STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Sampling methods and estimation

Research articles

Conference report

Book review

Volume 15, Number 3, Summer 2014

Editorial Office

ISSN 1234-7655

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

Address for correspondence

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

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 333–338

FROM THE EDITOR

This issue of the journal contains seven articles, four of which address matters classified to the first section, *Sampling methods and estimation,* while the other three fall under the second category, *Research articles*. These sections are complemented by a r elatively detailed report of the Small Area Estimation conference (SAE 2014) that took place in Poznan at the beginning of September. There is also the *Book review* section treating of a looked-for type of book by Vijay Verma, devoted to sampling of elusive populations in the context of child labour studies.

As a kind of announcement, there are few things worth making a note of here. The first relates to the aforementioned SAE 2014 conference. Namely, due to a big interest in the conference papers and as a response to the suggestions of several prominent participants of the conference, there is a joint special issue of the *Statistics in Transition new series* and the *Survey Methodology Journal* of Statistics Canada (edited by Mike Hidiroglou) under preparation. We have invited Raymond Chambers, Malay Ghosh, Graham Kalton, and Risto Lehtonen to provide leadership for this innovative venture as Guest Editors. We hope to contribute in this way to dissemination of the highest quality output of this important scientific event to the interested audience around the world.

The other thing concerns a technical improvement in accessing the journal before its final version is being made available − in both printed and online forms − due to introducing an *Early View* option, facilitating access and interaction between the editorial office and the authors, and allowing for making some minor changes or corrections in the meantime. As a way of bringing authors closer to readers, we have decided to include − starting with this issue - brief notes on authors/biosketches with some basic information about them.

From organizational point of view, it also deserves to be mentioned that the new Editorial Board held its first meeting on the occasion of the SAE2014 conference in Poznan - the minutes of the meeting concludes this note.

*

The first series of papers is opened by **Joseph W. Sakshaug's** and **Trivellore E. Raghunathan’s** paper ***Generating Synthetic Microdata to Estimate Small Area Statistics in the American Community Survey***. The authors propose a solution to practically important issue caused by certain constraints being imposed on information - as regards its scope and form - needed in the context of local-level studies. Although, on t he one hand, statistical agencies regularly collect data from small areas, they are prevented from releasing detailed geographical identifiers in public-use data sets due to disclosure concerns. On the other hand, data dissemination methods used in practice include releasing summary (aggregate) tables, suppressing detailed geographic information. Therefore, an alternative method for disseminating microdata with more geographical details than are currently being released in public-use data files is presented by the authors. Specifically, the method replaces the observed survey values with imputed or synthetic values simulated from a h ierarchical Bayesian model. Confidentiality protection is enhanced because no actual values are released. The method is demonstrated using restricted data from the 2005-2009 American Community Survey. The analytic validity of the synthetic data is assessed by comparing small area estimates obtained from the synthetic data with those obtained from the observed data.

The paper by **Kajal Dihidar** is devoted to ***Estimating Population Mean with Missing Data in Unequal Probability Sampling.*** It discusses the nonresponse problem as a ser ious obstacle to the validity of estimates. The question how to deal with missing values is complicated by the fact that they are deemed impossible to recover. One way of exploring a possible lack of representativity in missing data is to estimate the response probabilities which are usually done by logistic regression. However, the drawback of this model is that it requires knowledge of the explanatory variables for all nonrespondents. One way is to estimate response probabilities by weighting adjustment technique without having the individual data of the nonrespondents. The author considers the doubtful nature of nonresponse regarding possible existence of relationship with any of the covariates, and general unequal probability sampling scheme for selecting respondents. This paper presents the modification of Bethlehem (2012) proposal for unequal probability sampling to obtain the unbiased estimators for population total/average of a variable of interest and variance estimator, and compares them with the usual estimators through numerical simulations.

In the paper ***A Class of Two Phase Sampling Estimators for Ratio of Two Population Means Using Multi-Auxiliary Characters in the Presence of Non-Response*** by **B. B. Khare** and **R. R. Sinha** both the asymptotic bias and mean square error, as w ell as minimum mean square error of the proposed class of estimators have been obtained. The optimum values of the sample at the first and the second phases along with the sub-sampling fraction of the non-responding group have been determined for the fixed cost and for the specified precision. The efficiency of the proposed class of estimators has also been shown through the theoretical and empirical studies.

Sanjay Kumar Singh, **Umesh Singh** and **Manoj Kumar** in the paper ***Bayesian Inference for Exponentiated Pareto Model with Application to Bladder Cancer Remission Time*** discuss maximum likelihood and Bayes estimators of the unknown parameters. The expected experiment times of the exponentiated Pareto model have been obtained for progressive type-II censored data with binomial removal scheme. Markov Chain Monte Carlo (MCMC) method was used to compute the Bayes estimates of the parameters of interest. The generalized entropy loss function and squared error loss function have been considered for obtaining the Bayes estimators. Comparisons are made between Bayesian and maximum likelihood (ML) estimators via Monte Carlo simulation. The proposed methodology is illustrated for real data.

The *Research articles* section begins with **Anna Czapkiewicz's** and **Beata Basiura's** paper ***The Position of the WIG Index in Comparison with Selected Market Indices in Boom and Bust Periods.*** Its main objective is exploration of differences between the rank of the Polish stock market in the boom and the bust cycles. The daily stock market returns data for the twenty three major international indices from Europe, America and Asia are used for comparing two boom and two bust periods. The correlation coefficient obtained from Copula-GARCH model is a measure of similarity between the considered indices. The cluster analysis carried on for these series (in the boom and bust the cycles) allows us to identify the differences in the market behaviour. The empirical results indicate that the relationship of the Polish index with other indices is stronger during the bust sub-periods than during the boom ones. Through cluster analysis it is shown that the Polish index occurs in one subset with the Hungarian, Czech Republic, Turkish and Russian indices, regardless of the studied sub-periods.

In the paper by **Atanu Bhattacharjee** and **Dilip C. Nath*****, Joint Longitudinal and Survival Data Modelling: An Application in Anti-Diabetes Drug Therapeutic Effect,*** the longitudinal and survival analyses are shown to be useful tools in the exploration of drug trial data. In both cases the challenge is to deal with correlated repeated observations. Here, the joint modelling for longitudinal and survival data has been carried out via Markov Chain Monte Carlo (MCMC) method in type 2 d iabetes clinical trials to compare different combinations of drugs, viz. Metformin plus Pioglitazone and gliclazide plus pioglitazone. It has been found relatively easier to implement this model with Winbugs software, and the results were computed and compared with software R. In both types of the analyses it has been found that no estimates of treatment appear to have significant effect on the evolution of the matter of HBA1c, neither on the longitudinal part nor on the survival one. The Bayesian approach has also been considered as an extended tool with classical approach for estimation of clinical trial data analysis.

Henryk Gurgul's and **Pawel Zajac's** paper ***The Impact of Alterations in the Local Insolvency Legislation on Business Bankruptcy Rates in Poland*** analyses the effect of the major bankruptcy code novelization (that was enacted in the second quarter of 2009) on the number of insolvencies in Poland, using 'before-after' comparison. To this aim, a series of econometric models has been employed to analyze changes in bankruptcy rates using quarterly data for the period 2003-2013. Contrary to the expectations of lawmakers, while controlling for the variety of macroeconomic factors affecting insolvency rates, the authors conclude that the aggregate bankruptcy rates significantly increased after implementation of the new code (novelization of 2009). One of the reason is that entrepreneurs often do not use bankruptcy as a rational business formula due to its negative connotation in the colloquial language, and as a result they often start respective proceedings when it is too late to save their businesses. However, authors admit that this conclusion is pending for more detailed future assessment of the impact taking into account the effect of differences in firms' size and business sector on their failure rates.

*

Minutes of the Meeting of the SiTns' Editorial Board

Taking an exceptional opportunity provided by the fact that the overwhelming majority of members of the journal's new Editorial Board members attended the Small Are Estimation (SAE2014) conference in Poznan, an occasional meeting was organized by the SiTns Editor on the eve of the conference (i.e., on the 2nd of September). The following EB Members participated in the meeting: **Czesław Domanski, Malay Ghosh, Graham Kalton, Jan Kordos** *(Founder Editor)***, Janusz Witkowski, Janusz L. Wywial,** and the **Editor**.

At its outset, the Editor addressed some recent challenges and visions for possible improvements − such as moving to four issues per year, introducing an online *Early View* issue, increasing visibility of the journal, and multi-path efforts being under way for including the journal in the monitoring systems of the prestigious indexation bases toward obtaining the appropriate *impact factor(s)*. While sharing the journal's editorial policy and tasks being currently realized (by the Editorial Office), the EB Members provided several insightful observations and useful recommendations, which will underlay our efforts aimed at excelling the journal for making it increasingly attractive and needed for our key partners − potential authors, peer-reviewers and readers − and for the community of statisticians world-wide. The following suggestions are worthwhile mentioning here as accepted by all the EB Members:

- Janusz Witkowski stressed need to increase the visibility and accessibility of the journal and, as the President of the Central Statistical Office − the main sponsor of the journal − supported initiatives toward its greater international scope and rank in the global professional environment. Special issues, such as planned collection of papers based on the SAE 2014 conference presentations, would be a good means to achieve such goal.
- Graham Kalton indicated the need to have more research-based publications − meant as g iving greater preferences to articles presenting innovative applications of statistical methods in empirical research, and/or discussing statistical tools for such purposes. During supporting discussion further arguments were provided for bringing statistics closer to policy application (*policy research* articles, devoted, for instance, to policy and program evaluation), and the problem-solving implications were emphasized too. He also pointed to organizing special issues, as a strategy effective also in this context, illustrating this approach by one being currently under preparation (see *Call for papers* published in the last issue of the SiTns − special issue devoted to “The Measurement of Subjective Well-Being in Survey Research”).
- Malay Ghosh, who seconded this line of editorial policy, suggested to introduce a section *Review paper* on a systematic basis, as a p art of at least every other issue of the journal. It should be devoted to comprehensive discussion of the current state of selected areas of statistical research, with emphasis on new and important topics.

  In addition to such a review, it was also suggested that the *Book review* section should be a part of each issue too, to either complement the former or be used as its substitute.
- Czeslaw Domanski, the President of the Polish Statistical Association - under the aegis of which the journal is being issued − emphasized the unique role played by the journal as a platform for integration of the high level professionals across disciplines, world-wide.
- Janusz Wywial, commenting on the thematic profile of the journal, indicated the need of flexibility in this aspect, including papers on rarely presented matters, such as related for instance, to certain audit and finance statistics.

- Jan Kordos, supporting the suggestions, reflected a bit on the historical development of the journal, with optimist conviction of its further development in terms of quality and usability.

On behalf of the Journal and its Editorial Office, the Editor expressed the commitment to make all these observations, suggestions and recommendations the important input to efforts aimed at excelling the journal in all the aspects of its functioning as an organ serving professionals, statisticians and other readers from over the world.

Wlodzimierz Okrasa
Editor

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 339

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 m arket-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 sci entific domain (new research and data analysis methods) or as a domain of informational infrastructure of the economy, society and the state – are appropriate for *Statistics in Transition new series.*

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

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

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

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

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

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 341–368

GENERATING SYNTHETIC MICRODATA TO ESTIMATE SMALL AREA STATISTICS IN THE AMERICAN COMMUNITY SURVEY

Joseph W. Sakshaug[1], **Trivellore E. Raghunathan**[2]

ABSTRACT

Small area estimates provide a critical source of information used to study local populations. Statistical agencies regularly collect data from small areas but are prevented from releasing detailed geographical identifiers in public-use data sets due to disclosure concerns. Alternative data dissemination methods used in practice include releasing summary/aggregate tables, suppressing detailed geographic information in public-use data sets, and accessing restricted data via Research Data Centers. This research examines an alternative method for disseminating microdata that contains more geographical details than are currently being released in public-use data files. Specifically, the method replaces the observed survey values with imputed, or synthetic, values simulated from a hierarchical Bayesian model. Confidentiality protection is enhanced because no actual values are released. The method is demonstrated using restricted data from the 2005-2009 American Community Survey. The analytic validity of the synthetic data is assessed by comparing small area estimates obtained from the synthetic data with those obtained from the observed data.

Key words: counties, microdata, multiple imputation, data confidentiality.

1. Introduction

Demand for small area estimates is growing rapidly among a variety of stakeholders who use these data to advance the study of issues affecting local communities and the lives of their residents (Tranmer et al., 2005). Statistical agencies regularly collect data from small geographic areas and are therefore in a unique position to meet some of this demand. However, they are often prevented from releasing microdata for such areas because releasing detailed geographical identifiers for small areas may increase the risk of respondent re-identification

[1] Department of Statistical Methods, Institute for Employment Research, Germany. Program in Survey Methodology, University of Michigan, USA. E-mail: joesaks@umich.edu.

[2] Department of Biostatistics, University of Michigan, USA. E-mail: teraghu@umich.edu.

and inadvertent disclosure of confidential information (Mackie and Bradburn, 2000).

In order to minimize the risk of disclosure, statistical agencies commonly adopt one or more of the following data dissemination methods: 1) release summary tables that contain aggregate data for specific geographic areas (e.g., counties, census tracts, block groups); 2) suppress geographical details in public-use microdata sets for areas that fail to meet a predefined population threshold (e.g., 100,000) and; 3) release the unmasked confidential data set to data users via a secure data enclave or Research Data Center (RDC). Although these approaches are useful in many situations, each has limitations that preclude its ability to meet the growing demand for small area data that is being fuelled by researchers, analysts, policy-makers, and community planners.

For example, summary tables are useful tools for describing basic profiles of housing- and/or person-level characteristics for a wide variety of geographical areas, but their utility is limited to addressing complex scientific hypotheses that require customizable analytic approaches that are not feasible using existing aggregate data products. Releasing public-use microdata mitigates this issue by enabling users to perform customized analyses that go beyond the capabilities of published summary tables, but the suppression of identifiers for the smallest geographic areas limits their use for studying small area phenomenon. Releasing restricted microdata via a Research Data Center overcomes the limitations of the previous two by permitting users access to the full unmasked microdata, including all small area identifiers. In order to access data within an RDC, one must submit a research proposal, apply for special sworn status, pay a data usage fee, and travel to the nearest RDC facility. Unfortunately, these requirements are too restrictive for many analysts.

1.1. Synthetic data for small geographic areas

This article investigates a fourth approach that may permit statistical agencies to release more detailed geographical information in public-use data sets without compromising on data confidentiality. The approach extends the idea, originally proposed by Rubin (1993), of replacing the observed data values with multiply-imputed, or synthetic, values. The general idea is to treat the unobserved portion of the population as missing data to be multiply imputed using a predictive model fitted using the observed data. A random sample of arbitrary size is then drawn from each synthetic population which comprises the public-use data sets. Valid inferences are obtained by analyzing each synthetic data set separately and combining the point estimates and standard errors using combining rules developed by Raghunathan, Reiter, and Rubin (2003).

The synthetic data literature focuses on preserving statistics about the entire sample, but preserving small area statistics is usually ignored. Statistics about small areas can be extremely valuable to data users, but detailed geospatial information is almost always suppressed in public-use survey data. Research on model-based small area estimation has led to a greater understanding of how small area data can be summarized by statistical models (Platek et al., 1987; Rao, 2003), and such models could potentially be used for simulating small area microdata.

1.2. Fully synthetic versus partially synthetic data

There are two general synthetic data approaches: full synthesis and partial synthesis. Under a fully synthetic design all survey variables are synthesized and no real data is released. This approach provides the highest level of privacy and confidentiality protection (Drechsler, Bender, and Raessler, 2008), but the analytic validity of inferences drawn from the synthetic data may be poor if important relationships are omitted or mis-specified in the imputation model. Partial synthesis involves synthesizing a subset of variables or records that are pre-identified as being the most vulnerable to disclosure (Little, 1993; Kennickell, 1997; Liu and Little, 2002; Reiter, 2003). If implemented properly, this approach yields high analytic validity as inferences are less sensitive to misspecification of the imputation model. However, because the observed sample units and the majority of their data values are released to the public, it does not provide the same level of disclosure protection as full synthesis (Drechsler et al., 2008).

At the present time, the creation of partially synthetic data files is the most common application of synthetic data in large databases (Abowd, Stinson, and Benedetto, 2006; Rodriguez, 2007; Kinney et al., 2011). There are worthwhile reasons why fully synthetic data may be more appropriate for small area applications. Perhaps, the most important reason is that complete synthesis can offer stronger levels of disclosure protection than partial synthesis. Data disseminators are obligated by law to prevent data disclosures and may face serious penalties if they fail to do so. Maintaining high levels privacy protection should take precedence over maintaining high levels of analytic validity. This point is particularly important for small geographic areas, which may contain sparse subpopulations and higher proportions of unique cases that are especially susceptible to re-identification. A secondary benefit of fully synthetic data is that arbitrarily large sample sizes may be drawn from the synthetic populations, facilitating analysis for data users who would otherwise be forced to exclude areas with insufficient sample sizes, or apply complex indirect estimation procedures to compensate for the lack of sampled cases.

1.3. Organization of article

This article investigates an extension to Rubin’s synthetic data method for the purpose of creating fully synthetic, public-use microdata sets for small geographic areas. A hierarchical Bayesian model is used that accounts for multiple levels of geography and “borrows strength” across related areas. A sequential multivariate regression procedure is used to approximate the joint distribution of the observed data, which is then used to simulate synthetic values from the posterior predictive distribution (Raghunathan et al., 2001). How statistical agencies may generate fully synthetic data for small geographic areas is demonstrated using a subset of restricted data from the American Community Survey. Synthetic data is generated for several commonly used household- and person-level variables and their analytic validity is assessed by comparing inferences obtained from the synthetic data with those obtained from the actual data. The disclosure risk properties of the synthetic data methodology are not assessed here and are left to future work. Limitations of the app roach and possible extensions are discussed in the final section.

2. Review of fully synthetic data

2.1. Creation of fully synthetic data sets

The general framework for creating and analyzing fully synthetic data sets is described in Raghunathan et al. (2003) and Reiter (2005). Suppose a sample of size n is drawn from a finite population $\Omega = (X, Y)$ of size N, with $X = (X_i; i = 1,2, \dots, N)$ representing design, geographical, or other auxiliary information available for all N units in the population, and $Y = (Y_i; i = 1,2, \dots, N)$ representing the survey variables of interest. It is assumed that there is no confidentiality concern over releasing information about X and synthesis of these auxiliary variables is not needed, but the method can be extended to synthesize these variables if necessary. Let $Y_{obs} = (Y_i; i = 1,2, \dots, n)$ be the observed portion of Y corresponding to sampled units and $Y_{nobs} = (Y_i; i = n + 1, n + 2, \dots, N)$ be the unobserved portion of Y corresponding to the nonsampled units. The observed data set is $D = (X, Y_{obs})$. For simplicity, assume there are no item missing data in the observed data, but methods exist for handling this situation (Reiter, 2004).

Fully synthetic data sets are constructed in two steps. First, M synthetic populations $P^{(l)} = \{(X, Y^{(l)}); l = 1,2, \dots, M\}$ are generated by taking independent draws from the Bayesian posterior predictive distribution of $f(Y_{nobs} | X, Y_{obs})$ conditional on the observed data D. Alternatively, one can generate synthetic values of Y for all N units to ensure that no observed values of Y are released. The number of synthetic populations M is determined based on the desired accuracy for synthetic data inferences and the risk of disclosing confidential information. A modest number of fully synthetic data sets (e.g., 5 or 10) are usually sufficient to ensure valid inferences (Raghunathan et al., 2003). In the second step, a random sample of size n_{syn} is drawn from each of the $l = 1,2, \dots, M$ synthetic data populations, $D^{(l)} = \left(x_i, y_i^{(l)}, i = 1,2, \dots, n_{syn}\right)$. The corresponding M synthetic samples $D_{syn} = \left(D^{(l)}; l = 1,2, \dots, M\right)$ comprise the public-use data sets, which are released to, and analyzed by, data users. In practice, the first step of generating complete synthetic populations is unnecessary and we only need to generate values of Y for units in the synthetic samples. The complete synthetic population setup is useful for theoretical development of combining rules.

2.2. Obtaining inferences from fully synthetic data sets

From the publicly-released synthetic data sets, data users can make inferences about a scalar population quantity $Q = Q(X, Y)$, such as the population mean of Y or the population regression coefficients of Y on X. Suppose the analyst is interested in obtaining a point estimate q and an associated measure of uncertainty v of Q from a set of synthetic samples D_{syn} drawn from the synthetic populations $P_{syn} = \left(P^{(l)}; l = 1,2, \dots, M\right)$ under simple random sampling. The values of q and v computed on the M synthetic data sets are denoted by $\left(q^{(l)}, v^{(l)}, l = 1,2, \dots, M\right)$.

Consistent with the theory of multiple imputation for item missing data (Rubin, 1987; Little and Rubin, 2002), combining inferences about $Q = Q(X, Y)$ from a set of synthetic samples D_{syn} is achieved by approximating the posterior distribution of Q conditional on D_{syn}. The suggested approach, outlined by Raghunathan et al. (2003), is to treat $\left(q^{(l)}, v^{(l)}; l = 1,2, \dots, M\right)$ as sufficient summaries of the synthetic data sets D_{syn} and approximate the posterior density $f\left(Q|D_{syn}\right)$ using a normal distribution with the posterior mean Q computed as the average of the estimates,

$$\bar{q}_M = \sum_{l=1}^{M} q^{(l)} / M \tag{1}$$

and the approximate posterior variance is computed as,

$$T_M = (1 + M^{-1}) b_M - v_m \tag{2}$$

where$\bar{v}_M = \sum_{l=1}^{M} v^{(l)} / M$ is the overall mean of the estimated variances across all synthetic data sets ("within variance") and $b_M = \sum_{l=1}^{M} \left(q^{(l)} - \bar{q}_M\right)^2 / (M - 1)$ is the variance of $q^{(l)}$ across all synthetic data sets ("between variance").

Under certain regulatory conditions specified in Raghunathan et al. (2003), $\bar{q}_M$ is an unbiased estimator of Q and $b_M - v_m$ is an unbiased estimator of the variance of Q. The $\frac{1}{M} b_M$ adjusts for using only a finite number of synthetic data sets. It should be noted that the subtraction of the within imputation variance in T_M is due to the additional step of sampling units from the synthetic populations. Because of this extra sampling step, the between imputation variance contains the true between and nearly twice the amount of within variance needed to obtain an unbiased estimate of T.

When n, n_{syn}, and M are large, inferences for scalar Q can be based on normal distributions. For moderate M, inferences can be based on t-distributions with degrees of freedom $\gamma_M = (M-1)(1-r_m^{-1})^2$, where $r_m = (1+M^{-1})b_m/\bar{v}_M$, so that a $(1-\alpha)\%$ interval for Q is $\bar{q}_M \pm t_{\gamma_M}(\alpha/2)\sqrt{T_M}$ as described in Raghunathan and Rubin (2000). Extensions for multivariate Q are described in Reiter and Raghunathan (2007).

A limitation of the variance estimator T_M is that it can produce negative variance estimates. Negative values of T_M can generally be avoided by increasing Mor n_{syn}. Numerical routines can be used to calculate the integrals involved in the construction of T_M, yielding more precise variance estimates (Raghunathan et al., 2003). A simpler variance approximation that is always positive is shown in Reiter (2002).

3. Creation of synthetic data sets for small geographic areas

Hierarchical models have been used in several applications of small area estimation (Fay and Herriot, 1979; Malec et al., 1997). See Rao (2003) for a comprehensive review of design-based, empirical Bayes, and fully Bayesian approaches for small area estimation. Hierarchical models have also been used for multiple imputation of missing data in multilevel data structures (Reiter, Raghunathan, and Kinney, 2006; Yucel, 2008).

The approach considered here involves three stages. In the first stage, the joint density of the variables to be synthesized is approximated by fitting sequential regression models based on the observed data within each small area. In the second stage, the sampling distribution of the unknown regression parameters (estimated in the first stage) is approximated and the between-area variation is modelled using auxiliary information. In the third stage, the unknown regression parameters are simulated and used to draw synthetic microdata values from the posterior predictive distribution.

Two levels of geography are considered. For illustration, consider "small areas" as counties nested within states. In illustrating the approach, the models are kept relatively simple from a computational perspective to make the modelling practical. Despite the simplified presentation, the framework can be extended to handle more sophisticated modelling approaches.

3.1. Stage 1: Approximation of joint density via sequential regression

Suppose that a simple random sample of size n is drawn from a finite population of size N. Assuming units were sampled from each county, let n_{cs} and N_{cs} denote the respective sample and population sizes for county $c = (1,2,\ldots,C_s)$ nested within state $s = (1,2,\ldots,S)$. Let $Y_{cs} = (Y_{ics,p}; i = 1,2,\ldots,n_{cs};\ p = 1,2,\ldots,P)$ represent the $n_{cs} \times P$ matrix of survey variables collected from each survey respondent located in county c and state s. Let $X_{cs} = (X_{ics,j}; i = 1,2,\ldots,n_{cs},n_{cs}+1,\ldots,N_{cs};\ j = 1,2,..,J)$ represent the $N_{cs} \times J$ matrix of auxiliary or administrative variables known for every population member in a particular county and state. Here only the survey variables $Y_{cs,p}$ are synthesized, but it is straightforward to synthesize the auxiliary variables $X_{cs,j}$ as well.

