków, Poland. He has received his PhD degree in mathematical statistics from the Faculty of Mathematics and Computer Science of Jagiellonian University, Poland. His main areas of interest include multivariate statistical analysis, functional data analysis, and theory and economic applications of robust and nonparametric statistics.

Särndal Carl-Erik is a Professor Emeritus, Statistics Sweden. He has pursued an academic career in Canada and is an Honorary Member of the Statistical Society of Canada. He is the author or co-author of widely cited books and scientific articles. Among his distinctions are the Waksberg Award for contributions to theory and practice of Survey Methodology. His broad research interest in survey statistics is currently focused on issues of data collection and estimation under survey nonresponse.

Shabbir Javid is working as the Tenured Professor of Statistics at Quaid-i-Azam University of Islamabad, Pakistan. His main areas of interest include: randomized response, non-response, multistage and multiphase sampling. He has published more than 150 research papers in nationally and internationally reputed journals. He has produced 7 PhD students in the field of survey sampling. He is an associate editor of the Journal of Statistical Theory and Practice.

Sharma Prayas is an Assistant Professor at the Department of Decision Sciences, School of Business, University of Petroleum and Energy Studies, Derhadun. His research interests are sampling theory, estimation, statistical modelling and data analysis in particular. Dr. Sharma has published more than 30 research papers in international/national journals along with two chapters and one book. Dr. Sharma is a member of more than 10 editorial boards including Investigación Operacional (Cuba), Journal of Reliability and Statistical Studies (India), Cambridge Scholars Publishing (United Kingdom), etc. He is also serving as a reviewer for more than 20 reputed journals including Journal of Statistical Theory and Practice, Hacettepe Journal of Mathematics and Statistics, Clinical Epidemiology and Global Health, Statistics and Transition new Series, International Journal of Applied and Computational Mathematics etc. Dr. Sharma is also engaged in consulting private organizations and corporates.

Stępniak Czesław graduated in 1977 (Polish Academy of Sciences); habilitation in 1988 (Adam Mickiewicz University, Poznań); professor title in 1990. Starting his academic service at Agricultural University in Lublin (1972-2001), he accepted position of professor in mathematics at the University of Rzeszów in 1996 and headed the Department of Statistics and Econometrics of Maria Curie-Skłodowska University (2003-2009). During the academic year 1987/88 he was visiting Mathematical Sciences Institute of Cornell University. Recipient of two grants from Committee for Scientific Research and two awards of the Ministry of Higher Education. His research interests include two areas: mathematics and statistics, especially linear algebra and statistical experiments. Czesław Stępniak has authored over 80 publications; 36 of them in leading mathematical and statistical journals indexed by JRC (16 journals). He has supervised two Ph.D. students Marek Niezgoda and Zdzisław Otachel. Currently he is a Professor Emeritus of University of Rzeszów, Poland.

Subzar Mir is a research scholar in the Division of Agricultural Statistics, Sher Kashmir University of Agricultural Science and Technology - Kashmir, Shalimar, India. His area of research is sampling theory and statistical inference. He has published 18 research articles in reputed national and international journals of, Agricultural Statistical Sciences, statistical and mathematical sciences. Mir has also presented some reputed articles in national and international conferences.

Traat Imbi is an Associate Professor at the Institute of Mathematics and Statistics, University of Tartu. She is an Honorary Doctor of the University of Umeå. She is a coordinator of the Baltic-Nordic-Ukrainian Network on Survey Statistics, and a member of several statistical bodies. She has supervised a number of PhD theses and has published many scientific papers. Her research interests are multivariate distributions, theoretical aspects of sampling design, survey statistics, and estimation under nonresponse.

Yozgatligil Ceylan Talu is an Associate Professor at the Department of Statistics, Faculty of Science and Arts Middle East Technical University. Her main research interests include univariate and multivariate time series analysis, temporal aggregation of time series, statistical applications in meteorological variables, spatio-temporal analysis, temporal clustering methods, statistical inference, survival analysis and competing risks. She has published around 70 research papers in international/national journals and conferences.

GUIDELINES FOR AUTHORS

We will consider only original work for publication in the Journal, i.e. a submitted paper must not have been published before or be under consideration for publication elsewhere. Authors should consistently follow all specifications below when preparing their manuscripts.

Manuscript preparation and formatting

The Authors are asked to use *A Simple Manuscript Template (Word or LaTeX) for the Statistics in Transition Journal (published on our web page:* http://stat.gov.pl/en/sit-en/editorial-sit/).

- ***Title and Author(s)***. The title should appear at the beginning of the paper, followed by each author's name, institutional affiliation and email address. Centre the title in **BOLD CAPITALS**. Centre the author(s)'s name(s). The authors' affiliation(s) and email address(es) should be given in a footnote.
- ***Abstract.*** After the authors' details, leave a blank line and centre the word **Abstract** (in bold), leave a blank line and include an abstract (i.e. a summary of the paper) of no more than 1,600 characters (including spaces). It is advisable to make the abstract informative, accurate, non-evaluative, and coherent, as most researchers read the abstract either in their search for the main result or as a basis for deciding whether or not to read the paper itself. The abstract should be self-contained, i.e. bibliographic citations and mathematical expressions should be avoided.
- ***Key words***. After the abstract, *Key words* (in bold 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).