A desirable property of the synthetic data is that the multivariate relationships among the observed variables are maintained in the synthetic data, i.e., the joint distribution of variables given the auxiliary information $f(Y_{cs,1}, Y_{cs,2}, \dots, Y_{cs,P} | X_{cs,j})$ is preserved. Specifying and simulating from the joint conditional distribution can be difficult for complex data structures involving large numbers of variables representing a v ariety of distributional forms. Alternatively, one can approximate the joint density as a product of conditional densities (Raghunathan et al., 2001). That is, the joint density $f(Y_{cs,1}, Y_{cs,2}, \dots, Y_{cs,P} | X_{cs,j})$ can be factored into the following conditional densities: $f(Y_{cs,1} | X_{cs,j})$, $f(Y_{cs,2} | Y_{cs,1}, X_{cs,j})$,…,$f(Y_{cs,P} | Y_{cs,1}, \dots, Y_{cs,P-1}, X_{cs,j})$. In practice, a seq uence of generalized linear models are fit based on the observed county-level data where the variable to be synthesized comprises the outcome variable that is regressed on any auxiliary variables or previously fitted variables, e.g., $Y_{ics,1} = (X_{ics})\beta_{cs,1} + \varepsilon_{ics}$, $Y_{ics,2} = (X_{ics}, Y_{ics,1})\beta_{cs,2} + \varepsilon_{ics}$,…,$Y_{ics,P} = (X_{ics}, Y_{ics,1}, Y_{ics,2}, \dots, Y_{ics,P-1})\beta_{cs,P} + \varepsilon_{ics}$. The choice of model (e.g., Gaussian, binomial) is dependent on the type of variable to be synthesized (e.g., continuous, binary). It is assumed that any complex survey design features are incorporated into the generalized linear models and that each variable has been appropriately transformed to satisfy modelling assumptions. After fitting each conditional density, the vector of regression parameter estimates $\hat{\beta}_{cs,p}$, the corresponding covariance matrix $\hat{V}_{cs,p}$, and the residual variance $\hat{\sigma}^2_{cs,p}$ are extracted from each of the P regression models and incorporated into the hierarchical model described below. $p = (1,2, \dots, P)$is used to index the set of parameters associated with the p^{th} synthetic variable of interest and the p^{th} regression model from which the direct estimates are obtained.

3.2. Stage 2: Sampling distribution and between-area model

In the second stage, the joint sampling distribution of the design-based county-level regression estimates $\hat{\beta}_{cs,p}$ (obtained from each conditional model fitted in Stage 1) is approximated by a multivariate normal distribution,

$$\hat{\beta}_{cs,p} \sim MVN(\beta_{cs,p}, \hat{V}_{cs,p}) \tag{3}$$

where$\beta_{cs,p}$ is the $(J+p) \times 1$ matrix of unknown regression parameters and $\hat{V}_{cs,p}$ is the corresponding $(J+p) \times (J+p)$ estimated covariance matrix obtained from Stage 1. The unknown county-level regression parameters $\beta_{cs,p}$ are assumed to follow a multivariate normal distribution,

$$\beta_{cs,p} \sim MVN(\beta_p Z_s, \Sigma_p) \tag{4}$$

where$Z_s = \left(Z_{s,k}; k = 1,2, \dots, K\right)$ is a $K \times 1$ matrix of state-level covariates, β_p is a $(J + p) \times K$ matrix of unknown regression parameters, and Σ_p is a $(J + p) \times (J + p)$ covariance matrix. State-level covariates are incorporated into the hierarchical model in order to "borrow strength" from related areas. Prior distributions may be assigned to the unknown parameters $\beta_p$and Σ_p, but for computational simplicity it is assumed that β_p and Σ_p are fixed at their respective maximum likelihood estimates, a common assumption in hierarchical models for small area estimation (Fay and Herriot, 1979; Datta, Fay, and Ghosh, 1991; Rao, 1999). Details for obtaining the maximum likelihood estimates using the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) are provided in Appendix 1.

Based on standard theory of the normal hierarchical model (Lindley and Smith, 1972), the unknown regression parameters $\beta_{cs,p}$ can be drawn from the following posterior distribution,

$$\tilde{\beta}_{cs,p} \sim MVN\left[\left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_p^{-1}\right)^{-1}\left(\hat{V}_{cs,p}^{-1}\hat{\beta}_{cs,p} + \hat{\Sigma}_p^{-1}\hat{\beta}_p Z_s\right), \left(\hat{V}_{cs,p}^{-1} + \hat{\Sigma}_p^{-1}\right)^{-1}\right] \tag{5}$$

where$\tilde{\beta}_{cs,p}$ is a simulated vector of values for the unknown regression parameters $\beta_{cs,p}$.

3.3. Stage 3: Simulating from the posterior predictive distribution

The ultimate objective is to generate synthetic populations for each small area using an appropriate posterior predictive distribution. Simulating a synthetic variable $\tilde{Y}_{cs} = \left(\tilde{Y}_{lcs,p}; l = 1,2, \dots, N_{cs}; p = 1,2, \dots, P\right)$ for observed variable Y_{cs} for synthetic population unit $l = (1,2, \dots, N_{cs})$ is achieved by drawing, in sequential fashion, from the following posterior predictive distributions $f\left(\tilde{Y}_{cs,1}|X_{cs}, \tilde{\beta}_{cs,1}\right)$, $f\left(\tilde{Y}_{cs,2}|\tilde{Y}_{cs,1}, X_{cs}, \tilde{\beta}_{cs,1}\right)$, …, $f\left(\tilde{Y}_{cs,P}|\tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,P-1}, X_{cs}, \tilde{\beta}_{cs,1}\right)$. For example, if the first variable to be synthesized $Y_{cs,1}$ is normally distributed then $\tilde{Y}_{cs,1}$ can be drawn from a normal distribution with location and scale parameters $X_{cs}\tilde{\beta}_{cs,1}$and $\sigma_{cs,1}^2$, respectively, where $\sigma_{cs,1}^2$ may be drawn from an appropriate posterior predictive distribution, or fixed at its maximum likelihood estimate $\hat{\sigma}_{cs,1}^2$ (obtainable from Stage 1). Generating a second (normally distributed) synthetic variable $\tilde{Y}_{cs,2}$ from the posterior predictive distribution $f\left(\tilde{Y}_{cs,2}|\tilde{Y}_{cs,1}, X_{cs}, \tilde{\beta}_{cs,2}\right)$ is achieved by drawing $\tilde{Y}_{cs,2}$ from $N\left[\left(X_{cs}, \tilde{Y}_{cs,1}\right)\tilde{\beta}_{cs,2}, \sigma_{cs,2}^2\right]$, and so on up to $\tilde{Y}_{cs,P} \sim N\left[\left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,P-1}\right)\tilde{\beta}_{cs,P}, \sigma_{cs,P}^2\right]$. Alternatively, if the variable under synthesis $Y_{cs,p}$ is binary, then $\tilde{Y}_{cs,p}$ is drawn from a binomial distribution $Bin\left[1, \hat{p}\left\{\left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,P}\right\}\right]$, where $\hat{p}\left\{\left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,P}\right\}$ is the predicted probability computed from the inverse-logit of $\left\{\left(X_{cs}, \tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,p-1}\right)\tilde{\beta}_{cs,P}\right\}$. For polytomous variables, the same procedure is used to obtain posterior probabilities for each categorical response, which are then used to generate the synthetic values from a multinomial distribution. The iterative simulation process continues until all synthetic variables $(\tilde{Y}_{cs,1}, \tilde{Y}_{cs,2}, \dots, \tilde{Y}_{cs,P})$ are generated. The procedure is repeated M times to create multiple populations of synthetic variables $(\tilde{Y}^{(l)}_{cs,1}, \tilde{Y}^{(l)}_{cs,2}, \dots, \tilde{Y}^{(l)}_{cs,P}; l = 1,2, \dots, M)$. In addition, the entire cycle may be repeated several times to minimize ordering effects (Raghunathan et al., 2001).

The complete synthetic populations may be disseminated to data users, or simple random samples of arbitrary size may be drawn from each population and released. Stratified random sampling may be used if different sampling fractions are to be applied within small areas. Inferences for a variety of estimands can be obtained using the combining rules in Section 2.2.

4. American Community Survey (2005-2009)

The proposed methodology is applied to a subset of restricted county-level microdata from the 2005-2009 American Community Survey (ACS), obtained from the Michigan Census Research Data Center. The ACS is an ongoing national survey that provides yearly estimates on a variety of topics, including income and benefits, health insurance coverage, disabilities, family and relationships, and others. The ACS collects information on pe rsons living in housing units and group quarters facilities in all 3,141 counties in the United States. Data collection is conducted using a mixed-mode design. First, questionnaires are mailed to all sampled household addresses obtained from a Master Address File. Approximately six weeks after the questionnaire is mailed the Census Bureau attempts to conduct telephone interviews with all households that do not respond by mail. Following the telephone operation, a random sample is taken from the list of addresses where interviews have not been obtained and these addresses are visited by a field representative. Full details of the ACS methodology can be found in the technical documentation (U.S. Census Bureau, 2009).

Unlike the ACS public-use microdata files, the restricted data contain identifiers for all counties in the United States. For this application, we restrict the data to occupied housing units in the Northeast region. The Northeast region consists of 217 counties, all of which included households that completed ACS interviews. We use 5 years of restricted data to facilitate the disclosure review process and allow for the publication of estimates for all counties; the latter is not permitted with fewer years. Seven household- and seven person-level variables were selected for this analysis. The variables, shown in Table 1, were chosen by statisticians at the U.S. Census Bureau specifically for this project due to their common use among data users. Some variables (e.g., household tenure status, education, race) contained numerous categories. Ideally, each category would be preserved in the synthetic data; however, the decision was made to keep the number of categories at a minimum while maximizing the number of variables used in this small demonstration project. Thus, the few polytomous variables were recoded to reduce their number of categories. Transformations were applied to the continuous variables to meet normality assumptions during the model fitting and the synthetic data generation stages. After the synthesis was completed, the variables were transformed back to their original scales. The Census Bureau applies single imputation to missing ACS values in the restricted and public-use data files. We treat these imputations as actual observations in this application.

Table 1. List of ACS Variables Used in Synthetic Data Application. Variables Shown in the Order of Synthesis

Variable	Type	Range/Categories	Transformation
Household variables			
Household size	count	1 - 20	--
Sampling weight	continuous	1 - 201	log
Total bedrooms	count	0 - 5	--
Electricity bill/mo.	continuous	1 - 687	cube root
Total rooms (excl. bedrooms)	count	1 - 7	--
Income	continuous	0 – 3,999,996	cube root
Tenure	polytomous	recoded; mortgage/loan, own free and clear, rent	--
Person variables			
Sampling weight	continuous	1 - 341	log
Gender	binary	male, female	--
Education	polytomous	recoded; < 12 years, 12 years, 13-15 years, 16+ years	--
Hispanic ethnicity	binary	yes, no	--
Age	continuous	0 - 115	--
Race	polytomous	recoded; white, black, other	--
Living in poverty	binary	yes, no	--

Ten fully synthetic household- and person-level data sets were generated for each county. To ensure that each synthetic data set contained ample numbers of households and persons within each county, synthetic samples were created to be approximately equivalent to 20% of the total number of households based on the decennial census count. This yielded a total synthetic sample size of 3,963,715 households and 10,192,987 persons in the Northeast region.

The first survey variable to be synthesized was household size. Creating a household size variable facilitates the subsequent generation of synthetic person-level data. Household size was simulated using a Bayesian Poisson-Gamma model conditional on the observed household size variable with unknown hyperparameters fixed at their marginal maximum likelihood estimates obtained using the Newton-Raphson algorithm (see Appendix 2 for details). All subsequent variables were synthesized using the hierarchical modelling approach described in Section 3. S tate-level covariates $Z_S$that were incorporated into the hierarchical model included population size (2005 estimate: log-transformed) and the number of metropolitan and micropolitan areas. These covariates were obtained from the Census Bureau website.

For numerical variables (continuous, count), design-based estimates of regression parameters were obtained by fitting normal linear models within each county and synthetic values were drawn from the Gaussian posterior predictive distribution. For binary variables, logistic regression models were used to obtain the design-based parameter estimates and synthetic values were drawn from the binomial posterior predictive distribution. Logistic regression was also applied to polytomous variables after breaking them up i nto a series of conditional binary variables, estimating the propensity of a case belonging to a particular category versus all other categories, and using those propensities to predict case membership. We considered using multinomial regression for polytomous variables, but preliminary testing yielded convergence and stability problems for many counties. Therefore the decision was made to use the modified logistic regression approach. To increase the stability of the estimated regression coefficients, a minimum sample size rule of $10 \cdot p$ was applied within each county. If the target county did not meet this sample size threshold then nearby counties were pooled together until the criterion was met.

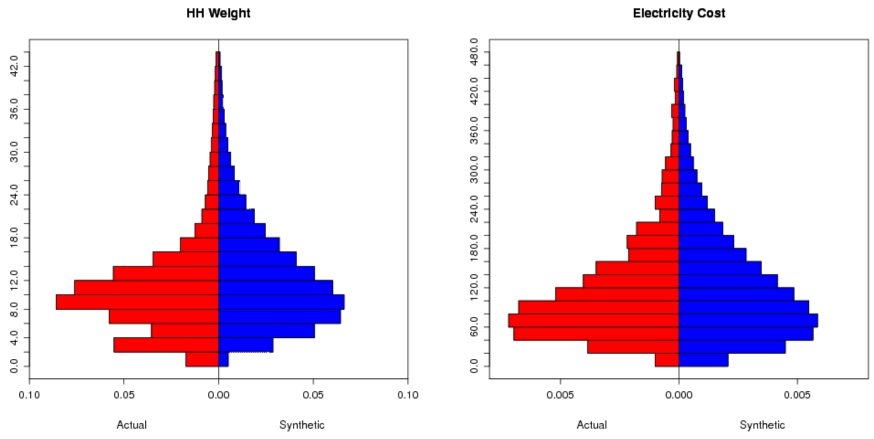
The household variables were synthesized first, followed by the person variables. After the synthetic household data sets had been created, they were converted to person-level data sets based on values of the synthetic household size variable. Taylor series linearization (Binder, 1993) was used to adjust the variances of the design-based regression estimates for the additional homogeneity due to persons clustered within households. To reduce the ordering effect induced by synthesizing the variables in a prescribed order, we repeat the entire synthetic data process 4 additional times, each time conditioning on the full set of synthetic variables generated from the previous implementations. Finally, it should be noted that the person-level variables were synthesized independently of the household-level variables. Although multiple imputation theory dictates that one should condition on all available information (Rubin, 1987), we found in preliminary runs that cycling between household- and person-level synthesis by aggregating person-level variables up to the household-level did not yield satisfactory inferences, possibly due to the non-standard distributions that the aggregation procedure produced. After applying several transformation procedures to the aggregated person-level variables, which did not significantly improve the imputations, we decided to keep the household and person levels separate for this demonstration project.

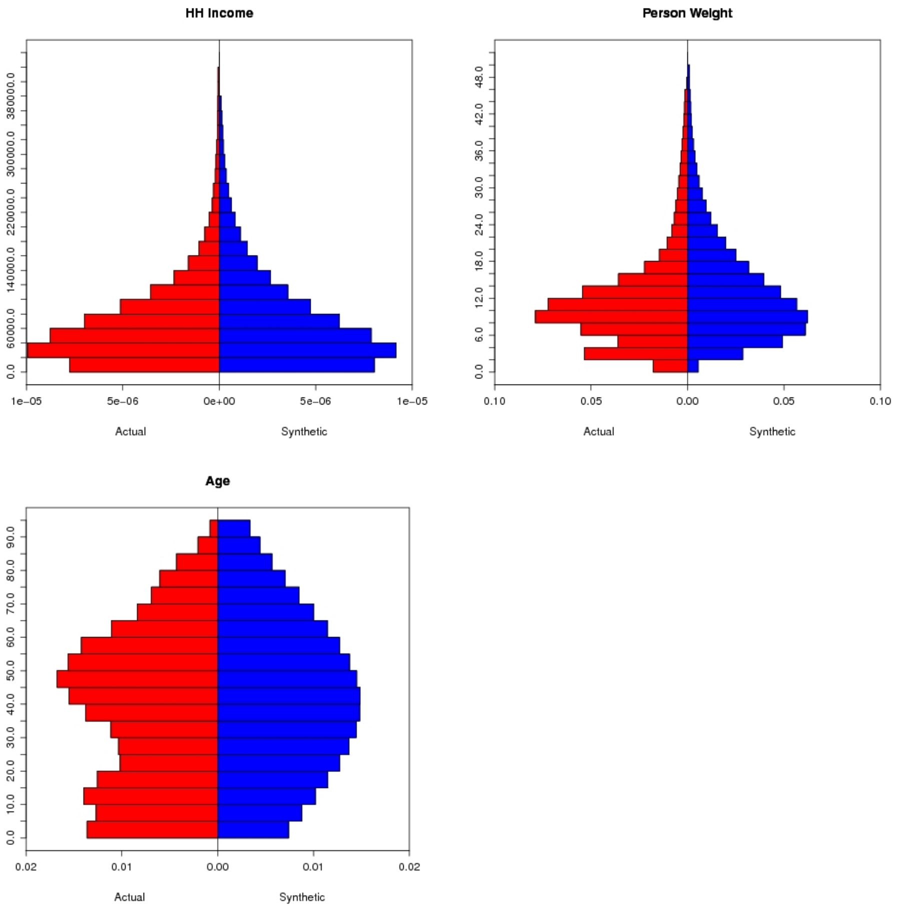
All results were reviewed and approved by the U.S. Census Bureau's Disclosure Review Board.

4.1. Validity of univariate estimates

Figure1 contains back-to-back histograms depicting the overall distributions for each continuous household- and person-level variable. The actual distributions are shown in the left panel and the synthetic distributions in the right panel. All variables are presented in their original scale. Visual comparisons show that for some variables, the synthetic data distribution corresponds to the actual data distribution reasonably well, but for others, the correspondence is poorer. Although the bulk of the distributions are generally maintained in the synthetic data, not every peak and valley is preserved. Those variables which do not follow a smooth parametric form tend to be most susceptible to a lack of correspondence. For example, the shape of the age distribution is bimodal denoting the highest frequency of people between the ages of 0-20 and 45-55. The synthetic age values, which are simulated from a normal distribution, fail to reflect the underlying bimodality. To a lesser degree, the sampling weight variables exhibit some bimodality at the left-most portion of their distributions, which is also not accounted for by the synthetic data. More sophisticated techniques, such as mixture modelling or nonparametric imputation may do a better job of preserving these non-standard distributional forms.

Figure 1. Back-to-Back Histograms of Actual (Left) and Synthetic (Right) Distributions for Continuous ACS Household- and Person-Level Variables in the Northeast Region.





While it is useful to compare synthetic and actual variable distributions for purposes of evaluation, data users are most interested in the validity of the small estimates obtained from the synthetic data. Table 2shows summary measures of univariate county-level estimands obtained from the synthetic and actual data. The first column contains the original set of ACS variables as w ell as recoded binary variables indicating overall income percentiles (50th, 75th and 90th) and specific subgroups (income x t enure; poverty x race/ethnicity). The second column shows the average county mean obtained from the synthetic and actual data, across all 217 counties. The third and fourth columns show the average standard deviation and standard error of the county means. The last column contains the intercept and slope values obtained from regressing the actual county means against the corresponding synthetic means. Intercept values close to zero and slope values close to one indicate strong correspondence between the synthetic and actual data estimates.

The synthetic data estimates, based on the original ACS variables, correspond roughly to the actual estimates, on a verage; out of the 9 hous ehold- and 12 person-level estimands, 5 and 10 of them yield synthetic point estimates that lie within two standard errors of the actual estimates, respectively, on average. The largest deviations occur for the tenure variable where the percentage of housing units being rented is overestimated by about two percentage points, on average, and the percentages of housing units owned free and clear and being financed through a mortgage or loan, are both underestimated in the synthetic data by about one and three percentage points, respectively, on average. These deviations are evident from examination of scatter plots of synthetic and actual county-level estimates (not shown, but available upon r equest). Similar over- and under-estimation effects appear in estimates of the other polytomous variables (education, race), but to a lesser extent. The cause of these effects is likely driven by two joint factors. The overestimation is likely due to the pooling of nearby counties to facilitate model fit for target counties that contained insufficient numbers of rented housing units; the rarest of the three membership categories. For the affected counties, the act of pooling at the estimation stage yields a higher rate of rented housing units in the synthetic data, which is closer to the population average. The underestimation in the other tenure estimates is driven by the fact that rental status was the first tenure category to be simulated, followed by ownership (conditional on not being rented) and mortgage/loan status (conditional on not being rented or owned). A consequence of this step-by-step conditional simulation approach is that the higher rates of rented housing units generated for the areas with inadequate samples sizes are offset by lower rates of ownership and mortgage/loan status for these smaller areas.

Aside from the positive/negative deviations among the polytomous estimates, the other estimates, based on continuous and binary ACS variables, appear to be reasonably valid as indicated by the diagnostic measures in Table 2. Many of the estimands yield intercept and slope values for the linear regression of actual county means against the synthetic means that are close to zero and one, respectively, indicating good correspondence between the actual and synthetic estimates. However, some of the continuous variables including electricity bill amount, household income, and, especially, age, yield larger deviations from the ideal intercept and slope values. The largest deviation occurs for the age estimates, which are likely due to the aforementioned bimodality of the age distribution that is reflected poorly in the synthetic data. The resulting synthetic county-level age estimates tend to be biased upward, particularly, for the counties with the highest average ages.

The validity of the percentile and subgroup estimates is mixed. The percentage of households with incomes exceeding the 50th percentile in the synthetic data corresponds closely to the actual percentages, on average. However, the estimates based on the 75th and 90th percentiles are higher in the synthetic data by about 1.5-2.0 percentage points, on average. Scatterplots of the county-level percentile means (not shown, but available upon request) indicate that the correspondence between synthetic and actual means becomes poorer as the percentile increases. Almost all of the income and poverty subgroup means lie within 1-2 standard errors of their corresponding actual means, on average. However, a positive and negative bias can be seen for synthetic estimates of mean income among mortgaged and rented housing units from scatterplots (not shown, but available upon request); a result that is likely due to the aforementioned under- and over-estimation of these tenure variables in the synthetic data, respectively.

A few remarks can be made about the uncertainty of the synthetic estimates. Based on multiple imputation theory, we would expect the synthetic standard deviations to be approximately the same and the standard errors to be larger than the actual standard deviations and standard errors, respectively, on average. This expectation is confirmed for some, but not all estimates. In most cases, the synthetic data standard deviations are close to their actual data counterparts. A particular exception is age, which yields larger standard deviations in the synthetic data, on average, due to the aforementioned bimodal age distribution, which is smoothed over in the synthetic data causing more age values to lie further away from the mean. On average, about half of the synthetic standard errors is equal to or greater than the corresponding actual standard errors. Estimates of income tend to have smaller standard errors in the synthetic data, on average, as a result of outlying observations being less preserved in the synthetic data. Moreover, the underestimated variances could be caused by misspecification of the imputation model and/or poor choice of transformation for preserving the tail-end of the distribution in the synthetic data, a problem which has been highlighted in earlier research on the estimation of imputed totals in skewed populations (Rubin, 1983). Another possible source of variation not accounted for in the synthetic data is due to the fact that the hyperparameters were fixed at their maximum likelihood estimates (see Section 3.2), rather than being randomly drawn from an *a priori* distribution.

Table 2. Summary Measures of Actual and Synthetic County Means

	Avg. Mean		**Avg. Standard Deviation**		**Avg. Standard Error of Mean**		**Regression of Actual Means on Synthetic Means**	
	Actual	Synthetic	Actual	Synthetic	Actual	Synthetic	Intercept	Slope
Household variables								
Household size	2.12	2.12	1.46	1.45	0.02	0.01	0.02	0.99
Sampling weight	9.99	10.20	7.21	7.04	0.11	0.11	0.01	0.98
Total bedrooms	2.88	2.82	0.96	1.09	0.02	0.01	0.15	0.97
Electricity bill/mo.	118.89	119.37	78.72	78.33	1.25	1.10	9.90	0.91
Total rooms	3.23	3.18	1.19	1.28	0.02	0.02	0.09	0.99
Income	67983.9	67382.4	68481.3	54081.9	1067.3	692.6	4681.7	0.94
Tenure (%)								
Mortgage/loan	49.00	47.03	49.38	49.30	0.82	0.74	0.04	0.95
Own free & clear	31.12	30.37	45.53	44.97	0.77	0.72	0.05	0.85
Rent	19.88	22.60	38.86	41.00	0.63	0.63	-0.05	1.09

Table 2. Summary Measures of Actual and Synthetic County Means (cont.)

	Avg. Mean		**Avg. Standard Deviation**		**Avg. Standard Error of Mean**		**Regression of Actual Means on Synthetic Means**	
	Actual	Synthetic	Actual	Synthetic	Actual	Synthetic	Intercept	Slope
Recoded variables								
Income > 50th pctile,%	44.65	44.56	48.24	48.19	0.80	0.56	0.01	0.97
Income > 75th pctile,%	19.34	21.49	37.34	38.69	0.59	0.43	-0.00	0.91
Income > 90th pctile,%	6.78	8.38	22.96	24.58	0.35	0.24	0.56	0.74
Income (Mortgage=1)	84667.0	86992.6	69019.2	58960.1	1536.0	1195.3	5460.0	0.91
Income (Own=1)	61076.6	60456.9	76053.1	45083.6	2132.8	1232.7	1717.0	0.98
Income (Rent=1)	38844.5	36921.9	37759.4	32527.3	1436.0	1166.5	3480.0	0.99
Person variables								
Sampling weight	10.27	10.67	7.59	8.02	0.08	0.14	-0.09	0.97
Gender (%)	48.63	48.63	49.97	49.97	0.53	0.44	0.04	0.91
Education (%)								
< 12 years	31.48	31.67	46.31	46.31	0.49	0.39	0.09	0.71
12 years	28.34	27.74	44.40	44.06	0.48	0.57	0.01	0.97
13-15 years	20.33	20.25	40.11	40.04	0.43	0.50	0.01	0.96
16+ years	19.85	20.35	38.72	39.14	0.40	0.51	-0.01	1.00
Hispanic (%)	3.85	4.23	15.72	16.99	0.14	0.26	-0.00	1.00
Age	40.89	41.16	22.98	30.34	0.25	0.27	22.02	0.46
Race (%)								
White	92.21	91.34	22.17	24.08	0.20	0.36	0.01	1.00
Black	3.55	4.01	14.54	16.26	0.13	0.26	-0.01	1.00
Other	4.24	4.65	14.54	18.61	0.16	0.27	-0.00	1.00
Poverty (%)	8.65	9.04	27.54	28.13	0.30	0.53	-0.00	1.00
Recoded variables								
Poverty (White=1; %)	7.93	8.19	26.41	26.84	0.30	0.51	-0.00	1.00
Poverty (Black=1; %)	20.48	21.30	36.86	37.03	4.62	3.52	-0.01	1.01
Poverty (Other=1; %)	16.62	17.84	35.37	36.07	2.96	4.38	0.01	0.87
Poverty (Hispanic=1; %)	19.92	21.11	37.08	37.96	3.52	5.54	-0.01	0.98

4.2. Validity of multivariate estimates

The next set of analyses examine the analytic validity of synthetic multivariate estimates obtained from multiple regression models. Table 3 shows average coefficient estimates (and their standard errors) for two regression models fit within each county. The first model fits a household-level linear regression of income (cube root) on the remaining ACS household covariates, and the second model fits a person-level logistic regression of poverty status on the remaining person covariates. Both models yield coefficient estimates based on the synthetic data that closely resemble those based on the actual data. Nearly all of the synthetic data coefficient estimates lie within one standard error of their corresponding actual data estimates, on average. Scatterplots of the synthetic and actual county regression coefficients (not shown, but available upon request) show that the synthetic data county estimates are in agreement with the actual county estimates as the points lie about the 45 degree line. However, there are clear biases associated with some coefficients, particularly, those associated with tenure variables that have already been shown to be affected by biases in the synthetic data. The standard errors of the synthetic data estimates appear to be on par, and in some cases, twice as large as those of the actual data estimates. In summary, the multivariate relationships examined here appear to be reasonably valid in the synthetic data. This is a reassuring result given that these relationships were explicitly accounted for in the synthetic data generation models.

Table 3. Summary Measures of Actual and Synthetic Linear and Logistic County Regression Coefficients

	Avg. Beta Coefficient		**Avg. Standard Error of Beta Coefficient**	
Linear regression of household income (cube root) on household-level covariates	Actual	Synthetic	Actual	Synthetic
Intercept	24.34	24.26	1.11	1.09
Household size	1.52	1.44	0.14	0.14
Sampling weight	-0.04	-0.05	0.24	0.26
Total bedrooms	1.15	1.23	0.19	0.18
Electricity bill/mo.	0.99	1.04	0.18	0.17
Total rooms	1.25	1.26	0.14	0.13
Tenure				
Mortgage/loan	Ref	Ref	Ref	Ref
Own free & clear	-3.47	-3.05	0.37	0.34
Rent	-6.01	-6.84	0.44	0.47
	Avg. Beta Coefficient		**Avg. Standard Error of Beta Coefficient**	
Logistic regression of poverty status on person-level covariates	Actual	Synthetic	Actual	Synthetic
Intercept	-2.39	-2.32	0.16	0.24
Sampling weight	0.25	0.25	0.07	0.10
Gender: Male	-0.33	-0.34	0.08	0.08
Education				
<12 years	Ref	Ref	Ref	Ref
12 years	-0.36	-0.35	0.12	0.13
13-15 years	-0.62	-0.63	0.13	0.15
16+years	-1.52	-1.59	0.18	0.30
Hispanic	0.36	0.27	0.29	0.63
Age	-0.00	0.01	0.00	0.07
Race				
White	Ref	Ref	Ref	Ref
Black	0.28	0.22	0.34	0.87
Other	0.41	0.41	0.25	0.56

5. ACS simulation

This section evaluates the repeated sampling properties of small area inferences drawn from the synthetic data based on a simulation study. In this simulation, we use public-use ACS microdata for the Northeast region for years 2005-2007. T he smallest geographical unit in the public-use microdata is a Public-Use Microdata Area (PUMA). PUMAs are defined as areas which contain at least 100,000 persons. In many cases, PUMAs overlap exactly with counties with the exception of very large counties, which are split into multiple PUMAs, and very small counties, which are combined with nearby counties to form a single PUMA. There are 405 PUMAs located in the Northeast region. For this simulation study, the ACS data is treated as a population from which subsamples are drawn. 500 stratified random subsamples are drawn from each PUMA with replacement. Each subsample accounts for approximately 30% of the total sample in each PUMA. Each ACS subsample is used as the basis for constructing a synthetic population from which 100 synthetic samples are drawn. This resulted in a total of 50,000 synthetic data sets.

Two types of inferences can be obtained from the synthetic data: conditional and unconditional. Conditional synthetic inferences are obtained from synthetic samples that are based on a single observed sample drawn from the population. This is the situation that most commonly occurs in practice, where a su rvey is carried out on a single population-based sample and the synthetic data is generated conditional on that sample. Unconditional inferences are obtained from synthetic samples that are based on multiple, or repeated, population-based samples. Obtaining unconditional inferences is not feasible in practice but is possible in the simulation study considered here.

To obtain conditional inferences, 500 sets of 10 synthetic samples are randomly selected (with replacement) from each of the 100 synthetic samples generated conditional on each of the 500 ACS subsamples. For each set of 10 synthetic samples, a synthetic estimate and associated 95% confidence interval are obtained for each variable in each PUMA using the combining rules of Section 2.2. To obtain unconditional inferences, 100 sets of 10 synthetic samples are randomly selected with replacement across each of the 100 ACS subsamples and point estimates and associated confidence intervals are again obtained using the relevant combining rules.

We use two evaluative measures to assess the validity of the synthetic data estimates. The first one is confidence interval coverage (CIC). For conditional inference, CIC is defined as t he proportion of times that the synthetic data confidence interval, computed at the 0.05 level,$[L_{\hat{q}_M,syn}, U_{\hat{q}_M,syn}]$ contains the actual estimate $\hat{y}_{act}$:

$$Q_{CIC} = I(\hat{y}_{act} \in \ [L_{\hat{q}_M,syn}, U_{\hat{q}_M,syn}])$$

where $I(\cdot)$ is an indicator function. $Q_{CIC} = 1 \text{if} L_{\hat{q}_M,syn} \leq \hat{y}_{act} \leq U_{\hat{q}_M,syn}$ and $Q_A = 0$ otherwise.

For unconditional inference, the only difference is that the CIC is calculated as the proportion of times that the synthetic data confidence interval contains the "true" population value Y_{pop}, i.e., $L_{\hat{q}_M,syn} \leq Y_{pop} \leq U_{\hat{q}_M,syn}$.

The second evaluative measure is referred to as the confidence interval overlap (CIO; Karr et al., 2006). CIO is defined as the average relative overlap between the synthetic and actual data confidence intervals. For every estimate the average overlap is calculated as,

$$Q_{CIO} = \frac{1}{2}\left(\frac{U_{over}-L_{over}}{U_{act}-L_{act}} + \frac{U_{over}-L_{over}}{U_{syn}-L_{syn}}\right),$$

where U_{act} and L_{act} denote the upper and the lower bound of the confidence interval for the actual estimate $\hat{y}_{act}$, U_{syn} and L_{syn} denote the upper and the lower bound of the confidence interval for the synthetic data estimate $\hat{q}_M$, and U_{over} and L_{over} denote the upper and lower bound of the overlap of the confidence intervals from the original and synthetic data for the estimate of interest. Q_{CIO}can take on any value between 0 and 1. A value of 0 means that there is no overlap between the two intervals and a value of 1 means that the synthetic interval completely covers the actual interval. Calculating the confidence interval overlap is only possible for conditional inferences. This measure yields a more accurate assessment of data utility in the sense that it accounts for the significance level of the estimate. That is, estimates with low significance might still have a high confidence interval overlap and therefore a high data utility even if their point estimates differ considerably from each other.

5.1. Validity of univariate estimates

Table 4 shows the average confidence interval coverage (CIC) and confidence interval overlap (CIO) across all PUMAs for univariate household-level estimands. The conditional CIC is high for non-recoded estimates ranging from 0.86-0.99. The income by tenure subgroup estimates also yield relatively high conditional CIC values (range: 0.89-0.97). The CIC values for income percentile estimates do not fare as well as they tend to decline monotonically as the percentiles increase. The same general trend is observed for the conditional CIO values, which closely resemble the CIC values. Regarding the unconditional inferences, the CIC values tend to be slightly higher than the corresponding values obtained from the conditional evaluation. The actual CIC values, obtained from the actual ACS subsamples, tend to be very close to the synthetic CIC values, if not slightly higher, except for the aforementioned percentile estimates which demonstrate weaker coverage for the most extreme percentiles.

Table 4. Simulation-Based Confidence Interval Results for PUMA Means

	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
Household variables				
Household size	0.99	0.97	0.98	0.98
Sampling weight	0.95	0.99	0.99	0.98
Bedrooms	0.89	0.87	0.93	0.98
Electricity cost/mo.	0.86	0.87	0.91	0.98
Rooms	0.97	0.93	0.98	0.98
Household income	0.90	0.91	0.94	0.98
Tenure				
Own free & clear	0.93	0.92	0.96	0.98
Rent	0.94	0.96	0.96	0.98

Table 4. Simulation-Based Confidence Interval Results for PUMA Means (cont.)

Recoded variables				
Income > 50th pctile	0.89	0.92	0.94	0.98
Income > 75th pctile	0.71	0.71	0.80	0.98
Income > 90th pctile	0.52	0.60	0.62	0.97
Income (Mortgage=1)	0.89	0.88	0.94	0.97
Income (Own=1)	0.91	0.98	0.96	0.96
Income (Rent=1)	0.97	0.93	0.99	0.96

5.2. Validity of multivariate estimates

Multivariate simulation results are shown in Table 5. T his table shows average CIC and CIO values for regression coefficient estimates obtained within each PUMA from a linear regression of income (cube root) on household-level covariates. The conditional CIC and CIO values are high and range from 0.93-0.99 and 0.90-0.98, respectively, indicating good analytic validity for these multivariate statistics. The unconditional CIC values range from 0.85-0.92, which are slightly below the actual CIC values obtained from the observed data (0.98). The lowest unconditional CIC values (0.85 and 0.87) are associated with the household tenure categories. Given that the analytic model being evaluated here is one of the same models used during the synthetic data generation process, it is not surprising that the analytic validity of the estimates is generally high. Overall, we believe this result is reassuring and underscores the importance of ensuring that the models used during the imputation process sufficiently overlap with the analytic models of interest.

Table 5. Simulation-Based Confidence Interval Results for PUMA Regression Coefficients

Linear regression of income (cube root) on	Conditional Inference		Unconditional Inference	
	CIC	CIO	CIC	CIC (Actual)
Intercept	0.98	0.97	0.92	0.98
Household size	0.98	0.95	0.91	0.98
Sampling weight	0.99	0.97	0.92	0.98
Total bedrooms	0.98	0.98	0.91	0.98
Electricity bill/mo.	0.99	0.97	0.91	0.98
Total rooms	0.98	0.97	0.92	0.98
Tenure				
Mortgage/loan	Ref	Ref	Ref	Ref
Own free & clear	0.95	0.90	0.87	0.98
Rent	0.93	0.96	0.85	0.98

6. Conclusions

Data users are increasingly interested in producing small area estimates, but statistical agencies are prevented from releasing these data due to disclosure concerns. In this article, a sy nthetic data methodology for generating and disseminating public-use microdata for small geographic areas was evaluated using restricted data from the U.S. Census Bureau. Compared with current practices of disseminating detailed geographical data, the synthetic data framework offers data users the flexibility of performing their own customizable geographic analyses using data that can presumably be released to the public without restriction.

The empirical evaluations show that the synthetic data generated from a Bayesian hierarchical model yields generally valid univariate and multivariate county-level estimates and repeated sampling properties. However, limitations of the method were apparent when simulating synthetic data for non-standard distributions and for polytomous variables when sample size limitations required pooling of nearby counties. Such limitations can potentially be overcome with more sophisticated modelling approaches, such as nonparametric imputation or mixture modelling, which was beyond the scope of this demonstration project. In addition, the "empirical" Bayesian approach considered here by fixing the hyperparameters at their maximum likelihood estimates may have underestimated the uncertainty of the synthetic data estimates, resulting in smaller standard errors and narrower confidence intervals. Although some underestimation of uncertainty might be welcomed in fully-synthetic data applications where standard errors are expected to be much higher relative to the observed standard errors, a more principled approach that accounts for all sources of variation might be viewed more favourably by sceptical data users.

Several extensions of this work are currently being considered. The preservation of skewed and non-standard distributions is an important issue that will need to be addressed prior to pubic release of synthetic small area microdata. Parametric modelling approaches are inherently limited in real-world applications where many of the most commonly used variables do not follow a smooth distributional form. The use of transformations to achieve normality is one possible solution; however, such transformations are not always effective for some types of distributions (e.g., bimodal). One must also consider the possibility that the same transformation might not work in all small areas. In this application, a single transformation was applied across all counties based on the overall distribution. Incorporating a tuning parameter in the hierarchical modelling approach that accounts for distributional differences across small areas might yield higher quality synthetic data and small area estimates with greater analytic validity. Another possible extension of this work is complex sample surveys. Although the ACS does not employ a complex sample design, most large-scale surveys do, and studies have shown that ignoring important design features during the imputation process can have drastic effects on the validity of the resulting estimates (Reiter, Raghunathan, and Kinney, 2006). Finally, the disclosure risk properties associated with fully synthetic data need to be studied in greater depth. Although we argue that fully synthetic data greatly enhances data confidentiality and prevents respondent re-identification because no observed data is released to the public, the extent to which confidentiality is protected needs to be systematically and empirically assessed.

Despite the potential for future improvements, the methodology examined here shows some promise and could be implemented by large-scale survey projects, such as t he American Community Survey, to release more geographically-relevant data to the public. Such efforts could potentially help meet the growing demand for small area microdata, which is expected to grow among a variety of data users across many disciplines.

Acknowledgements

An earlier version of this paper was presented at the 2011 Joint Statistical Meetings in Miami, FL and was selected for a Student Paper Award from the Government Statistics Section/Social Statistics Section/Survey Research Methods Section of the American Statistical Association. This research was supported by dissertation grants from the National Science Foundation (SES-0918942), the U.S. Census Bureau (YA132309SE0354), and the Centers for Disease Control and Prevention (1-R36-SH-000016-01). The authors are very grateful to Maggie Levenstein, Clint Carter, Arne Reznick, and Lynn Riggs for their assistance and support with accessing the restricted data from the Michigan Census Research Data Center.

REFERENCES

ABOWD, J. M., STINSON, M., BENEDETTO, G., (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. http://www.census.gov/sipp/SSAfinal.pdf.

BINDER, D. A., (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review,* 51, 279–292.

DATTA, G. S., FAY, R. E., GHOSH, M., (1991). Hierarchical and Empirical Bayes Analysis in Small-Area Estimation. *Proceedings of the Annual Research Conference*, U. S. Bureau of the Census, 63–78.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B., (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B,* 39, 1–38.

DRECHSLER, J., BENDER, S., RÄSSLER, S., (2008). Comparing Fully and Partially Synthetic Datasets for Statistical Disclosure Control in the German IAB Establishment Panel. *Transactions on Data Privacy,* 105–130.

FAY, R. E., HERRIOT, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association,* 74, 269–277.

KARR, A. F., KOHNEN, C. N., OGANIAN, A., REITER, J. P., SANIL, A. P., (2006). A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality. *The American Statistician,* 60, 224–232.

KENNICKELL, A. B. (1997). Multiple Imputation and Disclosure Protection: The Case of the 1995 Survey of Consumer Finances. In *Record Linkage Techniques*. W. Alvey and B. Jamerson (eds.) Washington D. C.: National Academy Press, 248–267.

KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S., ABOWD, J. M., (2011). Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database. *International Statistical Review,* 79, 362–384.

LINDLEY, D. V., SMITH, A. F. M., (1972). Bayes Estimates for the Linear Model. *Journal of the Royal Statistical Society, Series B,* 34, 1–41.

LITTLE, R. J. A., (1993). Statistical Analysis of Masked Data. *Journal of Official Statistics,* 9, 407–426.

LITTLE, R. J. A., RUBIN, D. B., (2002). *Statistical Analysis with Missing Data*. 2nd Edition. Wiley.

LIU, F., LITTLE, R. J. A., (2002). Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata. In *ASA Proceedings of the Joint Statistical Meetings,* 2, 2133–2138.

MACKIE, C., BRADBURN, N., (2000). *Improving Access to and Confidentiality of Research Data: Report of a Workshop.* Commission on Behavioral and Social Sciences and Education, National Research Council. National Academy Press, Washington, D. C.

MALEC, D., SEDRANKS, J., MORIARITY, C. L., LECLERE, F. B., (1997). Small Area Inference for Binary Variables in the National Health Interview Survey. *Journal of the American Statistical Association,* 92, 815–826.

PLATEK, R., RAO, J. N. K., SÄRNDAL, C. E., SINGH, M. P., (1987). *Small Area Statistics*. Wiley, New York.

RAGHUNATHAN, T. E., RUBIN, D. B., (2000). Bayesian Multiple Imputation to Preserve Confidentiality in Public-Use Data Sets. *ISBA 2000 The Sixth World Meeting of the International Society for Bayesian Analysis*.

RAGHUNATHAN, T. E., LEPKOWSKI, J. M., VAN HOEWYK, J., SOLENBERGER, P., (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a S equence of Regression Models. *Survey Methodology,* 27, 85–95.

RAGHUNATHAN, T. E, REITER, J. P., RUBIN, D. B., (2003). Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics,* 19, 1–16.

RAO, J. N. K., (1999). Some Recent Advances in Model-based Small Area Estimation. *Survey Methodology,* 25, 175–186.

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

REITER, J. P., (2002). Satisfying Disclosure Restrictions with Synthetic Data Sets. *Journal of Official Statistics,* 18, 531–544.

REITER, J. P., (2003). Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology,* 29, 181–188.

REITER, J. P., (2004).Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation. *Survey Methodology,* 30, 235–242.

REITER, J. P., (2005). Releasing Multiply-Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study. *Journal of the Royal Statistical Society, Series A,* 168, 185–205.

REITER, J. P., RAGHUNATHAN, T. E., KINNEY, S. K., (2006). The Importance of Modeling the Survey Design in Multiple Imputation for Missing Data. *Survey Methodology,* 32, 143–150.

REITER, J. P., RAGHUNATHAN, T. E., (2007).The Multiple Adaptations of Multiple Imputation. *Journal of the American Statistical Association,* 102, 1462–1471.

RODRIGUEZ, R., (2007). Synthetic Data Disclosure Control for American Community Survey Group Quarters. In *ASA Proceedings of the Joint Statistical Meetings,* 1439–1450.

RUBIN, D. B., (1983). A Case-Study of the Robustness of Bayesian/Likelihood Methods of Inference: Estimating the Total in a Finite Population using Transformations to Normality. In *Scientific Inference, Data Analysis and Robustness*. G.E.P. Box, T. Leonard, and C.F. Wu (eds.) New York: Academic Press, 213–244.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley: New York.

RUBIN, D. B., (1993). Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-Imputed Microdata. *Journal of Official Statistics,* 9, 461–468.

TRANMER, M., PICKLES, A., FIELDHOUSE, E., ELLIOT, M., DALE, A., BROWN, M., MARTIN, D., STEEL, D., GARDINER, C., (2005). The Case for Small Area Microdata. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168, 29–49.

U.S. CENSUS BUREAU, (2009). American Community Survey: Design and Methodology. http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf

YUCEL, R. M., (2008). Multiple Imputation Inference for Multivariate Multilevel Continuous Data with Ignorable Non-Response. *Philosophical Transactions of the Royal Society A,* 366, 2389–2403.

APPENDICES

Appendix 1. EM algorithm for estimating Bayesian hyperparameters

The EM algorithm is used to estimate the unknown population parameters $\beta_p$and $\Sigma_p$from the following setup,

$$\hat{\beta}_{cs,p} \sim MVN(\beta_{cs,p}, \hat{V}_{cs,p})$$

$$\beta_{cs,p} \sim MVN(\beta_p Z_s, \Sigma_p)$$

where$p = (1,2, \dots, P)$ is used to index the set of parameters associated with the p^{th} synthetic variable of interest and the p^{th} regression model from which the direct estimates $\hat{\beta}_{cs}$ and $\hat{V}_{cs}$ were obtained in Step 1.

The E step consists of solving the following expectations,

$$\beta^*_{cs,p} = E(\beta_{cs,p}) = \left[\left(\hat{\mathrm{V}}^{-1}_{cs,p} + \Sigma^{-1}_p\right)^{-1}\left(\hat{\mathrm{V}}^{-1}_{cs,p}\hat{\beta}_{cs} + \Sigma^{-1}_p\beta_p Z_s\right)\right]$$

$$\left[\beta_{cs,p}(\beta_{cs,p})^T\right]^* = E\left[\beta_{cs,p}\beta^T_{cs,p}\right] = \left(\hat{\mathrm{V}}^{-1}_{cs,p} + \Sigma^{-1}_p\right)^{-1} + \beta^*_{cs,p}\left(\beta^*_{cs,p}\right)^T$$

Once these expectations are computed they are then incorporated into the maximization (M-step) of the unknown hyperparameters β_p and$\hat{\Sigma}_p$ using the following equations,

$\hat{\beta}_p = \beta^*_{+s,p} Z_s (Z_s Z^T_s)^{-1}$, where $\beta^*_{+s} = \left(\sum^{C_s}_{c=1} \beta^*_{cs}\right)/C_s$, and

$$\hat{\Sigma}_p = \frac{\left[\sum^S_{s=1} \left[\sum^{C_s}_{c=1}\left(\beta^*_{cs,p} - \hat{\beta}_p Z_s\right)\left(\beta^*_{cs,p} - \hat{\beta}_p Z_s\right)^T\right] \Big/ C_s\right]}{S}$$

After convergence the maximum likelihood estimates are incorporated into the posterior distribution of $\beta_{cs,p}$ shown in equation [5].

Appendix 2. Creation of synthetic household size variable

Let Z_{hcs} be the number of people in household $h = (1,2,\dots,n_{cs})$ in county $c = (1,2,\dots,C_s)$ within state $s = (1,2,\dots,S)$. Assume that $Z_{hcs} \sim Poisson(\lambda_{cs})$and $\lambda_{cs} \sim Gamma(\alpha_s, \beta_s)$. Conditional on the data and $(\alpha_s, \beta_s; s = 1,2,\dots,S)$ it is straightforward to simulate values of Z_{hcs}.

First, obtain the marginal maximum likelihood estimates of $(\alpha_s, \beta_s; s = 1,2,\dots,S)$ through Newton-Raphson for each state independently. Also, obtain the covariance matrix $\hat{V_s} = Cov(\hat{\alpha}_s, \hat{\beta}_s)$ by inverting the observed Fisher Information matrix. The marginal likelihood is given by,

$$\int \left\{ \prod_{c=1}^{C_s} e^{-\beta_s \lambda_{cs}} \lambda_{cs}^{\alpha_s - 1} \left(\prod_{h=1}^{n_{cs}} e^{-\lambda_{cs}} \lambda_{cs}^{Z_{hcs}} \right) / \Gamma(\alpha_s) d\lambda_{cs} \right\}$$

$$= \prod_{c=1}^{C_s} \int e^{-(\beta_s + n_{cs})\lambda_{cs}} \lambda_{cs}^{Z_{+cs} + \alpha_s - 1} / \Gamma(\alpha_s) \beta_s^{\alpha_s} \, d\lambda_{cs}$$

$$= \prod_{c=1}^{C_s} \{\Gamma(Z_{+cs} + \alpha_s)\}(\beta_s + n_{cs})^{-(Z_{+cs} + \alpha_s)} / \Gamma(\alpha_s) \, \beta_s^{\alpha_s}$$

where $Z_{+cs} = \sum_{h=1}^{n_{cs}} Z_{hcs}$. Taking the logarithms, the quantity to be maximized with respect to α_s and b_s via the Newton-Raphson is,

$$L = \sum_{c=1}^{C_s} \{log\Gamma(Z_{+cs} + \alpha_s) - (Z_{+cs} + \alpha_s) log(\beta_s + n_{cs})\} - C_s log\Gamma(\alpha_s) + C_s \alpha_s log(\beta_s)$$

The first and second derivatives of this function are,

$$\frac{\partial L}{\partial \alpha_s} = \sum_{c=1}^{C_s} \{\psi(Z_{+cs} + \alpha_s) - log(\beta_s + n_s)\} - C_s \psi(\alpha_s) + C_s log(\beta_s)$$

$$\frac{\partial L}{\partial \beta_s} = -\sum_{c=1}^{C_s} \{(Z_{+cs} + \alpha_s)/(\beta_s + n_s)\} + C_s \alpha_s / \beta_s$$

$$\frac{\partial^2 L}{\partial \alpha_s^2} = \sum_{c=1}^{C_s} \psi'(Z_{+cs} + \alpha_s) - C_s \psi'(\alpha_s)$$

$$\frac{\partial^2 L}{\partial \beta_s^2} = \sum_{c=1}^{C_s} \{(Z_{+cs} + \alpha_s)/(\beta_s + n_s)^2\} - \alpha_s C_s / \beta_s^2$$

$$\frac{\partial^2 L}{\partial \beta_s \partial \alpha_s} = -\sum_{c=1}^{C_s} 1/(\beta_s + n_s) + C_s / \beta_s$$

The logarithm of the gamma function, its first and second derivatives can be accurately approximated as follows,

$$log\Gamma(z) = -log\sum_{i=1}^{26} c_i z^i$$

$$\psi(z) = \frac{\partial}{\partial z} log\Gamma(z) = -\frac{\sum_{i=1}^{26} ic_i z^{i-1}}{\sum_{i=1}^{26} c_i z^i}$$

$$\psi'(z) = \left(\frac{\sum_{i=1}^{26} ic_i z^{i-1}}{\sum_{i=1}^{26} c_i z^i}\right)^2 - \frac{\sum_{i=1}^{26} i(i-1)c_i z^{i-2}}{\sum_{i=1}^{26} c_i z^i}$$

The constants c_i can be found in Abramowitz and Stegun (1965). The Newton-Raphson method is applied iteratively to obtain maximum likelihood estimates of $\alpha_s$and β_s,

$$\begin{pmatrix}\alpha_{s,n+1}\\ \beta_{s,n+1}\end{pmatrix} = \begin{bmatrix} {}^{\partial^2 L}\!/_{\partial\alpha_{s,n}^2} & {}^{\partial^2 L}\!/_{\partial\alpha_{s,n}\partial\beta_{s,n}} \\ {}^{\partial^2 L}\!/_{\partial\beta_{s,n}\partial\alpha_{s,n}} & {}^{\partial^2 L}\!/_{\partial\beta_{s,n}^2} \end{bmatrix}^{-1} \begin{pmatrix} {}^{\partial L}\!/_{\partial\alpha_{s,n}} \\ {}^{\partial L}\!/_{\partial\beta_{s,n}} \end{pmatrix}$$

The logarithm of the estimates for α_s and β_s are then assumed to follow the hierarchical model,

$$\begin{pmatrix}log\,\hat{\alpha}_s\\ log\,\hat{\beta}_s\end{pmatrix} \sim N\left[\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix}, \begin{bmatrix}1/\hat{\alpha}_s & 0\\ 0 & 1/\hat{\beta}_s\end{bmatrix}\hat{V}_s\begin{bmatrix}1/\hat{\alpha}_s & 0\\ 0 & 1/\hat{\beta}_s\end{bmatrix}\right] = N\left[\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix}, \hat{\Sigma}_s\right]$$

$$\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix} \sim N\left[\begin{pmatrix}\theta\\ \phi\end{pmatrix}, \begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{22} & \Omega_{22}\end{bmatrix}\right] = N\left[\begin{pmatrix}\theta\\ \phi\end{pmatrix}, \Omega\right]$$

The Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) is used to obtain maximum likelihood estimates of (θ, ϕ, Ω). The E step is carried out by solving the following expectation equations,

$$\begin{pmatrix}log\,\alpha_s^*\\ log\,\beta_s^*\end{pmatrix} = E\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix} = \left[\left(\hat{\Sigma}_s^{-1} + \Omega^{-1}\right)^{-1}\left(\hat{\Sigma}_s^{-1}\begin{pmatrix}log\,\hat{\alpha}_s\\ log\,\hat{\beta}_s\end{pmatrix} + \Omega^{-1}\begin{pmatrix}\theta\\ \phi\end{pmatrix}\right)\right]$$

$$\left[\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix}\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix}^T\right]^* = E\left[\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix}\begin{pmatrix}log\,\alpha_s\\ log\,\beta_s\end{pmatrix}^T\right]$$
$$= \left(\hat{\Sigma}_s^{-1} + \Omega^{-1}\right)^{-1} + \begin{pmatrix}log\,\alpha_s^*\\ log\,\beta_s^*\end{pmatrix}\begin{pmatrix}log\,\alpha_s^*\\ log\,\beta_s^*\end{pmatrix}^T$$

and the *M* step is performed by solving the following maximization equations,

$$\begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix} = \frac{\left[\sum_{s=1}^{S}\begin{pmatrix}log\ \alpha_s^*\\ log\ \beta_s^*\end{pmatrix}\right]}{S}$$

$$\hat{\Omega} = \begin{bmatrix}\hat{\Omega}_{11} & \hat{\Omega}_{12}\\ \hat{\Omega}_{22} & \hat{\Omega}_{22}\end{bmatrix} = \frac{\left[\sum_{s=1}^{S}\left(\begin{pmatrix}log\ \alpha_s^*\\ log\ \beta_s^*\end{pmatrix} - \begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix}\right)\left(\begin{pmatrix}log\ \alpha_s^*\\ log\ \beta_s^*\end{pmatrix} - \begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix}\right)^T\right]}{S}$$

It is then straightforward using this setup to synthesize the number of members in each household by treating the parameter estimates of (θ, ϕ, Ω) as known and retracing back to simulate values of Z_{hcs} using the following 3 steps:

Step 1: Simulate Gamma parameters α_s and β_s from the bivariate normal distribution, $\begin{pmatrix}\tilde{\alpha}_s\\ \tilde{\beta}_s\end{pmatrix} \sim exp\left[N\left[\left(\hat{\Sigma}_s^{-1} + \hat{\Omega}^{-1}\right)^{-1}\left(\hat{\Sigma}_s^{-1}\begin{pmatrix}log\ \hat{\alpha}_s\\ log\ \hat{\beta}_s\end{pmatrix} + \Omega^{-1}\begin{pmatrix}\hat{\theta}\\ \hat{\phi}\end{pmatrix}\right), \left(\hat{\Sigma}_s^{-1} + \hat{\Omega}^{-1}\right)^{-1}\right]\right]$,

Step 2: Simulate Poisson parameter λ_{cs} from the Gamma distribution given the county population size, number of households, and simulated parameters obtained from Step 1,
$\tilde{\lambda}_{cs} \sim Gamma\left(Z_{+cs} + \tilde{\alpha}_s, \tilde{\beta}_s + \ n_{cs}\right)$,

Step 3: Simulate household size $Z_{hcs}$from the Poisson distribution,
$\tilde{Z}_{hcs} \sim Poisson\left(\tilde{\lambda}_{cs}\right)$.

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 369–388

ESTIMATING POPULATION MEAN WITH MISSING DATA IN UNEQUAL PROBABILITY SAMPLING

Kajal Dihidar [1]

ABSTRACT

Nonresponse problem is a serious obstacle to the validity of estimates in a survey. The estimates become biased due to the missing values in data. The problem is how to deal with missing values, once they have been deemed impossible to recover. One way of exploring a possible lack of representativity in missing data is to estimate the response probabilities which are usually done by logistic regression model. However, the drawback of the logit model is that this requires values of the explanatory variables of the model to be known for all nonrespondents. Bethlehem (2012) showed that the response probabilities can be estimated by some weighting adjustment technique without having the individual data of the nonrespondents. Here we consider the doubtful nature of nonresponse regarding possible existence of relationship with any of the covariates. Moreover, instead of simple random sampling, we consider general unequal probability sampling scheme for selecting respondents. This paper presents the modification of Bethlehem (2012) proposal for unequal probability sampling to obtain the unbiased estimators for population total/average of a variable of interest and variance estimator and compares them with the usual estimators through numerical simulations.

Key words: non-response, missing at random, missing completely at random, unequal probability sampling.

1. Introduction

Almost all large scale sample surveys suffer the problem with missing data. It may occur even if an investigator tries to have all questions fully responded to in a survey, or if the respondent is not available at home to answer the questionnaire. One of the effects of nonresponse is that the sample size is smaller than expected. This would lead to less accurate, but still valid estimates of population characteristics, which can be taken care of by taking the initial sample size larger. A far more serious effect of nonresponse is that estimates of population characteristics may be biased. This situation occurs if, due to nonresponse, some groups in the population are over- or under-represented, and these groups behave differently with respect

[1] Sampling and Official Statistics Unit, Indian Statistical Institute, Kolkata, West Bengal, India. E-mail: kajaldihidar@gmail.com, dkajal@isical.ac.in.

to the characteristics to be investigated. Consequently, wrong conclusions will be drawn from the survey data. The amount of bias created due to missing values often increases with the rate of occurence of nonresponse. Above all, the large number of missing values in the data set can also lead to computational difficulties.

Towards this problem, starting from the pioneered work by Hansen and Hurwitz (1946), many methods of attempts to re-collect the missing values in sample surveys are available in the literature. However, in most of the practical sample survey work, it is not possible to recover the actual missing values. In such situations, the problem is how to estimate the population parameters dealing with the missing values. The method of response modeling and imputation are popular to survey statisticians in this direction. Good details regarding this are given in Rubin (1987) and S$\ddot{a}$rndal, Swenson and Wretman (1992).

In general, obtaining the responses from the selected units is totally unknown in advance. For this reason, the probabilistic models are assumed to describe the unknown response distributions. Politz and Simmons (1949, 1950) obtained the response probability of a respondent as the proportion of time staying at home. The response probability may be directly related to the study variable and hence to the auxiliary variable, which is highly related to the study variable. For example, in the study of household income, the people with high income may respond with low probability and may be under represented in the sample. Similarly, if tax return is considered as an auxiliary variable, then the response probability of an individual may be inversely proportional to the amount of tax return.

Regarding the possible relatioship of missingness with any of the covariates, Rubin (1976) defined the concepts of missing at random (MAR) and missing completely at random (MCAR). Missing completely at random (MCAR) means that the missing data is not related to the values of any variable, neither to the response variable itself nor to other covariates, whether missing or observed; whereas missing at random (MAR) means that the missing data is unrelated to the actual missing values but is related either to observed covariates or to observed response variable itself or to both. Among many contributors in this area, Folsom (1991), Fuller et al. (1994), Kott (2006), Chang and Kott (2008) and Kott and Chang (2010) advocated the use of calibration weighting to adjust for unit nonresponse. In this regard, for more detailed clarification, interested researchers may see Heitzan and Basu (1996), Singh (2010).

In case the covariate relation is considered, the concept of the response propensity is introduced in Little (1986). The response propensity is the probability of response given the values of some auxiliary variables. The response propensities are also unknown, so they need to be estimated. For this purpose, the logistic model is used in practice. Of course, another model sometimes used is the probit model. Estimates of the coefficients in both the logit and probit models are obtained by maximum likelihood estimation. And the estimated response propensities in these two models are always in the interval [0, 1]. However, the drawback of the logit and probit models is that these require the values of the explanatory variables of the model to be known for all nonrespondents. Bethlehem (2012) showed that this condition can be relaxed by computing response probabilities from weights that have been obtained from some weighting adjustment technique. This technique produces weights that correct for the lack of representativity of the survey response. Since the weights can be seen as a kind of inverted response probabilities, they can be used to estimate response probabilities. Weights are computed following those techniques without having the individual data of the nonrespondents. We use this approach to estimate the response propensities from correlated auxiliary variables.

In this paper, we consider the situation where some of the respondents selected using an unequal probability sampling scheme fail to respond and the nature of non-response is uncertain as to whether it is MAR or MCAR. Moreover, instead of considering the simple random sampling, we consider any general unequal probability sampling scheme even without replacement for selecting the respondents because we believe that many of the practical cases of large-scale sample surveys require the selection of respondents with probability proportional to size measures of some auxiliary variable related to study variable. Under the consideration of doubtful nature of random nonresponse, we shall derive here unbiased estimators for population total/average of a variable of interest and variance estimators in unequal probability sampling scheme. The derived estimators will be compared with usual estimators in presence of random nonresponse through numerical simulations.

We organize our findings in the following sections.

2. Unbiased estimator of population mean and variance with missing data

Suppose in a finite survey population $U = (1, \ldots, i, \ldots, N)$ a person labelled i has the value y_i defined on a variable y of interest and has value $x_i > 0$ defined on an auxiliary variable x closely related to the study variable y. The values of x are all positive and known for all the population units in U. Our problem is to estimate $\bar{Y} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} y_i$ on the basis of a sample s of size n, selected with probability $p(s)$ according to a sampling design p.

Let π_i and π_{ij} be the first and second order inclusion probabilities of the units in U. Let us define a random variable δ_i as

$$\delta_i = \begin{cases} 1 & \text{if } i^{th} \text{ unit responds,} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Let E_P, V_P denote the expectation and variance operators with respect to the sampling design for selecting the respondents. Let E_R, V_R denote the expectation and variance operators with respect to obtaining a response from the selected respondent, and E, V denote the overall expectation and variance operators. In this setup, δ_i is a Bernoulli random variable with probability of success as δ_i^*, say, and it is known. So, $E_R(\delta_i)$ = Prob($\delta_i = 1$) = δ_i^*, and $V_R(\delta_i)$ = $\delta_i^*(1 - \delta_i^*)$. We first of all assume that the value of response probability depends on some auxiliary variables which are well correlated with the study variable, but the exact relationship of the response probability with the auxiliary variables is unknown to us. We can get possible relationships based on some statistical testing whether the available data is MCAR or MAR. For example, we can apply some variable selection method to see whether or not the response propensity depends (or not) on a set of auxiliary variables. However, a weighted sum of these two estimators (MCAR and MAR) would be an alternative to choosing one over the other to balance their degree of bias. This type of estimator, namely the 'composite estimator' is formed by compromising in between the MAR estimator and MCAR estimator, with a compromising factor $\lambda (0 < \lambda < 1)$.

The composite estimator of population total Y will be obtained as

$$\hat{Y}_{comp} = \lambda \hat{Y}_{MCAR} + (1-\lambda)\hat{Y}_{MAR},$$

where $\hat{Y}_{MCAR}$ and $\hat{Y}_{MAR}$ respectively denote the MCAR and MAR estimators for Y. We may get the optimal compromising factor by minimmizing the MSE of the composite estimator with respect to λ under the assumption that the covariance factor of $\hat{Y}_{MCAR}$ and $\hat{Y}_{MAR}$ is too small relative to the MSE of $\hat{Y}_{MAR}$ and then it can be negligible. In this situation, the optimal compromising factor λ_{opt} may be obtained as

$$\lambda_{opt} = \frac{MSE(\hat{Y}_{MAR})}{MSE(\hat{Y}_{MCAR}) + MSE(\hat{Y}_{MAR})}.$$

In practical situation, λ_{opt} can be estimated by substituting the estimates of $MSE(\hat{Y}_{MCAR})$ and of $MSE(\hat{Y}_{MAR})$ based on the sample survey data in above expression of λ_{opt}.

2.1. Unbiased estimator of population mean

Under the non-response setup, a homogeneous linear unbiased estimator for population mean is

$$\hat{\bar{Y}} = \frac{1}{N}\sum_{i\in s} y_i b_{si}\left(\frac{\delta_i}{\delta_i^*}\right) = \frac{1}{N}\sum_{i\in s} u_i b_{si}, \text{ where } u_i = y_i\frac{\delta_i}{\delta_i^*} \tag{2}$$

and b_{si}'s are free of y_i's and satisfy $\sum_{s\ni i} p(s) b_{si} = 1, \forall\, i \in U$.

This happens because

$$E_R(u_i) = \frac{y_i}{\delta_i^*}E_R(\delta_i) = \frac{y_i}{\delta_i^*}\delta_i^* = y_i, \tag{3}$$

and

$$\begin{aligned} E\left(\hat{\bar{Y}}\right) &= E_P E_R\left[\frac{1}{N}\sum_{i\in s} u_i b_{si}\right] = E_P\left[\frac{1}{N}\sum_{i\in s} b_{si}E_R(u_i)\right] \\ &= E_P\left[\frac{1}{N}\sum_{i\in s} y_i b_{si}\right] = \bar{Y}. \end{aligned} \tag{4}$$

2.2. Variance of the unbiased estimator of population mean

From the definition of u_i, we have

$$V_R(u_i) = \frac{y_i^2}{\delta_i^{*2}} V_R(\delta_i) = \frac{y_i^2(1-\delta_i^*)}{\delta_i^*}. \tag{5}$$

So, the variance of the estimator given in Eqn. (5) is

$$\begin{aligned} V\left[\hat{\bar{Y}}\right] &= V_P E_R\left[\frac{1}{N}\sum_{i\in s} u_i b_{si}\right] + E_P V_R\left[\frac{1}{N}\sum_{i\in s} u_i b_{si}\right] \\ &= V_P\left[\frac{1}{N}\sum_{i\in s} y_i b_{si}\right] + E_P\left[\frac{1}{N^2}\sum_{i\in s} b_{si}^2 V_R(u_i)\right] \\ &= \frac{1}{N^2}\left[\sum_{i=1}^N y_i^2 c_i + \sum_{i=1}^N \sum_{j=1,j\neq i}^N y_i y_j c_{ij} + E_P\left(\sum_{i\in s} b_{si}^2 \frac{y_i^2}{\delta_i^*}(1-\delta_i^*)\right)\right] \\ &= \frac{1}{N^2}\left[\sum_{i=1}^N y_i^2 c_i + \sum_{i=1}^N \sum_{j=1,j\neq i}^N y_i y_j c_{ij} + \left(\sum_{i=1}^N \frac{y_i^2 b_{si}}{\delta_i^*}(1-\delta_i^*)\right)\right], \end{aligned} \tag{6}$$

where $c_i = E_P(b_{si}^2 I_{si}) - 1$ and $c_{ij} = E_P(b_{si} b_{sj} I_{sij}) - 1$ where I_{si} and I_{sij} are defined as

$$I_{si} = \begin{cases} 1 & \text{if } i \in s, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

and $I_{sij} = I_{si} I_{sj}$.

2.3. Unbiased variance estimator for population mean

First of all, we find an unbiased estimator for $V_R(u_i)$. We note that $\delta_i^2 = \delta_i$ and so,

$$E_R(u_i^2) = E_R\left[\frac{y_i^2\delta_i^2}{\delta_i^{*2}}\right] = E_R\left[\frac{y_i^2\delta_i}{\delta_i^{*2}}\right] = \frac{y_i^2}{\delta_i^{*2}} E_R[\delta_i] = \frac{y_i^2}{\delta_i^*},$$

and so

$$E_R[u_i^2\delta_i^*] = y_i^2. \tag{8}$$

Now,

$$\begin{aligned} V_R(u_i) &= E_R(u_i^2) - (E_R(u_i))^2 = E_R(u_i^2) - y_i^2 \\ &= E_R(u_i^2) - E_R[u_i^2\delta_i^*] = E_R[u_i^2(1-\delta_i^*)] \end{aligned} \tag{9}$$

implies that

$$v_R(u_i) = \hat{V}_R(u_i) = u_i^2(1-\delta_i^*). \tag{10}$$

Let c_{si} and c_{sij} be such that $E_P(c_{si}I_{si}) = c_i$ and $E_P(c_{sij}I_{sij}) = c_{ij}$.
We define

$$v_1 = \frac{1}{N^2}\left[\sum_{i\in s} u_i^2 c_{si} + \sum_{i\in s}\sum_{j\in s, j\neq i} u_i u_j c_{sij} + \sum_{i\in s} v_R(u_i)(b_{si}^2 - c_{si})\right], \tag{11}$$

and

$$v_2 = \frac{1}{N^2}\left[\sum_{i\in s} u_i^2 c_{si} + \sum_{i\in s}\sum_{j\in s, j\neq i} u_i u_j c_{sij} + \sum_{i\in s} v_R(u_i) b_{si}\right]. \tag{12}$$

Following Raj (1966), we have $E_P E_R(v_1) = V(\hat{\hat{Y}}) = E_P E_R(v_2)$, and so v_1 and v_2 are two unbiased estimators for $V(\hat{\hat{Y}})$.

3. Estimation of response probability

The true response probability δ_i^* as discussed in Section 2 is practically unknown in advance. So, we need to use an estimator for this.

If no covariate relation is considered, the missing data is considered as missing completely at random (MCAR), then the probability of response (assuming same for all units) is estimated by $\frac{r}{n}$, where n is the sample size and r is the number of responses obtained out of n persons sampled.

If the covariate relation is considered, the concept of the response propensity is introduced in Little (1986). He has defined the response propensity of element i as

$$\delta_i^*(\mathbf{X}) = P(\delta_i = 1|\mathbf{X}_i), \tag{13}$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})'$ is a vector of values of, say, p auxiliary variables. So, the response propensity is the probability of response given the values of some auxiliary variables. The response propensities are also unknown, so they need to be estimated.

3.1. Traditional models

The most frequently used model to estimate the response propensities is the logistic regression model. It assumes the relationship between response propensity and auxiliary variables as

$$\text{logit}(\delta_i^*(\mathbf{X})) = \log\left(\frac{\delta_i^*(\mathbf{X})}{1-\delta_i^*(\mathbf{X})}\right) = \sum_{j=1}^{p} X_{ij}\beta_j, \tag{14}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ is a vector of p regression coefficients.

Of course, another model, the probit model can also be used. It assumes the relationship between response propensity and auxiliary variables as

$$\text{probit}(\delta_i^*(\mathbf{X})) = \Phi^{-1}\left(\delta_i^*(\mathbf{X})\right) = \sum_{j=1}^{p} X_{ij}\beta_j, \tag{15}$$

where Φ^{-1} is the inverse of the $N(0,1)$ distribution function.

Estimates of the coefficients in both the logit and probit models can be obtained by maximum likelihood estimation. And the estimated response propensities in these two models are always in the interval [0, 1].

However, the drawback of the logit and probit models is that these require the values of the explanatory variables of the model to be known for all nonrespondents. But this is not the situation in many cases. To overcome this drawback, we follow Bethlehem (2012) model to estimate the response propensities and this is described below.

3.2. Bethlehem Model

Bethlehem (2012) showed how to estimate the response probabilities from weights that have been obtained from some weighting adjustment technique without having the individual data of the nonrespondents. The basic idea is to assign weights to responding elements in such a way that over-represented groups get a weight smaller than 1 and under-represented groups get a weight larger than 1. There is a relationship between response probabilities and weights: large weights correspond to small response probabilities, and vice versa. Therefore, it should be possible to transform weights into estimates for response probabilities.

There are several types of weighting techniques. The most frequently used ones are post-stratification, generalized regression estimation and raking ratio estimation. Weighting is based on the use of auxiliary information. Auxiliary information is defined here as a set of variables that have been measured in the survey, and for which the distribution in the population, or in the complete sample, is available. The individual values of the auxiliary variables are not required for the nonresponding elements. Among several weighting techniques, we adopt here the generalized regression estimation technique for simplicity. The generalized regression estimator is based on a linear model that attempts to explain a target variable of the survey from one or more auxiliary variables. The weights resulting from generalized regression estimation make the response representative with respect to the auxiliary variables in the model (S$\ddot{a}$rndal, 2011).

In principle, the auxiliary variables in the linear model have to be continuous variables, i.e. they measure a size, value or duration. However, it is also possible to use categorical variables. The trick is to replace a categorical variable by a set of dummy variables, where each dummy variable represents a category, i.e. it indicates whether or not a person belongs to a specific category. Suppose there are p

(continuous) auxiliary variables available. The p-vector of values of these variables for element i is $\mathbf{X}_i$.

Let Y be the N-vector of all values in the population of the target variable, and let $\mathbf{X}$ be the $N \times p$-matrix of all values of the auxiliary variables. The vector of population means of the p auxiliary variables is defined by

$$\bar{\mathbf{X}} = (\bar{\mathbf{X}}_1, \bar{\mathbf{X}}_2, \ldots, \bar{\mathbf{X}}_p)'.$$

We assume that this vector representing the population information is available, based on some expert guess or on the result of some prior survey. If the auxiliary variables are correlated with the target variable, then for a suitably chosen vector $\mathbf{B} = (B_1, B_2, \ldots, B_p)'$ of regression coefficients for a best fit of Y on $\mathbf{X}$, the residuals $E = (E_1, E_2, \ldots, E_N)'$, defined by $E = Y - \mathbf{XB}$ will vary less than the values of the target variable itself. The population regression coefficient B obtained by applying ordinary least squares technique is

$$B = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \left(\sum_{i=1}^{N} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i \mathbf{Y}_i\right). \tag{16}$$

The vector $\mathbf{B}$ can be estimated by

$$\mathbf{b} = \left(\sum_{i \in s} \pi_i^{-1} \mathbf{x}_i \mathbf{x}_i' \delta_i\right)^{-1} \left(\sum_{i \in s} \pi_i^{-1} \mathbf{x}_i \mathbf{y}_i \delta_i\right), \tag{17}$$

where π_i is the first order inclusion probability of unit i in sample s.

Let $\mathbf{X}_s$, $\mathbf{Y}_s$ be $n \times p$ and $n \times 1$ versions of $\mathbf{X}$ and $\mathbf{Y}$ for the units $i \in s$ where n is the sample size. Let $\mathbf{W}_s$ be the $n \times n$ diagonal matrix with the weights w_i for the units $i \in s$ on the diagonal. The Horvitz-Thompson (1952) weights are $w_i = 1/\pi_i$. Also let δ_s be the $n \times n$ diagonal matrix with values δ_i for the units $i \in s$ on the diagonal. The vector $\mathbf{b}$ can then be written in matrix form as

$$\mathbf{b} = (\mathbf{X_s'} \mathbf{W_s} \delta_\mathbf{s} \mathbf{X_s})^{-\mathbf{1}} (\mathbf{X_s'} \mathbf{W_s} \delta_\mathbf{s} \mathbf{Y_s}). \tag{18}$$

The generalized regression estimator is now defined by

$$\bar{y}_{GR} = \frac{1}{N}\left[\sum_{i \in s} \frac{y_i \delta_i}{\pi_i} + (\mathbf{X} - \sum_{i \in s} \pi_i^{-1} \mathbf{x_i} \delta_\mathbf{i})'\mathbf{b}\right]. \tag{19}$$

Following Bethlehem and Keller (1987), the generalized regression estimator can be rewritten in the form of the weighted estimator as

$$\bar{y}_{GR} = \sum_{i \in s} w_i y_i \delta_i, \tag{20}$$

where the weights are

$$w_i = \bar{\mathbf{X}} \left[\sum_{j \in s} \pi_j^{-1} \mathbf{x}_j' \mathbf{x}_j \delta_j\right]^{-1} \pi_i^{-1} \mathbf{x}_i', \tag{21}$$

where $\mathbf{x}_j$ is the k-dimensional vector of control variables, $\bar{\mathbf{X}}$ is the row vector of population totals of the control variables, the first element of $\mathbf{x}_j$ is always one, and the first element of $\bar{\mathbf{X}}$ is one.

Following Bethlehem (2012), we get the adjusted weight w_i for observed element i for unequal probability sampling, as equal to $w_i = \nu'\mathbf{X}_i$, where ν is a vector of weight coefficients defined by

$$\nu = (\sum_{i\in s} \pi_i^{-1}\delta_i) \left(\sum_{j\in s} \pi_j^{-1}\mathbf{x}_j\mathbf{x}_j'\delta_j \right)^{-1} \bar{X}. \tag{22}$$

So, it is clear that computation of the weight does not require the individual values of the nonresponding elements. It is sufficient to have the population means of the auxiliary variables.

As an illustration, the case of one auxiliary variable with C categories is considered. Then C dummy variables $X^{(1)}$,$X^{(2)}$, ..., $X^{(C)}$ are defined. For an observation in a category H, the corresponding dummy variable is assigned the value 1, and all other dummy variables are set to 0. Consequently, the vector of population means of these dummy variables is equal to

$$\bar{X} = \left(\frac{N_1}{N}, \frac{N_2}{N}, ..., \frac{N_C}{N} \right), \tag{23}$$

where N_j is the number of elements in category j (in the population), for $j = 1, 2, ..., C$. The vector ν of weight coefficients is equal to

$$\nu = \frac{\sum_{i\in s} \pi_i^{-1}\delta_i}{N} \left(\frac{N_1}{\sum_{i\in s} \pi_i^{-1}\delta_i X^{(1)}}, \frac{N_2}{\sum_{i\in s} \pi_i^{-1}\delta_i X^{(2)}}, ..., \frac{N_C}{\sum_{i\in s} \pi_i^{-1}\delta_i X^{(C)}}, \right)'. \tag{24}$$

Now we see how the weights computed by means of generalized regression estimation can be transformed into response propensities. Let there be p categorical auxiliary variables. The continuous variables can also be transformed into categorical variables by forming several meaningful groups. The values of these variables for unit i are denoted by the vector

$$X_i = (X_i^{(1)}, X_i^{(2)}, ..., X_i^{(p)})'.$$

The number of categories of variable $X^{(j)}$ is denoted by C_j, say, for $j = 1, 2, ..., p$. So, for variable $X^{(j)}$, the categories are numbered as $1, 2, ..., C_j$.

We note from the above adjusted regression weight formula that all responding units with the same set of values for the auxiliary variables will be assigned the same weight. Suppose a unit is in category number k_1 of the first variable, category k_2 of the second variable,..., and category k_p of the p^{th} variable. Let $w(k_1, k_2, ..., k_p)$ denote the corresponding weight. Furthermore, we assume that there are $r(k_1, k_2, ..., k_p)$ respondents in this group. The number of sample units

$n(k_1, k_2, ..., k_p)$ in the group can now be estimated by

$$\hat{n}(k_1, k_2, ..., k_p) = \frac{\sum_{i \in s} \pi_i^{-1}}{\sum_{i \in s} \pi_i^{-1} \delta_i} \times w(k_1, k_2, ..., k_p) \times r(k_1, k_2, ..., k_p). \qquad (25)$$

The response propensity for all elements in the group can be estimated by

$$\hat{\rho}(k_1, k_2, ..., k_p) = \frac{r(k_1, k_2, ..., k_p)}{\hat{n}(k_1, k_2, ..., k_p)} = \frac{\sum_{i \in s} \pi_i^{-1} \delta_i}{\sum_{i \in s} \pi_i^{-1}} \times \frac{1}{w(k_1, k_2, ..., k_p)}. \qquad (26)$$

So, it is clear that the response propensities are inversely proportional to the weights. We note that the response propensities can only be estimated for respondents and not for nonrespondents.

Following Chaudhuri (2010), we can now obtain several competitive estimators and variance estimators for population mean of a variable of interest by replacing the response probabilities δ_i^* with their estimates $\hat{\delta}_i^*$ obtained by whatever means using MCAR or the logit/probit models or the $\hat{\rho}(k_1, k_2, ..., k_p)$s of Bethlehem model in the respective equations shown in Section 2.

4. Illustrative simulation based findings

In this section, we present the results of numerical comparison of our different estimators based on sample drawn using unequal probability sampling scheme. To perform the comparison simulation, we use the data of a real population. The population considered is the Labor Force Population obtained from the September 1976 Current Population Survey (CPS) conducted in the United States and this data set was studied by Valliant et al. (2000). This population data contains information on demographic and economic variables from the persons chosen in that labor force survey. This is basically a clustered population of individuals, where the clusters are compact geographic areas used as one of the stages of sampling in the CPS and are typically composed of about four nearby households. The units within clusters for this illustrative population are individual persons. For our numerical illustration, we use all of the observations of one stratum containing information of $N = 210$ persons. This data set contains information of persons about their usual number of hours of working per week, usual amount of their weekly wages along with their demographic and social charateristics like their age, sex, race (non-black, black). We consider the usual amount of their weekly wages as the main variable y of interest and the usual number of hours of working per week as the size measure variable x for drawing sample of persons. Our objective is to estimate the average weekly wage taking into account the doubtful missing information obtained from the selected respondents chosen by varying probability sampling scheme and to study the performance behaviour of alternative estimators. We use the logistic model as $\phi(x_i) = \frac{1}{1+e^{(-1.65+.5 \times race_i + .08 \times sex_i + 0.05 \times age_i)}}$ to generate the true probabilities δ_i^* s.

4.1. Application in two specific unequal probability sampling schemes

For illustration in practical sample survey situation, we consider two different unequal probability sampling schemes. The first one is Midzuno's (1952) scheme and the second one is a modification of Brewer's (1963) scheme. The choice of these two different types of sampling schemes is based on the knowledge of having a constant effective sample size and uniformly non-negative variance estimator for Midzuno's scheme and the knowledge of having varying effective sample size and uniformly non-negative variance estimator for the modified Brewer's scheme. We now describe briefly these two sampling schemes.

4.1.1. Midzuno's scheme

Midzuno (1952) suggested this scheme first by drawing one unit by probability proportional to the size measure of an auxiliary variable with known $x_i > 0$, for $i = 1, 2, \ldots, N$. Then, keeping the selected unit aside, the remaining $(n-1)$ units should be chosen by simple random sampling without replacement (SRSWOR) out of $(N-1)$ units. Let $X = \sum_{i=1}^{N} x_i$. Then, under this scheme,

$$\pi_i = \frac{x_i}{X} + \frac{X - x_i}{X}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} = \frac{x_i}{X}\frac{N-n}{N-1} + \frac{n-1}{N-1} \;\; \forall i = 1, 2, \ldots, N, \tag{27}$$

and

$$\begin{aligned}\pi_{ij} &= \frac{x_i}{X}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{x_j}{X}\frac{\binom{N-2}{n-2}}{\binom{N-1}{n-1}} + \frac{X - x_i - x_j}{X}\frac{\binom{N-3}{n-3}}{\binom{N-1}{n-1}} \\ &= \frac{x_i + x_j}{X}\frac{(N-n)(n-1)}{(N-1)(N-2)} + \frac{(n-1)(n-2)}{(N-1)(N-2)},\end{aligned} \tag{28}$$

$\forall i \neq j \in U$. For this scheme, $\pi_i\pi_j > \pi_{ij} \forall i \neq j \in U$, and so for the Horvitz and Thompson (1952)'s estimator $\sum_{i\in s} \frac{y_i}{\pi_i}$ for population total Y of y variable, the Yates and Grundy (1953) form of variance estimator
$V_{YG} = \sum_{i\in s}\sum_{j\in s, j>i} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$ is always non-negative.

Now, keeping in mind that all the y_i's may not be available for all $i \in s$, so with respect to the response probabilities δ_i^*, an unbiased estimator for population mean is

$$e_M = \frac{1}{N}\sum_{i\in s} \frac{y_i}{\pi_i}\left(\frac{\delta_i}{\delta_i^*}\right) = \frac{1}{N}\sum_{i\in s}\frac{u_i}{\pi_i}, \text{ where } u_i = y_i\frac{\delta_i}{\delta_i^*}, \tag{29}$$

since $E_R(u_i) = y_i$ and $E_P E_R(e_M) = E_P(\frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i}) = \bar{Y}$.

The variance of e_M is obtained as

$$V(e_M) = V_P E_R(e_M) + E_P V_R(e_M) = \frac{1}{N^2}\left[V_P\left(\sum_{i\in s}\frac{y_i}{\pi_i}\right) + E_P\left(\sum_{i\in s}\frac{1}{\pi_i^2}V_R(u_i)\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + E_P\left(\sum_{i\in s}\frac{1}{\pi_i^2}\frac{y_i^2(1-\delta_{ik}^*)}{\delta_i^*}\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N}\frac{1}{\pi_i}\frac{y_i^2(1-\delta_i^*)}{\delta_i^*}\right]. \quad (30)$$

Following Chaudhuri, Adhikary and Dihidar (2000), an unbiased estimator of the variance of e_M is :

$$v(e_M) = \frac{1}{N^2}\left[\sum_{i\in s}\sum_{j\in s,j>i}\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right)^2 + \sum_{i\in s}\frac{v_R(u_i)}{\pi_i}\right]. \quad (31)$$

Next, as δ_i^* is unknown to us, so following Chaudhuri (2010), we can have the required estimators and variance estimators of the population mean as $\hat{e}_M = e_M|\{\delta_i^* = \hat{\delta}_i^*\}$ and $v(e_M)|\{\delta_i^* = \hat{\delta}_i^*\}$, i.e. replacing δ_i^* in e_M and $v(e_M)$ throughout by its estimate $\hat{\delta}_i^*$ obtained by any means as discussed earlier.

4.1.2. Modified Brewer's scheme

We consider the following scheme of Brewer (1963), modified by Seth (1966) and further modified by Chaudhuri and Pal (2002). Let us call the normed size measues of auxiliary variable as $p_i = \frac{x_i}{X}$'s for $i = 1, 2, \ldots, N$. In this scheme, on the first draw, the unit i is chosen with a probability proportional to $q_i = \dfrac{p_i(1-p_i)}{1-2p_i}$ and leaving aside the unit i so chosen, a second unit $j(\neq i)$ is chosen in the second draw from the remaining units with the probability $\dfrac{p_j}{1-p_i}$. Writing $D = \displaystyle\sum_{i=1}^{N}\frac{p_i}{1-2p_i}$, from Brewer (1963) it is known that the inclusion probability of i and that of the pair $(i, j), i \neq j$ in the sample of 2 draws are respectively

$$\pi_i(2) = 2p_i, \quad \text{and} \quad \pi_{ij}(2) = \left[\frac{2p_ip_j}{1+D}\right]\left(\frac{1}{1-2p_i} + \frac{1}{1-2p_j}\right). \quad (32)$$

It is further known that

$$\Delta_{ij}(2) = \pi_i(2)\pi_j(2) - \pi_{ij}(2) \geq 0 \;\; \forall i, j(i \neq j) \in U. \quad (33)$$

We use '2' within parenthesis to emphasize that this scheme uses 2 draws. Let the sample chosen as above be augmented by adding to the 2 distinct units so drawn as above, $(r-2)$ further distinct units from the remaining $(N-2)$ units of U by simple random sampling without replacement (SRSWOR). For such a scheme introduced by Seth (1966) admitting r distinct units in each sample, the inclusion probabilities $\pi_i(r)$ for i and $\pi_{ij}(r)$ for $(i,j)(i \neq j)$, involving r draws, are respectively

$$\pi_i(r) = \frac{1}{N-2}\left[(r-2) + (N-r)\pi_i(2)\right], \tag{34}$$

$$\pi_{ij}(r) = \pi_{ij}(2) + \left(\frac{r-2}{N-2}\right)(\pi_i(2) + \pi_j(2) - 2\pi_{ij}(2)) + \left(\frac{r-2}{N-2}\right)\left(\frac{r-3}{N-3}\right)(1 - \pi_i(2) - \pi_j(2) + \pi_{ij}(2)). \tag{35}$$

Chaudhuri and Pal (2002) modified this sampling scheme of Seth (1966) by allowing $(r-2)$ to be (1) a number $(n-2)$ to be chosen with a pre-assigned probability $w(0 < w < 1)$ and (2) a number $(n-1)$ to be chosen with the complementary probability $(1-w)$. Then, a sample s so drawn will have a size n with probability w and $(n+1)$ with probability $(1-w)$. So, the effective sample size is either n or $(n+1)$. So for this modified sampling scheme if π_i^* and π_{ij}^* denote the first and second order inclusion probabilities, then

$$\pi_i^* = w\pi_i(n) + (1-w)\pi_i(n+1), \tag{36}$$

and

$$\pi_{ij}^* = w\pi_{ij}(n) + (1-w)\pi_{ij}(n+1). \tag{37}$$

Chaudhuri and Pal (2002) also showed that $\pi_i^*\pi_j^* \geq \pi_{ij}^*, \forall i,j \in U(i \neq j)$.

Under this scheme, for the Horvitz and Thompson (1952) estimator $\sum\limits_{i\in s}\frac{y_i}{\pi_i^*}$ for population total Y of y variable, the variance estimator is given by Chaudhuri and Pal (2002)

$$v_{CP} = \sum_{i\in s}\sum_{j\in s, j>i}\frac{\pi_i^*\pi_j^* - \pi_{ij}^*}{\pi_{ij}^*}\left(\frac{y_i}{\pi_i^*} - \frac{y_j}{\pi_j^*}\right)^2 + \sum_{i\in s}\frac{\alpha_i y_i^2}{\pi_i^{*2}}, \tag{38}$$

where $\alpha_i = 1 + \frac{1}{\pi_i^*}\sum\limits_{j=1, j\neq i}^{N}\pi_{ij}^* - \sum\limits_{i=1}^{N}\pi_i^*$. They also showed that $\alpha_i > 0$ for all $i \in U$ and so v_{CP} is always non-negative.

Now, keeping in mind that all the y_i's may not be available for all $i \in s$, so with respect to the response probability δ_i^*, an unbiased estimator for population mean is

$$e_B = \frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i^*}\left(\frac{\delta_i}{\delta_i^*}\right) = \frac{1}{N}\sum_{i\in s}\frac{u_i}{\pi_i^*}, \text{ where } u_i = y_i\frac{\delta_i}{\delta_i^*}, \tag{39}$$

since $E_R(u_i) = y_i$ and $E_P E_R(e_B) = E_P(\frac{1}{N}\sum_{i\in s}\frac{y_i}{\pi_i^*}) = \bar{Y}$.

The variance of e_B is obtained as

$$V(e_B) = V_P E_R(e_B) + E_P V_R(e_B)$$

$$= \frac{1}{N^2}\left[V_P\left(\sum_{i\in s}\frac{y_i}{\pi_i^*}\right)+E_P\left(\sum_{i\in s}\frac{1}{\pi_i^{*2}}V_R(u_i)\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i^*\pi_j^*-\pi_{ij}^*)\left(\frac{y_i}{\pi_i^*}-\frac{y_j}{\pi_j^*}\right)^2+\sum_{i=1}^{N}\frac{\alpha_i y_i^2}{\pi_i^*}+E_P\left(\sum_{i\in s}\frac{1}{\pi_i^{*2}}\frac{y_i^2(1-\delta_i^*)}{\delta_i^*}\right)\right]$$

$$= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j=1,j>i}^{N}(\pi_i^*\pi_j^*-\pi_{ij}^*)\left(\frac{y_i}{\pi_i^*}-\frac{y_j}{\pi_j^*}\right)^2+\sum_{i=1}^{N}\frac{\alpha_i y_i^2}{\pi_i^*}+\sum_{i=1}^{N}\frac{1}{\pi_i^*}\frac{y_i^2(1-\delta_i^*)}{\delta_i^*}\right]. \tag{40}$$

Following Chaudhuri, Adhikary and Dihidar (2000), an unbiased estimator of the variance of e_B is:

$$v(e_B) = \frac{1}{N^2}\left[\sum_{i\in s}\sum_{j\in s,j>i}\frac{\pi_i^*\pi_j^*-\pi_{ij}^*}{\pi_{ij}^*}\left(\frac{u_i}{\pi_i^*}-\frac{u_j}{\pi_j^*}\right)^2+\sum_{i\in s}\frac{\alpha_i u_i^2}{\pi_i^{*2}}+\sum_{i\in s}\frac{v_R(u_i)}{\pi_i^*}\right]. \tag{41}$$

Next, as δ_i^* is unknown to us, so following Chaudhuri (2010), we can have the required estimators and variance estimators of the population mean as $\hat{e}_B = e_B|\{\delta_i^* = \hat{\delta}_i^*\}$ and $v(e_B)|\{\delta_i^* = \hat{\delta}_i^*\}$ i.e. replacing δ_i^* in e_B and $v(e_B)$ throughout by its estimate $\hat{\delta}_i^*$ obtained by any means as discussed earlier.

4.2. Efficiency comparison

To get some ideas about the estimates and the measure of errors obtained in a practical sample survey situation, we perform the simulation by drawing samples of size equal to 15% of the population size by each of above mentioned sampling schemes taking the usual number of hours of working per week (x) as the size measure. Let us denote the estimators based on the two sampling designs by $\hat{e}_M$ and $\hat{e}_B$, the subscripts M and B being for Midzuno's and Modified Brewer's sampling schemes respectively. The notations used for different competitive estimators for population mean concerned are described as below.

(1) $\hat{e}_M(1)$ and $\hat{e}_B(1)$: Based on the estimate of δ_i^* as $\hat{\delta}_{i1}^* = \frac{r}{n}$, the traditional MCAR estimator.

(2) $\hat{e}_M(2)$ and $\hat{e}_B(2)$: Based on the estimate of δ_i^* as $\hat{\delta}_{i2}^*$ obtained from usual logit model.

(3) $\hat{e}_M(3)$ and $\hat{e}_B(3)$: Based on the estimate of δ_i^* as $\hat{\delta}_{i3}^*$ obtained from Bethlehem (2012) model.

(4) $\hat{e}_M(4, \lambda = ...)$ and $\hat{e}_B(4, \lambda = ...)$: Based on compromising in between traditional MCAR and Bethlehem (2012) model.

We compare the estimators using measures based on confidence intervals for the parameters they are meant to estimate. For each sampling scheme, the sampling is replicated a large number of times, say, 10000 times and the corresponding estimator is computed for each such sample. The standardized pivotal, namely, $\tau = \frac{\hat{\theta}-\theta}{\sqrt{v(\hat{\theta})}}$ is assumed to be a standard normal deviate. Then,

$$\left(\hat{\theta} - 1.96\sqrt{v(\hat{\theta})},\ \ \hat{\theta} + 1.96\sqrt{v(\hat{\theta})}\right)$$

is used as a 95% confidence interval for θ based on the estimator $\hat{\theta}$.

Two measures based on this confidence interval are often used to compare the performance of the alternative estimators. One is the ACP, i.e., the Average Coverage Percentage, which is the percent of the replicated samples for which θ is covered by the above confidence interval. The second measure is the AL, i.e., the average length, which is the length of the confidence interval (=2 × $1.96\sqrt{v(\hat{\theta})}$) averaged over all the replicates.

We consider another measure, namely the simulation coefficient of variation (or in short SimCV, say) defined by

$$SimCV(\hat{\theta}) = 100 \times \frac{\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_l - \left(\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_l\right)\right)^2}}{|\theta|},$$

where

L = the number of replications in the simulation study,

$\hat{\theta}_l$ = the value of the estimator in the l^{th} iteration $(l = 1, 2, \ldots, L)$,

θ = the value of the population parameter computed based on the whole population dataset.

As the simulation CV is not sufficient to compare the accuracies of the estimators, additionally some more values are computed. These are defined below.

(i) Simulation relative biases of the estimators given by:

$$rB(\hat{\theta}) = 100 \times \frac{\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_l - \theta}{|\theta|},$$

(ii) Simulation relative Root Mean Squared Errors given by:

$$rRMSE(\hat{\theta}) = 100 \times \frac{\frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_l - \theta\right)^2}{|\theta|},$$

(iii) Simulation relative biases of variance estimators given by:

$$rB(\hat{D^2}(\hat{\theta})) = 100 \times \frac{\frac{1}{L}\sum_{l=1}^{L}\hat{D^2}(\hat{\theta}_l) - \hat{D^2}(\hat{\theta})}{\hat{D^2}(\hat{\theta})},$$

where

$\hat{D^2}(\hat{\theta}_l)$ = the value of the variance estimator in the l^{th} iteration $(l = 1, 2, \ldots, L)$, and

$\hat{D^2}(\hat{\theta}) = \frac{1}{L}\sum_{l=1}^{L}\left(\hat{\theta}_l - \left(\frac{1}{L}\sum_{l=1}^{L}\hat{\theta}_l\right)\right)^2.$

A good estimator is the one with a high value of ACP; the closer this value is to 95%, the better the estimator. Again, with respect to AL, a good estimator should have a small value of AL. Similarly, the small values for the criteria SimCV, Simulation relative biases of the estimators, Simulation relative Root Mean Squared Errors, Simulation relative biases of variance estimators are also desirable for a good estimator. We present the results of these comparison criteria in Tables 1 to 4. Almost all the above stated criteria show good performances. More specifically, it is interesting to note that all values of the biases of the estimators are negative, and they are quite small, and the only exception is for the biases of the variance estimators. It is important to note that they are not so quite small, and this inspires us to investigate for other variance estimators in the future research. However, the overall results show that the estimator based on Bethlehem (2012) model used in unequal probability sampling scheme is a good competitor of the traditional estimators. Moreover, the compromised estimators based on the MCAR and Bethlehem (2012) model may also be tried with several compromising factors in order to achieve further improvement.

Table 1. Simulation results for alternative estimators (Midzuno's scheme)

Estimator	$ACP(\hat{\theta})$ (%)	$AL(\hat{\theta})$	$SimCV(\hat{\theta})$ (%)	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D^2}(\hat{\theta}))$
MCAR($\hat{e}_M(1)$)	95.8	195.7	14.38	-2.37	14.55	131.99
Logistic($\hat{e}_M(2)$)	97.0	240.1	14.83	-1.82	14.92	136.56
Bethlehem($\hat{e}_M(3)$)	96.3	220.8	13.13	-6.53	14.65	195.54

Table 2. Simulation results for compromised estimators (Midzuno's scheme)

Estimator	$ACP(\hat{\theta})$ (%)	$AL(\hat{\theta})$	$SimCV(\hat{\theta})$ (%)	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D^2}(\hat{\theta}))$
$\hat{e}_M(4, \lambda = 0.1)$	95.7	196.5	13.87	-4.31	14.50	134.78
$\hat{e}_M(4, \lambda = 0.2)$	95.2	190.7	13.45	-5.83	14.64	142.56
$\hat{e}_M(4, \lambda = 0.3)$	94.9	190.8	13.11	-7.00	14.84	150.96
$\hat{e}_M(4, \lambda = 0.4)$	95.7	189.2	12.82	-7.86	15.02	159.28
$\hat{e}_M(4, \lambda = 0.5)$	94.8	196.4	12.58	-8.44	15.13	168.10
$\hat{e}_M(4, \lambda = 0.6)$	94.7	199.7	12.40	-8.74	15.16	177.06
$\hat{e}_M(4, \lambda = 0.7)$	95.1	203.2	12.29	-8.76	15.08	185.81
$\hat{e}_M(4, \lambda = 0.8)$	95.4	206.7	12.28	-8.46	14.90	193.31
$\hat{e}_M(4, \lambda = 0.9)$	96.0	219.5	12.46	-7.77	14.67	197.93

Table 3. Simulation results for alternative estimators (Modified Brewer's scheme)

Estimator	$ACP(\hat{\theta})$ (%)	$AL(\hat{\theta})$	$SimCV(\hat{\theta})$ (%)	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D^2}(\hat{\theta}))$
MCAR($\hat{e}_B(1)$)	96.3	245.23	14.49	-3.14	14.81	137.28
Logistic($\hat{e}_B(2)$)	96.8	263.70	15.43	-1.60	15.50	182.50
Bethlehem($\hat{e}_B(3)$)	98.3	262.08	13.72	-6.46	15.15	219.12

Table 4. Simulation results for compromised estimators (Modified Brewer's scheme)

Estimator	$ACP(\hat{\theta})$ (%)	$AL(\hat{\theta})$	(%) (%)	$rB(\hat{\theta})$	$rRMSE(\hat{\theta})$	$rB(\hat{D^2}(\hat{\theta}))$
$\hat{e}_B(4, \lambda = 0.1)$	97.60	239.21	14.05	-4.82	14.84	145.24
$\hat{e}_B(4, \lambda = 0.2)$	97.40	237.26	13.71	-6.12	15.00	153.24
$\hat{e}_B(4, \lambda = 0.3)$	97.00	236.40	13.43	-7.11	15.19	161.64
$\hat{e}_B(4, \lambda = 0.4)$	98.40	236.53	13.22	-7.82	15.35	170.53
$\hat{e}_B(4, \lambda = 0.5)$	98.00	237.61	13.06	-8.27	15.45	179.91
$\hat{e}_B(4, \lambda = 0.6)$	98.20	239.70	12.96	-8.48	15.48	189.63
$\hat{e}_B(4, \lambda = 0.7)$	98.20	242.89	12.93	-8.43	15.43	199.41
$\hat{e}_B(4, \lambda = 0.8)$	98.20	247.39	13.00	-8.11	15.32	208.66
$\hat{e}_B(4, \lambda = 0.9)$	97.80	253.55	13.22	-7.48	15.18	216.20

5. Concluding remarks

This paper presents a general framework to estimate the population mean in the presence of auxiliary variables and non-response under the unequal probability sampling scheme. It is shown that the good competitive estimators can be obtained by estimating the response probabilities postulating good models keeping in mind that the values of the possible correlated variables may also not be available for the non-respondents. Finally, the doubtful missing data can also be profitably handled with the use of compromised estimator. Moreover, we need to examine the performance of the suggested estimators with some other estimators like Kott and Chang (2010), Chang and Kott (2008). Our research is in progress to see if the results of the proposed estimators in this paper show better performance in comparison with Kott and Chang (2010), Chang and Kott (2008) estimators.

Acknowledgement

The author is indebted to the anonymous referees for many constructive and insightful comments and suggestions which led to an improved version of the manuscript. The author gratefully acknowledges their kind attempt to draw her attention to valuable pioneering research papers in this direction and to inspire her for future research.

REFERENCES

BETHLEHEM, J. G., (2012). Using response probabilities for assessing representativity. *Statistics Netherlands, Discussion Paper.*

BETHLEHEM, J. G., KELLER, W. A., (1987). Linear weighting of sample survey data. *Journal of Official Statistics.* 3, 141-153.

BREWER, K. R. W., (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics.* 5, 5-13.

CHANG, T., KOTT, P. S., (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika.* 95, 557–571.

CHAUDHURI, A., (2010). Essentials of Survey Sampling. PHI Learning Private Limited. New Delhi.

CHAUDHURI, A., ADHIKARY, A., DIHIDAR, S., (2000). Mean square error estimation in multi-stage sampling. *Metrika.* 52, 115-131.

CHAUDHURI, A., PAL, S., (2002). Estimating proportions from unequal probability samples using randomized responses by Warner's and other devices. *Journal of the Indian Society of Agricultural Statistics.* 55(2), 174-183.

FOLSOM, R. E., (1991). Exponential and logistic weight adjustment for sampling and nonresponse error reduction. *Proceedings of Social Statistics Section, Washington, DC: American Statistical Association.* 197-202.

FULLER, W. A., LOUGHIN, M. M., BAKER, H. D., (1994). Regression weighting for the 1987-88 National Food Consumption Survey. *Survey Methodology.* 20, 75-85.

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association.* 41, 517-529.

HEITJAN, D. F., BASU, S., (1996). Distinguishing 'Missing at Random' and 'Missing Completely at Random'. *The American Statistician.* 50, 207-213.

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

KOTT, P. S., (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology.* 32, 133-142.

KOTT, P. S. CHANG, T., (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association.* 105(491), 1265-1275.

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

MIDZUNO, H., (1952). On the sampling system with probabilities proportionate to sum of sizes. *Annals of the Institute of Statistical Mathematics.* 3, 99-107.

POLITZ, A. N., SIMMONS, W. R., (1949). An Attempt to Get 'Not-at-Homes' into the Sample Without Call-Backs. *Journal of the American Statistical Association.* 44, 9-31.

POLITZ, A., SIMMONS, W., (1950). Note on an Attempt to Get 'Not-at-Homes' into the Sample Without Call-Backs. *Journal of the American Statistical Association.* 45, 136-137.

RAJ, D., (1966). Some remarks on a simple procedure of sampling without replacement. *Journal of the American Statistical Association.* 61, 391-396.

RUBIN, D. B., (1987). Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

RUBIN, D. B., (1976). Inference and missing data. *Biometrika.* 63, 581-592.

SÄRNDAL, C. E., (2011). Dealing with survey nonresponse in data collection, in estimation. *Journal of Official Statistics.* 27, 1-21.

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

SETH, G. R., (1966). On estimators of variance of estimate of population total in varying probabilities. *Journal of the Indian Society of Agricultural Statistics.* 18, 52-56.

SINGH, S., (2010). Layman's understanding of non-response: How Michael and Amy adjust a missing phone call. *LIAISON, Statistical Society of Canada.* 24(3), p. 67.

VALLIANT, R., DORFMAN, A. H., ROYALL, R. M., (2000). Finite Population Sampling and Inference: A Prediction Approach. Wiley Series in Survey Methodology. New York.

YATES, F., GRUNDY, P. M., (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the American Statistical Association.* 75, 206-211.

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 389–402

A CLASS OF TWO PHASE SAMPLING ESTIMATORS FOR RATIO OF TWO POPULATION MEANS USING MULTI-AUXILIARY CHARACTERS IN THE PRESENCE OF NON-RESPONSE

B. B. Khare[1], **R. R. Sinha**[2]

ABSTRACT

In this paper, a class of two phase sampling estimators for estimating the ratio of two population means using multi-auxiliary characters with unknown population means has been proposed in presence of non-response. The asymptotic bias, mean square error and minimum mean square error of the proposed class of estimators have been obtained. The optimum values of the sample at the first and the second phases along with the sub-sampling fraction of the non-responding group have been determined for the fixed cost and for the specified precision. The efficiency of the proposed class of estimators has also been shown through the theoretical and empirical studies.

Key words: two phase sampling, ratio of two means, bias, mean square error, auxiliary characters.

1. Introduction

The estimation of the ratio of two population means with known population mean of auxiliary character(s) has been discussed by Hartley and Ross (1954), Singh (1965), Tripathi (1970), Tripathi and Chaurvedi (1979) and Khare (1991). It has been well known that the ratio, product and regression types of estimators are used to increase the efficiency of the estimates when population mean of the auxiliary character is known in advance. But sometimes it has been observed in sample surveys that the population means of available auxiliary characters are not known in advance [sea Rao (1990)], in this condition it is customary to use two phase sampling for estimating the population means of the auxiliary characters. By introducing the two phase sampling scheme, Tripathi (1970), Singh (1982)

[1] Department of Statistics, Banaras Hindu University, Varanasi, India.
E-mail: bbkhare56@yahoo.com.

[2] Department of Mathematics, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India. E-mail: raghawraman@gmail.com.

and Khare (1983, 91) have proposed the estimators for estimating the ratio of two population means $R = \bar{Y}_1/\bar{Y}_2$ using an auxiliary character with unknown population mean. Using two auxiliary characters with unknown population means, the estimators for estimating R have been proposed by Tripathi and Sinha (1976) and Srivasvata *et al.* (1988). Further Khare (1993) has proposed a class of estimators for R by using multi-auxiliary characters with unknown population means when the information is available on all selected units in the sample for main and auxiliary characters. But it has been observed in practice while conducting a sample survey related to human that we do not collect complete information for all the units selected in the sample due to the problem of non-response on s tudy characters. Khare and Sinha (2002, 2004) have proposed classes of two phase sampling estimators for estimating the ratio of two population means using auxiliary character in presence of non-response while Khare and

Sinha (2012) have suggested the general classes of estimators using multi-auxiliary characters with subsampling the non-respondents.

In this paper, we have proposed a class of two phase sampling estimators for estimating the ratio of two population means $R = \bar{Y}_1/\bar{Y}_2$ of the study characters in presence of non-response using multi-auxiliary characters when their population means are not known. The expressions of bias, mean square error and minimum mean square error of the proposed class of estimators have been obtained. The optimum values of first phase sample, second phase sample and sub-sampling fraction of the non-responding group have been determined for the fixed cost and for the specified precision. The efficiency of the proposed class of estimators has also been shown through theoretical and empirical studies.

2. The proposed class of estimators

Consider a finite population which consists of N identifiable units $U_N = (u_1, u_2, \dots\dots\dots, u_N)$ in which (y_1, y_2) are the variables under study and $(x_1, x_2, \dots\dots, x_p)$ are the p auxiliary characters having population means $\bar{Y}_i$ $(i = 1, 2)$ of study characters and $\bar{X}_j$ $(j = 1, 2, \dots\dots, p)$ of auxiliary characters respectively. In many practical situations when the list of the sampling units is available but the population means of the auxiliary characters are not known then we use two phase sampling scheme to estimate the unknown population means of the auxiliary characters. In such situations, the estimate of population mean $\bar{X}_j$ $(j = 1, 2, \dots\dots, p)$ is furnished by taking a large first phase sample of size n' from the population of N units using simple random sampling without replacement (SRSWOR) method. Let the estimate of $\bar{X}_j$ $(j = 1, 2, \dots\dots, p)$ be the sample means $\bar{x}_j'$ $(j = 1, 2, \dots\dots, p)$ based on the information available on n' units. Again a seco nd phase sample of size $n\,(< n')$ is drawn from the first phase selected units n' by SRSWOR method of sampling and collect the information on the study characters y_i $(i = 1,2)$. We observe for the study characters y_i $(i = 1,2)$ that only n_1 units are responding and $n_2(= n - n_1)$ units are not responding in the sample of size n. In this case, it has been assumed that the whole population U_N is divided into two non-overlapping strata U_{N_1} and U_{N_2} of responding and non-responding soft-core groups; however they are not known in advance. The stratum weights of responding and non-responding groups are given by $P_1 = N_1/N$ and $P_2 = N_2/N$, and their estimates are respectively given by $\hat{P}_1 = p_1 = n_1/n$ and $\hat{P}_2 = p_2 = n_2/n$. Further, from the non-responding units n_2, we draw a su bsample of size $r\,(= n_2\,k^{-1},\ k > 1)$ using SRSWOR technique of sampling and collect the information by the direct interview for yi $(i = 1,2)$. Now using the approach of Hansen and Hurwitz (1946), the unbiased estimator for $\bar{Y}_i$ $(i = 1,2)$ based on the information of $(n_1 + r)$ units is given by

$$\bar{y}_i^* = p_1\bar{y}_{i1} + p_2\bar{y}_{in_{(2r)}}, \quad i = 1,2 \tag{2.1}$$

where $\bar{y}_{i1}$ and $\bar{y}_{in_{(2r)}}$ are the sample means of y_i based on n_1 and r units respectively.

The variance of the estimator $\bar{y}_i^*$ up to the terms of order (n^{-1}) is given by

$$V(\bar{y}_i^*) = V_i = \theta\, S_{y_i}^2 + \theta_k\, S_{y_{i(2)}}^2, \tag{2.2}$$

where $S_{y_i}^2$ and $S_{y_{i(2)}}^2$ denote the population mean square of y_i for the entire and non-responding part of the population, and $\theta = \frac{N-n}{Nn}$, $\theta_k = \frac{P_2(k-1)}{n}$.

If the ratio of two population means is $R = \bar{Y}_1/\bar{Y}_2$ and we have incomplete information on t he study characters (y_1, y_2), then the usual estimator for estimating R may be given by

$$\hat{R} = \frac{\bar{y}_1^*}{\bar{y}_2^*} \ . \tag{2.3}$$

The bias and mean square error of $\hat{R}$ under SRSWOR up to the terms of order (n^{-1}) are given by

$$B(\hat{R}) = R\{\theta\ \nabla_{21} +\ \theta_k\nabla_{21}'\}, \tag{2.4}$$

$$M(\hat{R}) = R^2\{\theta\,\Delta_{12} + \theta_k\Delta_{12}'\}, \tag{2.5}$$

where $\nabla_{21} = \frac{S_{y_2}^2}{\bar{Y}_2^2} - \rho\frac{S_{y_1}}{\bar{Y}_1}\frac{S_{y_2}}{\bar{Y}_2}$, $\nabla_{21}' = \frac{S_{y_{2(2)}}^2}{\bar{Y}_2^2} - \rho_2\frac{S_{y_{1(2)}}}{\bar{Y}_1}\frac{S_{y_{2(2)}}}{\bar{Y}_2}$, $\Delta_{12} = \frac{S_{y_1}^2}{\bar{Y}_1^2} + \frac{S_{y_2}^2}{\bar{Y}_2^2} - 2\rho\frac{S_{y_1}}{\bar{Y}_1}\frac{S_{y_2}}{\bar{Y}_2}$, $\Delta_{12}' = \frac{S_{y_{1(2)}}^2}{\bar{Y}_1^2} + \frac{S_{y_{2(2)}}^2}{\bar{Y}_2^2} - 2\rho_2\frac{S_{y_{1(2)}}}{\bar{Y}_1}\frac{S_{y_{2(2)}}}{\bar{Y}_2}$, ρ and ρ_2 are the correlation coefficients between (y_1, y_2) for the entire and non-responding group of the population respectively.

Hence, when we have incomplete information on the study characters y_1, y_2 but complete information on the auxiliary characters $x_1, x_2, \ldots, x_p$ for the sample of size n [See Rao (1986) p. 220], we propose a class of two phase sampling estimators for estimating the ratio of two population means $R(=\bar{Y}_1/\bar{Y}_2)$ of study characters using multi-auxiliary characters in presence of non-response on study characters only as

$$T = f(\bar{y}_1^*/\bar{y}_2^*,\ \underline{z}') = f(m,\ \underline{z}') \tag{2.6}$$

such that $f(R,\ \underline{e}') = R$, and $f_{1(R,\ \underline{e}')} = \left(\frac{\partial}{\partial m} f(m,\ \underline{z}')\right)_{(R,\ \underline{e}')} = 1,$ (2.7)

where $\underline{z}$ and $\underline{e}$ are the column vectors of $(z_1, z_2, \dots\dots, z_p)'$ and $(1, 1, \dots\dots, 1)'$ respectively and $z_j = \frac{\bar{x}_j}{\bar{x}_j'}$, $(j = 1,2, \dots, p)$. Here we assume that the function $f(m,\ \underline{z}')$ is continuous and bounded in $(p+1)$ dimensional real space S^* containing the point $(R,\ \underline{e}')$ and the first and second order partial derivatives of $f(m,\ \underline{z}')$ exist and are continuous and bounded in S^*.

3. Bias and mean square error (MSE)

Let the conventional estimator of R be $\hat{R}(=\bar{y}_1^*/\bar{y}_2^*)$. Since the number of possible samples is finite, so the bias and mean square error of the estimator T may be obtained. Now, expanding the function $f(m,\ \underline{z}')$ about the point $(R, \underline{e}')$ in a second order Taylor's series and using the condition (2.7), we have

$$T = R + D + \underline{D}' f_{2\,(R,\ \underline{e}')} + D\underline{D}' f_{12(m^*,\ \underline{z}^{*\prime})} + \frac{1}{2}\left[D^2 f_{11(m^*,\ \underline{z}^{*\prime})} + \underline{D}' f_{22(m^*,\ \underline{z}^{*\prime})} \underline{D}\right], \tag{3.1}$$

where $D = (m - R)$, $\underline{D}' = (\underline{z} - \underline{e})'$, $m^* = R + \emptyset(m - R)$, $\underline{z}^* = \underline{e} + \underline{\emptyset}\ (\underline{z} - \underline{e})$ such that $0 < \emptyset, \emptyset_j < 1$; $j = 1, 2, \dots, p$ and $\underline{\emptyset}$ is a $p \times p$ diagonal matrix having j^{th} diagonal elements ϕ_j.

Here, $f_{1(m, \underline{z}')}$ and $f_{2(m,\ \underline{z}')}$ denote the first partial derivatives of $f(m,\ \underline{z}')$ with respect to m and $\underline{z}'$ respectively. The second partial derivative of $f(m,\ \underline{z}')$ with respect to $\underline{z}'$ is denoted by $f_{22(m,\ \underline{z}')}$ and the first partial derivative of $f_{2(m,\ \underline{z}')}$ with respect to m is denoted by $f_{12(m,\ \underline{z}')}$.

The expressions for bias and mean square error of T for any sampling design up to the terms of order n^{-1} $[O(n^{-1})]$ are given by

$$B(T) = B(\hat{R}) + E(D\underline{D}') f_{12(m^*, \underline{z}^{*\prime})} + \frac{1}{2}\ E(\underline{D}' f_{22(m^*, \underline{z}^{*\prime})} \underline{D}) \tag{3.2}$$

and $$M(T) = M(\hat{R}) + 2E(D\underline{D}') f_{2(R, \underline{e}')} + E(f_{2(R, \underline{e}')})' \underline{D}\,\underline{D}' f_{2(R,\ \underline{e}')}. \tag{3.3}$$

The mean square error of T is minimized for

$$f_{2(R,\ \underline{e}')} = -[E(\underline{D}\ \underline{D}')]^{-1}E(D\underline{D}) \tag{3.4}$$

and the resulting minimum mean square error of R up to the terms of $O(n^{-1})$ is given by

$$M(T)_{min.} = M(\hat{R}) + E(D\underline{D}')[E(\underline{D}\ \underline{D}')]^{-1}E(D\underline{D}). \tag{3.5}$$

To find the bias and mean square error of T under SRSWOR, we use the large sample approximation by assuming

$\bar{y}_i^* = \bar{Y}_i(1 + \epsilon_{0i})$, $\bar{x}_j = \bar{X}_j(1 + \epsilon_j')$, $\bar{x}_j' = \bar{X}_j(1 + \epsilon_j'')$ with $E(\epsilon_{0i}) = E(\epsilon_j') = E(\epsilon_j'') = 0$ and $|\epsilon_{0i}| < 1$, $|\epsilon_j'| < 1$, $|\epsilon_j''| < 1$ $\forall\, i = 1, 2; j = 1, 2, \ldots \ldots, p$.

We also assume that the contribution of the terms involving the powers in ϵ_{0i}, ϵ_j' and ϵ_j'' of order higher than two in the bias and mean square error are assumed to be negligible.

Let $\rho_{jj'}$, ρ_{ij}^* be the correlation coefficients between $(x_j, x_{j'})$ and (y_i, x_j) respectively for the entire population and $\rho_{jj'(2)}$, $\rho_{ij(2)}^*$ be the correlation coefficients between $(x_j, x_{j'})$ and (y_i, x_j) for the non-responding group of the population.

So, the expressions of bias and mean square error of R in SRSWOR method of sampling up to the terms of $O(n^{-1})$ are given by

$$B(T) = B(\hat{R}) + R(\theta - \theta')\underline{\mathbb{B}}' f_{12(m^*, \underline{z}^{*\prime})} + \frac{(\theta - \theta')}{2}\ trace\ \underline{M}\, f_{22(m^*, \underline{z}^{*\prime})} \tag{3.6}$$

and

$$M(T) = M(\hat{R}) + (\theta - \theta')\left(f_{2(R,\ \underline{e}')}\right)' \underline{M}\, f_{2(R, \underline{e}')} + 2R\,(\theta - \theta')\,\underline{\mathbb{B}}' f_{2(m^*, \underline{z}^{*\prime})}, \tag{3.7}$$

where $\theta' = \frac{N - n'}{Nn'}$, $\underline{\mathbb{B}} = (\mathbb{B}_1, \mathbb{B}_2, \ldots \ldots, \mathbb{B}_p)'$is a column vector of order $(p \times 1)$ having the j^{th} element $\mathbb{B}_j = \frac{S_{x_j}}{\bar{X}_j}\left(\rho_{1j}^* \frac{S_{y_1}}{\bar{Y}_1} - \rho_{2j}^* \frac{S_{y_2}}{\bar{Y}_2}\right)$, $\underline{M} = [m_{jj'}]_{p \times p}$ is a $(p \times p)$ positive definite matrix having $m_{jj'} = \rho_{jj'} \frac{S_{x_j}}{\bar{X}_j} \frac{S_{x_{j'}}}{\bar{X}_{j'}}$; $\forall\, j \neq j' = 1, 2, \ldots \ldots, p$ and $S_{x_j}^2$ denotes the mean square error of x_j for the entire part of the population.

Since the objective of this paper is to suggest a generalized class of estimators $T = f(m,\ \underline{z}')$ for estimating R and study its properties, so we may consider the following exponential, chain ratio and chain ratio cum regression types of estimators as members of T, which are as follows:

$$T_e = m\, e^{\sum_{j=1}^{p} a_j log z_j}, \tag{3.8}$$

$$T_r = m\left[\omega_1 {z_1}^{b_1/\omega_1} + \omega_2 {z_2}^{b_2/\omega_2} + \cdots \dots \dots + \omega_p {z_p}^{b_p/\omega_p}\right];\ \sum_{j=1}^{p} \omega_j = 1 \qquad (3.9)$$

and $$T_{crr} = \sum_{j=1}^{p} \{m + \varphi_j (z_j - 1)\} \left(\omega_j {z_j}^{c_j/\omega_j}\right), \qquad (3.10)$$

where a_j, b_j and c_j $(j = 1, 2, \dots, p)$ are the scalar constants.

Now we state the following theorems:

Theorem 1. *Up to the terms of order $O(n^{-1})$ under SRSWOR, the mean square error of T is minimized for*

$$f_{2(R,\underline{e}')} = -R\underline{M}^{-1}\underline{\mathbb{B}} \qquad (3.11)$$

and minimum mean square error of T is given by

$$M(T)_{min.} = R^2\{(\theta\, \Delta_{12} + \theta_k \Delta'_{12}) - (\theta - \theta')\left(\underline{\mathbb{B}}' \underline{M}^{-1} \underline{\mathbb{B}}\right)\}. \qquad (3.12)$$

Since the estimators T_e, T_r and T_{crr} are the members of T, so the values of the constants involved in them can be obtained by the condition (3.11) and their minimum mean square error will be equal to $M(T)_{min.}$. Sometimes this condition involves unknown parameters, so one may use the values of the parameters from past data or experience for obtaining the required value of the constants involved in (3.11). Reddy (1978) has shown that such values are stable not only over time but also over different regions. Srivastava and Jhajj (1983) have shown that the efficiency of such type of estimators does not decrease up to the terms of order $O(n^{-1})$ if we replace the optimum values of the constants by their estimates based on the sample values.

On comparing the proposed class of estimator T with $\hat{R}$ in terms of precision from (2.5) and (3.12), we have derived the following theorem:

Theorem 2. *Up to the terms of order $O(n^{-1})$,*

$$M(T) < M(\hat{R}) \text{ and } M(\hat{R}) - M(T) = R^2\{(\theta - \theta')\left(\underline{\mathbb{B}}' \underline{M}^{-1} \underline{\mathbb{B}}\right)\} > 0.$$

Theorem 3. *Up to the terms of order $O(n^{-1})$,*

$$M(T) < M(\hat{R}) \text{ iff } -M(\hat{R}) < \left\{(\theta - \theta')\left(f_{2(R,\ \underline{e}')}\right)\right\}\left\{\left(f_{2(R,\ \underline{e}')}\right)' \underline{M} + 2R\underline{\mathbb{B}}'\right\} < 0.$$

If we compare the efficiency of proposed class of estimators (T) with the class of estimators suggested by Khare and Sinha (2012), we find that the Khare and Sinha (2012) estimator gives equal precision to T under the condition of known population mean of auxiliary characters.

It is also to be noted here that for $W_2 = 0$, i .e. when we have complete information on the study characters as well as on auxiliary characters for the sample of size n, then the proposed class of estimators T is equally efficient to the class of estimators for R as proposed by Khare (1993). Hence, it is clear that all the members of the proposed class of estimators T will attain minimum mean square error for one, two or p-auxiliary characters if the condition (3.11) is satisfied.

It is very important to know whether the reduction in variance would be worth the extra expenditure on the additional sample required to estimate the population mean of the auxiliary characters used in the case of two phase sampling. Hence, a rational approach is found by minimizing the mean square error of T for the fixed cost and obtaining the optimum values of n', n and k. Therefore, we determine the size of the first phase sample (n'), second phase sample (n) and the value of subsampling proportion (k^{-1}) which will minimize the mean square error of the proposed class of estimators T for the fixed cost $C \leq \boldsymbol{C_0}$.

4. Optimum sample size for the fixed cost C≤ C₀

The minimum value of the mean square error of T depends upon the values of n' n and k. Let the fixed total cost apart from overhead cost be $C \leq \boldsymbol{C_0}$. Let C_1' and C_1 are be the cost per unit of identifying and observing auxiliary characters and the cost per unit of mailing questionnaire/visiting the unit at the second phase respectively while C_2 and C_3 be the cost per unit of collecting/processing data for the study characters y_1, y_2 obtained from n_1 responding units and the cost per unit of obtaining and processing data for the study characters y_1, y_2 (after extra efforts) from the subsampled units. Now, the cost function under these assumptions is given by

$$C' = C_1'n' + C_1n + C_2n_1 + C_3r. \tag{4.1}$$

Since C' will vary from sample to sample, so we consider the expected cost C to be incurred in the survey apart from overhead expenses, which is given by

$$C = E(C') = C_1'n' + n[C_1 + C_2P_1 + C_3P_2k^{-1}]. \tag{4.2}$$

Let $R^2\Psi_{0r}$, $R^2\Psi_{1r}$ and $R^2\Psi_{2r}$ be the coefficients of the terms n^{-1}, $(n')^{-1}$ and kn^{-1} respectively in the expressions of $M(T)$, then $M(T)$ can be expressed as

$$M(T) = (n^{-1})R^2\Psi_{0r} + (n')^{-1}R^2\Psi_{1r} + (kn^{-1})R^2\Psi_{2r} + I, \tag{4.3}$$

where I is the terms independent of n, n' and k in the expressions of $M(T)$.

Now, let us define a function φ for minimizing the $M(T)$ for the fixed cost $C \leq \boldsymbol{C_0}$ and to obtain the optimum sample sizes as

$$\varphi = M(T) + \lambda_r\{C_1'n' + n(C_1 + C_2P_1 + C_3P_2k^{-1}) - \boldsymbol{C_0}\}, \tag{4.4}$$

where λ_r is a Lagrange's multiplier.

Differentiating φ with respect to n', n and k and equating to zero, we have

$$n' = R\sqrt{\frac{\Psi_{1r}}{\lambda_r C_1'}}, \tag{4.5}$$

$$n = R\sqrt{\frac{\Psi_{0r}+k\Psi_{2r}}{\lambda_r(C_1+C_2P_1+C_3P_2k^{-1})}} \tag{4.6}$$

and $$k_{opt.} = R\sqrt{\frac{C_3P_2\Psi_{0r}}{(C_1+C_2P_1)\ \Psi_{2r}}}\ . \tag{4.7}$$

Now, putting the values of $n'$and n from (4.5) and (4.6) and using the value of $k_{opt.}$ from (4.7) in (4.2), we have

$$\sqrt{\lambda_r} = \frac{R}{\boldsymbol{C_0}}\left[\sqrt{\Psi_{1r}C_1'} + \sqrt{\left(\Psi_{0r} + k_{opt.}\Psi_{2r}\right)\left(C_1 + C_2P_1 + C_3P_2k_{opt.}^{-1}\right)}\right]. \tag{4.8}$$

It has also been observed that the determinant of the matrix of the second order derivative of φ with respect to n', n and k is positive for the optimum values of n', n and k, which shows that the solutions for n', n given by (4.5), (4.6) and the optimum value of k under the condition $C \leq \boldsymbol{C_0}$ minimize the variance of T. It is also important to note here that the subsampling fraction $k_{opt.}^{-1}$ will decrease as $\sqrt{C_3/(C_1 + C_2P_1)}$ increases.

Hence, for the optimum values of n', n and k, the minimum value of $M(T)$ is given by

$$M(T)_{min.} = \boldsymbol{C_0}\lambda_r - R^2\Delta_{12}N^{-1}. \tag{4.9}$$

5. Determination of sample sizes for the specified variance M_0

Let $\boldsymbol{M_0}$ be the variance of the estimator T fixed in advance and we have

$$\boldsymbol{M_0} = (n^{-1})R^2\Psi_{0r} + (n')^{-1}R^2\Psi_{1r} + (kn^{-1})R^2\Psi_{2r} + R^2\Delta_{12}N^{-1}. \tag{5.1}$$

For minimizing the average total cost C for the specified variance of the estimator T (i.e. $M(T) = \boldsymbol{M_0}$), we define a function φ^* which is given as

$$\varphi^* = C_1'n' + n(C_1 + C_2P_1 + C_3P_2k^{-1}) - \mu(M(T) - M_0) \tag{5.2}$$

where μ is a Lagrange's multiplier.

Now, for obtaining the optimum values of n', n and k, differentiating φ^* with respect to n', n and k and equating to zero, we have

$$n' = R\sqrt{\frac{\mu\Psi_{1r}}{C_1'}} \tag{5.3}$$

$$n = R\sqrt{\frac{\mu(\Psi_{0r}+k\Psi_{2r})}{(C_1+C_2P_1+C_3P_2k^{-1})}} \tag{5.4}$$

and
$$k_{opt.} = \sqrt{\frac{C_3P_2\Psi_{0r}}{(C_1+C_2P_1)\,\Psi_{2r}}}. \tag{5.5}$$

Again by putting the values of n' and n from (5.3) and (5.4) and utilizing the optimum value of k in (5.1), we get

$$\sqrt{\mu} = \frac{\left[\sqrt{\Psi_{1r}C_1'}+\sqrt{(\Psi_{0r}+k_{opt.}\Psi_{2r})(C_1+C_2P_1+C_3P_2k_{opt.}^{-1})}\right]}{[\boldsymbol{M_0}+R^2\Delta_{12}N^{-1}]}. \tag{5.6}$$

The minimum expected total cost incurred in attaining the specified variance $\boldsymbol{M_0}$ by the estimator T is then given by

$$C(T)_{min.} = \frac{\left[\sqrt{C_1'V_{11}}+\sqrt{(V_{01}+k_{opt.}V_{21})\left(C_1+C_2W_1+C_3\frac{W_2}{k_{opt.}}\right)}\right]^2}{[\boldsymbol{M_0}+R^2\Delta_{12}N^{-1}]}. \tag{5.7}$$

6. An empirical study

109 Village/Town/ward population of urban area under Police-station – Baria, Tahasil – Champua, Orissa has been taken under consideration from District Census Handbook, 1981, O rissa, published by Govt. of India. The last 25% villages (i.e. 27 villages) have been considered as n on-response group of the population. Here we have considered the study characters and auxiliary characters given as follows:

y_1: Number of literate persons in the village,

y_2: Number of main workers in the village,

x_1: Number of non-workers in the village,

x_2: Total population of the village and

x_3: Number of cultivators in the village.

The values of the parameters of the population under study are as follows:

$\bar{Y}_1 = 145.3028$	$\bar{Y}_2 = 165.2661$	$\bar{X}_1 = 259.0826$	$\bar{X}_2 = 485.9174$	$\bar{X}_3 = 100.5505$
$S_{y_1} = 111.3891$	$S_{y_2} = 112.8437$	$S_{x_1} = 198.0687$	$S_{x_2} = 320.2197$	$S_{x_3} = 73.5426$
$S^2_{y_{1(2)}} = 100.2444$	$S^2_{y_{2(2)}} = 95.3420$	$\rho^*_{11} = 0.905$	$\rho^*_{12} = 0.905$	$\rho^*_{13} = 0.648$
$\rho = 0.816$	$\rho_2 = 0.787$	$\rho^*_{21} = 0.819$	$\rho^*_{22} = 0.908$	$\rho^*_{23} = 0.841$
	$\rho_{12} = 0.946$	$\rho_{13} = 0.732$	$\rho_{23} = 0.801$	

Let the costs at the different processing stages be $C_1' =$ Rs. 0.15, $C_1 =$ Rs. 5.00, $C_2 =$ Rs. 25.00 and $C_3 =$ Rs. 65.00.

To show the efficiency of the proposed class of estimators T for the ratio of two population means [i.e. $R = \bar{Y}_1/\bar{Y}_2$] using the auxiliary characters x_1, x_2 and x_3, we have considered $T_e = m\, e^{\sum_{j=1}^{p} a_j log z_j}$ as a member of the proposed class of estimators T.

The optimum values of the constants a_j, mean square error and the percentage relative efficiency (PRE) of T_e with respect to $\hat{R}$ for fixed sample sizes $n' = 80$, $n = 20$ and for the fixed cost $\boldsymbol{C_0} =$ Rs. 280 are shown in Table 1. The expected cost of $\hat{R}$ and T_e in case of specified precision $\boldsymbol{M_0} = 1250 \times 10^{-5}$ are also given in Table 1.

7. Conclusions

From Table 1 – see Appendix 2, it has been observed that the estimator T_e is more efficient than $\hat{R}$ for all the different values of the sub-sampling fraction k^{-1} and its efficiency increases as the value of sub-sampling fraction increases. The mean square error of the estimator T_e decreases while the relative efficiency of the estimator T_e with respect to $\hat{R}$ increases with the increase in the numbers of auxiliary characters used. Regarding the performance of the estimator T_e over $\hat{R}$ in case of fixed cost, we observe that the relative efficiency of T_e increases as the number of the auxiliary characters increases. We also observe that the values of $k_{opt.}$ and $n_{opt.}$ decrease while the value of $n'_{opt.}$ increases with the increase in the numbers of auxiliary characters used. Further, in case of specified variance, the expected cost incurred by T_e decreases with the increases in the numbers of auxiliary characters used. It has been also observed that $n'_{opt.}$increases while $n_{opt.}$ decreases by increasing the numbers of the auxiliary characters. Hence, on the basis of theoretical and empirical studies, we may recommend the proposed class of estimators T for the use in practice under its respective circumstances as discussed in the text.

Acknowledgements

Authors are grateful to the referee and the editor for their invaluable suggestions which helped in further improvement in the paper.

REFERENCES

HANSEN, M. H., HURWITZ, W. N., (1946). The problem of nonresponse in sample surveys, Jour. Amer. Statist. Assoc., 41, 517–529.

HARTLEY, H. O., ROSS, A., (1954). Unbiased ratio estimators, Nature, 174, 270–271.

KHARE, B. B., (1983). Some problems of estimation using auxiliary character, Ph.D. Thesis submitted to B.H.U., Varanasi, India.

KHARE, B. B., (1991). Determination of sample sizes for a class of two phase sampling estimators for ratio and product of two population means using auxiliary character, Metron, 49(1-4), 185–197.

KHARE, B. B., (1993). On a class of two phase sampling estimators for ratio of two population means using multi-auxiliary characters, Proc. Nat. Acad. Sci. India, 63(A) III, 513–519.

KHARE, B. B., SINHA, R. R., (2002). Estimation of the ratio of two population means using auxiliary character with unknown population mean in presence of non-response, Progress of Mathematic, 36, 337–348.

KHARE, B. B., SINHA, R. R., (2004). Estimation of finite population ratio using two phase sampling scheme in the presence of non-response, Aligarh J. Stat, 2004, 24, 43–56.

KHARE, B. B., SINHA, R. R., (2012). Improved classes of estimators for ratio of two means with double sampling the non respondents, Statistika, 49(3), 75–83.

REDDY, V. N., (1978). A study of use of prior knowledge on certain population parameters in estimation, Sankhaya, C, 40, 29–37.

RAO, P. S. R. S., (1986). Ratio estimation with subsampling the nonrespondents, Survey Methodology, 12(2), 217–230.

RAO, P. S. R. S., (1990). Regression estimators with subsampling of nonrespondents, In-Data Quality Control, Theory and Pragmatics, (Eds.) Gunar E. Liepins and V.R.R. Uppuluri, Marcel Dekker, New York, (1990), 191–208.

SINGH, M. P., (1965). On the estimation of ratio and product of population parameters, Sankhya, Ser.C, 27, 321–328.

SINGH, R. K., (1982). Generalized double sampling estimators for the ratio and product of population parameters, Jour. Ind. Statist. Assoc., 20, 39–49.

SRIVASTAVA, S. R., KHARE, B. B., SRIVASTAVA, S. R., (1988). On generalised chain estimator for ratio and product of two population means using auxiliary characters, Assam. Stat. Rev., 1, 21–29.

SRIVASTAVA, S. K., JHAJJ, H. S., (1983). A class of estimators of the population mean using multi-auxiliary information, Cal. Stat. Assoc. Bull., 32, 47–56.

TRIPATHI, T. P., (1970). Contribution to the sampling theory using multivariate information, Ph.D. Thesis submitted to Punjabi University, Patiyala, India.

TRIPATHI, T. P., CHATURVEDI, D. K., (1979). Use of multivariate auxiliary information in estimating the population ratio, Stat-Math. Tech. Report No. 24/79, I.S.I., Calcutta, India, Abs. J. Indian Soc. Agricultural Statist., bf 31.

TRIPATHI, T. P., SINHA, S. K. P., (1976). Estimation of ratio on successive occasions, Proceedings of Symposium on Recent Developments in Survey Methodology, held at I.S.I., Calcutta, March 1976.

APPENDIX 1

Expand the function $f\left(m,\ \underline{z}'\right)$ given in (2.6) about the point $\left(R,\ \underline{e}'\right)$ using Taylor's series up to the second order partial derivatives, we have

$$\begin{aligned} T &= f\left(R,\underline{e}'\right) + Df_{1\,(R,\ \underline{e}')} + \underline{D}'f_{2\,(R,\ \underline{e}')} \\ &\quad + \tfrac{1}{2}\left[D^2 f_{11(m^*,\ \underline{z}^{*'})} + 2D\underline{D}'f_{12(m^*,\ \underline{z}^{*'})} + \underline{D}'f_{22(m^*,\ \underline{z}^{*'})}\underline{D}\right] \end{aligned}$$

Using condition (2.7), we get

$$T = R + D + \underline{D}'f_{2\,(R,\ \underline{e}')} + 2D\underline{D}'f_{12(m^*,\ \underline{z}^{*'})} + \tfrac{1}{2}\left[D^2 f_{11(m^*,\ \underline{z}^{*'})} + \underline{D}'f_{22(m^*,\ \underline{z}^{*'})}\underline{D}\right]$$

Now,

$$\begin{aligned} B(T) &= E(T-R) \\ &= B\left(\hat{R}\right) + E\left(D\underline{D}'\right)f_{12(m^*,\underline{z}^{*\prime})} + \tfrac{1}{2}\ E\left(\underline{D}'f_{22(m^*,\underline{z}^{*\prime})}\underline{D}\right) \\ M(T) &= E(T-R)^2 \\ &= M\left(\hat{R}\right) + 2E\left(D\underline{D}'\right)f_{2(R,\underline{e}')} + E(f_{2(R,\underline{e}')})'\underline{D}\,\underline{D}'f_{2(R,\ \underline{e}')} \end{aligned}$$

Differentiating $M(T)$ with respect to $f_{2(R,\underline{e}')}$ and equating it to zero, we have

$$f_{2(R,\ \underline{e}')} = -\ [E\left(\underline{D}\,\underline{D}'\right)]^{-1}E\left(D\underline{D}\right)$$

Putting this $f_{2(R,\ \underline{e}')}$ in $M(T)$, we get

$$M(T)_{min.} = M\left(\hat{R}\right) + E\left(D\underline{D}'\right)[E\left(\underline{D}\,\underline{D}'\right)]^{-1}E\left(D\underline{D}\right).$$

Under simple random sampling without replacement (SRSWOR), we have obtained

$$E\left(\underline{D}\ \underline{D}'\right) = E\left[\left(\underline{z} - \underline{e}\right)\left(\underline{z} - \underline{e}\right)'\right]$$

Consider

$$E[(z_1 - 1)(z_2 - 1)] = E\left[\left(\frac{\bar{x}_1}{\bar{x}_1^{'}} - 1\right)\left(\frac{\bar{x}_2}{\bar{x}_2^{'}} - 1\right)\right]$$

$$= E\left[\left(\frac{\bar{X}_1(1+\epsilon_1^{'})}{\bar{\bar{X}}_1(1+\epsilon_1^{''})} - 1\right)\left(\frac{\bar{X}_2(1+\epsilon_2^{'})}{\bar{\bar{X}}_2(1+\epsilon_2^{''})} - 1\right)\right]$$

$$= E\left[\left\{\left(1 +\epsilon_1^{'}\right)\left(1 +\epsilon_1^{''}\right)^{-1} - 1\right\}\left\{\left(1 +\epsilon_2^{'}\right)\left(1 +\epsilon_2^{''}\right)^{-1} - 1\right\}\right]$$

Neglecting the terms involving powers in $\epsilon_j^{'}, \epsilon_j^{''}; j = 1, 2$ of order higher than two, we have

$$E[(z_1 - 1)(z_2 - 1)] = E\left[\epsilon_1^{''}\epsilon_2^{''} - \epsilon_1^{''}\epsilon_2^{'} - \epsilon_1^{'}\epsilon_2^{''} + \epsilon_1^{'}\epsilon_2^{'}\right]$$

Since $\quad E[\epsilon_1''\epsilon_2''] = E[\epsilon_1''\epsilon_2'] = (\theta - \theta')\rho_{12}\frac{S_{x_1}}{\bar{\bar{X}}_1}\frac{S_{x_2}}{\bar{\bar{X}}_2}$

Therefore, $E[(z_1 - 1)(z_2 - 1)] = E[\epsilon_1'\epsilon_2'] - E[\epsilon_1'\epsilon_2'']$

$$= \theta\rho_{12}\frac{S_{x_1}}{\bar{\bar{X}}_1}\frac{S_{x_2}}{\bar{\bar{X}}_2} - \theta'\rho_{12}\frac{S_{x_1}}{\bar{\bar{X}}_1}\frac{S_{x_2}}{\bar{\bar{X}}_2}$$

$$= (\theta - \theta')\rho_{12}\frac{S_{x_1}}{\bar{\bar{X}}_1}\frac{S_{x_2}}{\bar{\bar{X}}_2} = (\theta - \theta')m_{12}$$

Similarly, we can define $\quad m_{jj'} = \rho_{jj'}\frac{S_{x_j}}{\bar{\bar{X}}_j}\frac{S_{x_{j'}}}{\bar{\bar{X}}_{j'}}.$

Now, $\quad E\left(D\underline{D}'\right) = E(m - R)\left(\underline{z} - \underline{e}\right)$

Consider

$$E(m - R)(z_1 - 1) = E\left[\left\{\frac{\bar{y}_1^*}{\bar{y}_2^*} - R\right\}\left\{\frac{\bar{x}_1}{\bar{x}_1^{'}} - 1\right\}\right]$$

$$= E\left[\left\{\frac{\bar{Y}_1(1+\epsilon_{01})}{\bar{Y}_2(1+\epsilon_{02})} - 1\right\}\left\{\frac{\bar{X}_1(1+\epsilon_1^{'})}{\bar{\bar{X}}_1(1+\epsilon_1^{''})} - 1\right\}\right]$$

$$= R\, E\left[\{(1 +\epsilon_{01})(1 +\epsilon_{02})^{-1} - 1\}\left\{\left(1 +\epsilon_1^{'}\right)\left(1 +\epsilon_1^{''}\right)^{-1} - 1\right\}\right]$$

Neglecting the terms involving powers in $\epsilon_{01}, \epsilon_{02}, \epsilon_1^{'}$ and ϵ_1'' of order higher than two, we have

$$E(m - R)(z_1 - 1) = R[\{E(\epsilon_{01}\epsilon_1') - E(\epsilon_{01}\epsilon_1'')\} - \{E(\epsilon_{02}\epsilon_1') - E(\epsilon_{02}\epsilon_1'')\}]$$

$$= R\left[\left\{\theta\rho_{11}^*\frac{S_{y_1}}{\bar{Y}_1}\frac{S_{x_1}}{\bar{X}_1} - \theta^{'}\rho_{11}^*\frac{S_{y_1}}{\bar{Y}_1}\frac{S_{x_1}}{\bar{X}_1}\right\} - \left\{\theta\rho_{21}^*\frac{S_{y_2}}{\bar{Y}_2}\frac{S_{x_1}}{\bar{X}_1} - \theta^{'}\rho_{21}^*\frac{S_{y_2}}{\bar{Y}_2}\frac{S_{x_1}}{\bar{X}_1}\right\}\right]$$

$$= R\left[\left(\theta - \theta^{'}\right)\rho_{11}^*\frac{S_{y_1}}{\bar{Y}_1}\frac{S_{x_1}}{\bar{X}_1} - \left(\theta - \theta^{'}\right)\rho_{21}^*\frac{S_{y_2}}{\bar{Y}_2}\frac{S_{x_1}}{\bar{X}_1}\right]$$

$$= R\left(\theta - \theta^{'}\right)\left[\frac{S_{x_1}}{\bar{X}_1}\left(\rho_{11}^*\frac{S_{y_1}}{\bar{Y}_1} - \rho_{21}^*\frac{S_{y_2}}{\bar{Y}_2}\right)\right]$$

$$= R\left(\theta - \theta^{'}\right)B_1$$

Similarly, we can define $\quad B_j = \frac{S_{x_j}}{\bar{X}_j}\left(\rho_{1j}^*\frac{S_{y_1}}{\bar{Y}_1} - \rho_{2j}^*\frac{S_{y_2}}{\bar{Y}_2}\right)$

The expressions given in theorems can be obtained from (3.4) and (3.5).

APPENDIX 2

Table 1. Mean square error (MSE) and the percentage relative efficiency (PRE) of T_e with respect to $\hat{R}$ for fixed sample sizes, cost and variance

Estimators	Auxiliary character(s)	MSE in 10^{-5} and P.R.E. of T_e with respect to $\hat{R}$ for fixed $n' = 80$ and $n = 20$			P.R.E. with respect to $\hat{R}$ for the fixed cost $\mathbf{C_0}$ = Rs. 280.00					Expected cost of the estimators $\hat{R}$ and T_e for the specified variance $\boldsymbol{M_0} = 1250 \times 10^{-5}$		
		k^{-1}			$k_{opt.}$	$n'_{opt.}$ (approx.)	$n_{opt.}$ (approx.)	MSE(.) in 10^{-5}	P.R.E.(.) in %	$n'_{opt.}$ (approx.)	$n_{opt.}$ (approx.)	E.C. (.) in Rs.
		4^{-1}	3^{-1}	2^{-1}								
$\hat{R}$	–	100.00 (1153)*	100.00 (979)	100.00 (804)	1.5199	-	8	1977	100.00	-	12	426.19
T_e	x_1 $a_1 = -0.1759$	104.72 (1101)	105.61 (927)	106.92 (752)	1.4280	36	8	1881	105.10	52	11	406.79
T_e	x_1, x_2 $a_1 = -0.8046$, $a_2 = 0.7711$	112.82 (1022)	115.45 (848)	119.47 (673)	1.2782	60	7	1650	119.82	78	10	360.47
T_e	x_1, x_2, x_3 $a_1 = -0.6947$, $a_2 = -0.8046$, $a_3 = 0.7711$	127.83 (902)	134.48 (728)	145.39 (553)	1.0072	95	7	1240	159.44	94	7	278.02

*Figures in parenthesis give the MSE(·).

STATISTICS IN TRANSITION new series, Summer 2014
Vol.15, No. 3, pp. 403–426

BAYESIAN INFERENCE FOR EXPONENTIATED PARETO MODEL WITH APPLICATION TO BLADDER CANCER REMISSION TIME

Sanjay Kumar Singh [1], **Umesh Singh** [2], **Manoj Kumar** [3]

ABSTRACT

Maximum likelihood and Bayes estimators of the unknown parameters and the expected experiment times of the exponentiated Pareto model have been obtained for progressive type-II censored data with binomial removal scheme. Markov Chain Monte Carlo (MCMC) method is used to compute the Bayes estimates of the parameters of interest. The generalized entropy loss function and squared error loss function have been considered for obtaining the Bayes estimators. Comparisons are made between Bayesian and maximum likelihood (ML) estimators via Monte Carlo simulation. The proposed methodology is illustrated through real data.

Key words: PT-II CBR, MLE, bayes estimators, average experiment time.

1. Introduction

The exponentiated Pareto model (EPM) was proposed by Gupta, Gupta and Gupta (1998). The probability density function (pdf) and cumulative distribution function (cdf) of the EPM are given by

$$f(x,\alpha,\theta) = \alpha\theta\left[1-(1+x)^{-\alpha}\right]^{\theta-1}(1+x)^{-(\alpha+1)} \quad ;x>0,\alpha>0,\theta>0 \quad (1)$$

and

$$F(x,\alpha,\theta) = \left[1-(1+x)^{-\alpha}\right]^{\theta} \quad ;x>0,\alpha>0,\theta>0 \quad (2)$$

respectively, where α and θ are the shape parameters of the model. The reliability function takes the following form:

$$S(x) = 1-F(x,\alpha,\theta) = 1-\left[1-(1+x)^{-\alpha}\right]^{\theta}, x>0,\alpha>0,\theta>0. \quad (3)$$

[1] Department of Statistics, Banaras Hindu University,Varanasi-221005.

[2] Department of Statistics and DST-CIMS, Banaras Hindu University,Varanasi-221005.

[3] Assistant Prof. (Statistics), Department of School of Basic Science and Research, Sharda University, Greater Noida, UP. E-mail: manustats@gmail.com.

A distinguished feature of EPM is that because it accommodates all types of failure rates (i.e both monotone and non-monotone). Therefore, it can be effectively used for analyzing various types of data. It may also be noted that a number of distributions can be obtained as particular cases of it. For the shape parameter $\theta = 1$, the EPM is reduced to standard Pareto distribution of second kind (see, Johnson Kotz and Balakrishnan, 1994). For more details about EPM, we refer to Gupta Gupta and Gupta (1998). Some statistical properties of this distribution and the estimators of the parameters of EPM have been discussed by Shawky and Abu-Zinadah (2009) under different estimation procedures for complete sample case. In general life testing experiments, situations do arise when units are lost or removed from the experiment while they are still functioning, i.e. we get censored data from the experiment. The loss of units may occur due to time constraints, giving type-I censored data. In such a censoring scheme, the experiment is terminated at some specified time. Sometimes, the experiment is terminated after a prefixed number of observations due to cost constraints and we get type-II censored data. The estimation of parameters of EPM has also been attempted by Afify (2010) under type-I and type-II censoring scheme. Besides the above two controlled causes, units may drop out of the experiment randomly due to some uncontrolled causes such type of situation progressive censoring arises.

For example, consider that a doctor performs an experiment with n bladder cancer patients with remission times (in months), i.e. a period during which symptoms of disease are reduced (partial remission) or disappear (complete remission) with regard to cancer, remission means there is no sign of it on scans or when the doctor examines you. Doctors use the word 'remission' instead of cure when talking about cancer because they cannot be sure that there are no cancer cells at all in the body. So the cancer could come back in the future. But the complete remission would therefore be better than partial remission. Because with partial remission the chances of occurrence of bladder cancer are higher, its means remission times (in months) are the minimum that represents partial remission, when remission times (in months) are longer, say complete remission. So the doctor performs an experiment on bladder cancer patients with partial and complete remission times (in months) are very costly and time-consuming. Due to cost constraint the experiment is terminated after a prefixed number of bladder cancer patients and we get type-II censored data. After type-II censoring another situation of bladder cancer patients with remission times (in months) may arise, the first bladder cancer patient has died due to some other unforeseen circumstances such as heart attack, accident, damage of lever, depletion of funds, etc.; some patients leave the experiment and go for treatment to other doctor/hospital. Similarly, after the second death a few more leave and so on. Finally, the doctor stops taking observation as soon as the predetermined number of deaths (say m) is recorded. Which has arise a scenario of progressive type-II censoring with random/binomial removals. For further details, readers are referred to Balakrishnan (2007). In last few years, the estimation of parameters of different life time distribution based on progressive censored samples have been studied by several authors such as Childs and Balakrishnan (2000), Balakrishnan and Kannan (2001), Mousa and Jheen (2002), Ng, Chn and Balakrishnan (2002). The progressive type-II censoring with binomial removal were considered by Yang, Tse and Yuen (2000) for Weibull distribution, Wu and Chang (2002) for Exponential distribution. Under the progressive type-II censoring with random removals, Wu and Chang (2003) and Yuen and Tse (1996) developed the estimation problem for the Pareto distribution and Weibull distribution respectively, when the number of units removed at each failure time has a discrete uniform distribution, the expected time of this censoring plan is discussed and compared numerically. Mathematically, this experiment is similar to a life test experiment which starts with n units. At the first failure X_1, r_1 (random) units are removed randomly from the remaining $(n-1)$ surviving units. At the second failure X_2, r_2 units from remaining $n-2-r_1$ units are removed, and so on; untill m^{th} failure is observed, i.e. at m^{th} failure all the remaining $r_m = n-m-r_1-r_2\cdots r_{m-1}$ units are removed. Note that here m is pre-fixed and $r_i^{,}s$ are random. Such a censoring mechanism is termed as progressive type-II censoring with random removal scheme. If we assume that probability of removal of a unit at every stage is p for each unit then r_i can be considered to follow a binomial distribution i.e, $r_i \approx B(n-m-\sum_{l=0}^{i-1} r_l, p)$ for $i = 1, 2, 3, \cdots m-1$ and with $r_0 = 0$. The main aim of this article is concerned with the problem of obtaining Bayes estimates for the two parameter EPM based on progressive type-II censoring with binomial removals (PT-II CBR). Bayes estimators are obtained based on under square error loss function (SELF) and generalized entropy loss function (GELF). The results are obtained to PT-II CBRs, and compare the expected test times for PT-II CBR with complete sampling scheme. However, no attempt has been made to develop estimators for the parameters of EPD under PT-II CBR and its applications are discussed based on real illustration. Therefore, we propose to develop such an estimation procedure. The rest of the paper is organized as follows.

Section 2, provides the likelihood function. The ML estimators of the unknown parameters are presented in section 3. Section 4 contains the loss functions, prior distributions, the Bayes estimates using the MCMC via Gibbs sampling scheme. An algorithm for simulating the PT-II CBR is presented in section 5. We compare the expected test times under PT-II CBRs with complete sample which are given in section 6. The comparison of ML estimators and corresponding Bayes estimators are presented in section 7. These comparisons are based on simulated risk (average loss over sample space) of the estimators and discussion of results is presented. In section 8, we provide an application of the EPD distribution to remission time of bladder cancer. Finally, some conclusions are drawn in section 9.

2. Likelihood function

Let $(X_1, R_1), (X_2, R_2), (X_3, R_3), \cdots, (X_m, R_m)$ denote a progressive type-II censored sample, where $X_1 < X_2 < X_3, \cdots, X_m$.With pre-determined number of removals, say $R_1 = r_1, R_2 = r_2, R_3 = r_3, \cdots, R_m = r_m$, the conditional likelihood function can be written as, Cohen(1963)

$$L(\alpha;\theta;x|R=r) = c^{*}\prod_{i=1}^{m} f\left(x_i\right)\left[S\left(x_i\right)\right]^{r_i}, \tag{4}$$

where $c^{*} = n(n-r_1-1)(n-r_1-r_2-2)(n-r_1-r_2-r_3-3)\cdots(n-r_1-r_2-r_3,\cdots,r_m-m+1)$, and $0 \leq r_i \leq (n-m-r_1-r_2-r_3\cdots r_{i-1})$, for $i=1,2,3\ldots,m-1$. Substituting (1) and (3) into (4), we get

$$L(\alpha,\theta;x|R=r) = \prod_{i=1}^{m} \alpha\theta\left[1-(1+x_i)^{-\alpha}\right]^{\theta-1}\left\{1-\left[1-(1+x_i)^{-\alpha}\right]^{\theta}\right\}^{r_i}(1+x_i)^{-(\alpha+1)}. \tag{5}$$

Suppose that an individual unit being removed from the test at the i^{th} failure, $i=1,2,\cdots(m-1)$ is independent of the others but with the same probability p. That is the number R_i of the unit removed at i^{th} failure $i=1,2,\cdots(m-1)$ follows a binomial distribution with parameters $\left(n-m-\sum_{l=1}^{i-1} r_i, p\right)$ therefore,

$$P\left(R_1=r_1;p\right) = \binom{n-m}{r_1} p^{r_1}(1-p)^{n-m-r_1}, \tag{6}$$

and for $i=2,3,\cdots,m-1$,

$$\begin{aligned} P\left(R_{;}p\right) &= P\left(R_i=r_i|R_{i-1}=r_{i-1},\cdots R_1=r_1\right) \\ &= \binom{n-m-\sum_{l=0}^{i-1} r_l}{r_i} p^{r_i}(1-p)^{n-m-\sum_{l=0}^{i-1} r_l}. \end{aligned} \tag{7}$$

Now, we further assume that R_i is independent of X_i for all i. Then, using above equations, we can write the full likelihood function as in the following form

$$L\left(\alpha,\theta,p;x,r\right) = AL_1\left(\alpha,\theta\right)L_2\left(p\right), \tag{8}$$

where

$$L_1(\alpha;\theta) = \prod_{i=1}^{m} \alpha\theta\left[1-(1+x_i)^{-\alpha}\right]^{\theta-1}\left\{1-\left[1-(1+x_i)^{-\alpha}\right]^{\theta}\right\}^{r_i}(1+x_i)^{-(\alpha+1)}, \tag{9}$$

$$L_2\left(p\right) = p^{\sum_{i=1}^{m-1} r_i}\left(1-p\right)^{(m-1)(n-m)-\sum_{i=1}^{m-1}(m-i)r_i}. \tag{10}$$

and $A = \frac{c^{*}(n-m)!}{\left(n-m-\sum_{l=1}^{i-1} r_i\right)!\prod_{i=1}^{m-1} r_i!}$ does not depend on the parameters α,θ and p.

3. ML estimation

The ML estimations of α and θ are the simultaneous solutions of following normal equations

$$\begin{aligned} &\frac{m}{\alpha} + (\theta - 1) \sum_{i=1}^{m} \frac{(1+x_i)^{-\alpha} \ln (1+x_i)}{1-(1+x_i)^{-\alpha}} - \sum_{i=1}^{m} \ln (1+x_i) - \\ &\theta \sum_{i=1}^{m} \frac{r_i \left[1-(1+x_i)^{-\alpha}\right]^{\theta-1} (1+x_i)^{-\alpha} \ln (1+x_i)}{1-\left[1-(1+x_i)^{-\alpha}\right]^{\theta}} = 0, \end{aligned} \tag{11}$$

and

$$\begin{aligned} &\frac{m}{\theta} + \sum_{i=1}^{m} \ln \left[1-(1+x_i)^{-\alpha}\right] \\ &\sum_{i=1}^{m} \frac{r_i \left[1-(1+x_i)^{-\alpha}\right]^{\theta} \ln \left[1-(1+x_i)^{-\alpha}\right]}{1-\left[1-(1+x_i)^{-\alpha}\right]^{\theta}} = 0. \end{aligned} \tag{12}$$

It may be noted that (11) and (12) cannot be solved simultaneously to provide a nicely closed form for the estimators. Therefore, we propose to use fixed point iteration method for solving these equations. For details about the proposed method readers may refer to Jain, Iyengar and Jain (1984).

4. Bayesian estimation

This section is concerned with prior distributions for unknown parameters, symmetric and asymmetric loss function and Bayes estimates using the Gibbs sampling scheme.

Prior and posterior distributions

In order to obtain the Bayes estimators of unknown parameters α and θ based on PT-II CBRs, we must assume that the parameters α and θ are random variables. The model under consideration has shapes and censoring parameters, and continuous conjugate priors for these parameters do not exist. We further assume that these random variables α and θ have independently distributed informative prior distribution with respective prior pdfs

$$g_1(\alpha) = \frac{{\lambda_1}^{\nu_1} e^{-\lambda_1 \alpha} \alpha^{\nu_1 - 1}}{\Gamma \nu_1} \quad ; \quad 0 < \alpha < \infty, \quad \lambda_1 > 0, \quad \nu_1 > 0 \tag{13}$$

$$g_2(\theta) = \frac{{\lambda_2}^{\nu_2} e^{-\lambda_2 \theta} \theta^{\nu_2 - 1}}{\Gamma \nu_2} \quad ; \quad 0 < \theta < \infty, \quad \lambda_2 > 0, \quad \nu_2 > 0 \tag{14}$$

respectively. But Arnold and Press (1983) had all ready discussed that there is no clear cut way in which one can say that one prior is better than the other. But for purpose of Bayesian analysis, the gamma prior $g_1(\alpha)$ and $g_2(\theta)$ are chosen instead of the exponential prior of α and θ used by Eissa and Nassar (2004) and Jung, Chung and Kim (2011) because the gamma prior is wealthy enough to cover the prior belief of the experimenter for different shapes. On the basis of the above stated assumptions, the joint prior pdf of α and θ is

$$g(\alpha,\theta) = g_1(\alpha)\, g_2(\theta) \quad ; \quad \alpha > 0, \quad \theta > 0 \tag{15}$$

Combining the priors given by (13) and (14) with likelihood given by (9), we can easily obtain joint posterior pdf of (α,θ) as $\pi(\alpha,\theta|x,r) = \frac{J_1}{J_0}$ where

$$\begin{aligned} J_1 = \alpha^{m+\nu_1-1}\theta^{m+\nu_2-1}e^{-\left(\sum_{i=1}^{m}\lambda_1\alpha+\sum_{i=1}^{m}\lambda_2\theta\right)}\prod_{i=1}^{m}\left[1-(1+x_i)^{-\alpha}\right]^{\theta-1} \\ \left\{1-\left[1-(1+x_i)^{-\alpha}\right]^{\theta}\right\}^{r_i}(1+x_i)^{-(\alpha+1)}, \end{aligned} \tag{16}$$

and $J_0 = \int_0^\infty \int_0^\infty J_1 d\alpha d\theta$. Hence, the respective marginal posterior pdfs of α and θ are given by

$$\pi_1(\alpha|x,r) = \int_0^\infty \frac{J_1}{J_0} d\theta, \tag{17}$$

and

$$\pi_2(\theta|x,r) = \int_0^\infty \frac{J_1}{J_0} d\alpha. \tag{18}$$

Loss functions

In order to select the best decision in decision theory, an appropriate loss function must be specified. For this purpose, we use symmetric as well as asymmetric loss function. The Bayes estimators are obtained under SELF

$$l_1(\phi,\hat{\phi}) = \in_1 \left(\phi-\hat{\phi}\right)^2 ; \quad \in_1 > 0 \tag{19}$$

where $\hat{\phi}$ is the estimate of the parameter ϕ and the Bayes estimator $\hat{\phi}_S$ of ϕ comes out to be $E_\phi[\phi]$, where E_ϕ denotes the posterior expectation. However, this loss function is symmetric loss function and can only be justified if over-estimation and under-estimation of equal magnitudes are of equal seriousness. A number of asymmetric loss functions are also available in the statistical literature. Let us consider the GELF, proposed by Calabria and Pulcini (1996), defined as follows :

$$l_2(\phi,\hat{\phi}) = \in_2 \left(\left(\frac{\hat{\phi}}{\phi}\right)^\delta - \delta\ln\left(\frac{\hat{\phi}}{\phi}\right) - 1\right) ; \quad \in_2 > 0 \tag{20}$$

The constant δ, involved in (20), is its shape parameter. It reflects departure from symmetry. When $\delta > 0$, it considers over-estimation (i.e., positive error) to be more serious than under-estimation (i.e., negative error) and converse for $\delta < 0$. The Bayes estimator $\hat{\phi}_E$ of ϕ under GELF is given by

$$\hat{\phi}_E = \left[E_\phi\left(\phi^{-\delta}\right)\right]^{\left(-\frac{1}{\delta}\right)} \tag{21}$$

provided the posterior expectation exits. It may be noted here that for $\delta = -1$, the Bayes estimator under loss (19) coincides with the Bayes estimator under SELF l_1. Expressions for the Bayes estimators $\hat{\alpha}_E$ and $\hat{\theta}_E$ for α and θ respectively, under GELF can be given as

$$\widehat{\alpha_E} = \left[\int_0^\infty \alpha^{-\delta}\pi_1\left(\alpha|x,r\right)d\alpha\right]^{\left(-\frac{1}{\delta}\right)}, \tag{22}$$

and

$$\widehat{\theta_E} = \left[\int_0^\infty \theta^{-\delta}\pi_1\left(\theta|x,r\right)d\theta\right]^{\left(-\frac{1}{\delta}\right)}, \tag{23}$$

It is to mention here that from equation (22) and (23), the Bayes estimators $\hat{\alpha}_E$ and $\hat{\theta}_E$ are not reducible in a nice closed form. Therefore, we use the numerical techniques for obtaining the estimates. We, therefore, propose to consider Gibbs sampling procedure.

MCMC method via Gibbs sampling

In this subsection, we use the Gibbs sampling procedure to obtain the Bayes estimates α and θ under SELF and GELF. It is clear from equations (22) and (23) that the Bayes estimators of α and θ are not obtained analytically and numerical techniques must be used for computations. To compute Bayes estimators of the parameters α and θ we propose to use MCMC technique, via Gibbs sampler along with Metropolis-Hastings algorithms to generate samples from posterior distributions and then compute Bayes estimates. The Gibbs is an algorithm for simulating from the full conditional posterior distributions while the Metropolis-Hastings algorithm generates samples from an (essentially) arbitrary proposal distribution. For more details about the MCMC methods see, for example, Vasishta, Smith and Upadhyay (2001) and Gupta and Upadhyay (2010). The full conditional posterior distributions of the parameters α and θ are, respectively, given as

$$\tau_1(\alpha|x,r) \propto \alpha^{m+\nu_1-1}e^{-(\Sigma_{i=1}^m \lambda_1\alpha)}\prod_{i=1}^m\left[1-(1+x_i)^{-\alpha}\right]^{\theta-1}\left\{1-\left[1-(1+x_i)^{-\alpha}\right]^\theta\right\}^{r_i} (1+x_i)^{-(\alpha+1)} \tag{24}$$

$$\tau_2(\theta|x,r) \propto \theta^{m+\nu_2-1}e^{-(\Sigma_{i=1}^m \lambda_2\theta)}\prod_{i=1}^m\left[1-(1+x_i)^{-\alpha}\right]^{\theta-1}\left\{1-\left[1-(1+x_i)^{-\alpha}\right]^\theta\right\}^{r_i} \tag{25}$$

The following MCMC algorithm is used to generate the posterior samples and then to obtain the Bayes estimates of α and θ.

I. Start with initial guesses of α and θ say α_0 and θ_0.

II. Set j=1.

III. Generate α_1 from $\tau_1(\alpha|\theta, \mathbf{x}, r)$ and θ_1 from $\tau_2(\theta|\alpha, \mathbf{x}, r)$.

IV. Repeat steps 2-3, N times.

V. Now, the Bayes estimates of α and θ under GELF are, respectively, given as

$$\hat{\alpha}_E = \left[\frac{1}{N-M} \sum_{j=M+1}^{N} {\alpha_j}^{-\delta}\right]^{-1/\delta} \tag{26}$$

$$\hat{\theta}_E = \left[\frac{1}{N-M} \sum_{j=M+1}^{N} {\theta_j}^{-\delta}\right]^{-1/\delta} \tag{27}$$

VI. Put $\delta = -1$ in above step 5, then the Bayes estimator under GELF coincides with Bayes estimator under SELF.
where M is the burn-in period (i.e, the number of iterations before the stationary distribution is achieved).

5. Algorithm for PT-II CBR

We need to simulate PT-II CBR from specified EPD. To get such a sample, we propose the use of the following algorithm:

I. Specify the value of n.

II. Specify the value of m.

III. Specify the value of parameters α, θ and p.

IV. Generate random number r_i from B$\left(n - m - \sum_{l=0}^{i-1} r_l, p\right)$, for $i = 1, 2, 3, \cdots,$ $m - 1$.

V. Set r_m according to the following relation.

VI. $r_m = \begin{cases} n - m - \sum_{l=1}^{m-1} r_l & \text{if } n - m - \sum_{l=1}^{m-1} r_l > 0 \\ 0 & \text{otherwise} \end{cases}$

VII. Generate m independent $U(0,1)$ random variables $W_1, W_2, \cdots, W_m$.

VIII. For given values of the progressive type-II censoring scheme $r_i(i = 1, 2, \cdots, m)$ set $V_i = W_i^{1/(i+r_m+\cdot+r_{m-i+1})}(i = 1, 2, \cdots, m)$.

IX. Set $U_i = 1 - V_m V_{m-1} \cdots V_{m-i+1}(i = 1, 2, \cdots, m)$, then $U_1, U_2, \cdots, U_m$ are progressive type-II censored samples with binomial removals of size m from $U(0,1)$.

X. Finally, for given values of parameters α and λ, we set $x_i = F^{-1}(U)(i = 1, 2, \cdots, m)$. Then,$(x_1, x_2, \cdots, x_m)$ is the required from progressive censoring with binomial removals sample of size m from the EPD.

6. Average Experiment Time

In practical situations, an Experimenter may be interested to know whether the test can be completed within a specified time. This information is important for an experimenter to choose an appropriate sampling plan because the time required to complete a test is directly related to cost. Under Progressive censoring with a fixed number of removal the time is given by X_m. According to Balakrishnan and Aggarwalla (2000), the expected value of X_m is given by

$$E\left[X_m|R\right]=C(r)\sum_{l_1=0}^{r_1}\sum_{l_2=0}^{r_2}\cdots\sum_{l_m=0}^{r_m}(-1)^B\frac{C_{l_1=0}^{r_1}\cdots C_{l_m=0}^{r_m}}{\prod_{i=1}^{m-1}h(l_i)}\int_0^\infty xf(x)F^{h(l_m)-1}(x)\partial x. \tag{28}$$

where $B=\sum_{i=1}^m l_i$, $h(l_i)=l_1+l_2+\cdots+l_i+i$, $C(r)=n(n-r_1-1)(n-r_1-r_2-2)\cdots[n-\sum_{i=1}^{m-1}(r_i+1)]$ and i is the number of live units removed from experiment (number of failure units). Using the p.d.f and c.d.f of EPD, the equation will be

$$E\left[X_m|R\right]=C(r)\sum_{l_1=0}^{r_1}\sum_{l_2=0}^{r_2}\cdots\sum_{l_m=0}^{r_m}(-1)^B\frac{C_{l_1=0}^{r_1}\cdots C_{l_m=0}^{r_m}}{\prod_{i=1}^{m-1}h(l_i)}$$
$$\int_0^\infty x_i\alpha\theta\left[1-(1+x)^{-\alpha}\right]^{\theta-1}(1+x)^{-(\alpha+1)}\left\{\left[1-(1+x)^{-\alpha}\right]^\theta\right\}^{(h(l_m)-1)} \tag{29}$$

Let

$$\begin{aligned}S_1&=\alpha\theta\int_0^\infty x_i\left[1-(1+x)^{-\alpha}\right]^{\theta-1}(1+x)^{-(\alpha+1)}\left\{\left[1-(1+x)^{-\alpha}\right]^\theta\right\}^{(h(l_m)-1)}\\&=\alpha\theta\int_0^\infty x_i(1+x)^{-(\alpha+1)}\left[1-(1+x)^{-\alpha}\right]^{(h(l_m)\theta-1)}\partial x_i.\\&=\alpha\theta\sum_{k=0}^{h(l_m)\theta-1}(-1)^k\binom{h(l_m)\theta-1}{k}\int_0^\infty\frac{x_i}{(1+x_i)^{(\alpha(k+1)+1)}}\partial x_i\\&=\alpha\theta\sum_{k=0}^{h(l_m)\theta-1}(-1)^k\binom{h(l_m)\theta-1}{k}B_{II}(2,\alpha(k+1)-1)\end{aligned}$$

Putting this value in to the right hand of equation (29), the expected test time is given by

$$E\left[X_m|R\right] = C(r)\alpha\theta \sum_{l_1=0}^{r_1}\sum_{l_2=0}^{r_2}\cdots\sum_{l_m=0}^{r_m}(-1)^B \frac{C_{l_1=0}^{r_1}\cdots C_{l_m=0}^{r_m}}{\prod_{i=1}^{m-1}h(l_i)}$$
$$\alpha\theta\sum_{k=0}^{h(l_m)\theta-1}(-1)^k\binom{h(l_m)\theta-1}{k}B_{II}(2,\alpha(k+1)-1) \qquad (30)$$

The expected test time for PT-II CBRs is evaluated by taking expectation on both sides (29) with respect to the R. That is

$$\begin{aligned} E\left[X_m\right] &= E_R\left[E\left[X_m|R=r\right]\right] \\ &= \sum_{r_1=0}^{g(r_1)}\sum_{r_2=0}^{g(r_2)}\cdots\sum_{r_{m-1}=0}^{g(r_{m-1})}P(R,p)E\left[X_m|R=r\right]. \end{aligned} \qquad (31)$$

where $g(r_i) = n - m - r_1 - \cdots - r_{i-1}$ and $P(R;p)$ is given in equation (7). For the expected time a complete sampling case with n test units is obtained by taking $m = n$ and $r_i = 0$ for all $i = 1, 2, \cdots, m$, in (30). We have

$$E\left[X_n^*\right] = n\alpha\theta\sum_{k=0}^{n-1}\binom{n-1}{k}(-1)^k B_{II}(2,\alpha(k+1)-1). \qquad (32)$$

Also, the expected time of a type-II censoring without removal is defined by the expected value of the m^{th} failure time, then

$$E\left[X_m^*\right] = m\alpha\theta\binom{n}{m}\sum_{k=0}^{m-1}\binom{m-1}{k}(-1)^k B_{II}(2,\alpha(k+1)-1), \qquad (33)$$

The ratio of the expected experiment time (REET)δ_{REET} is computed between PT-II CBR and the complete sampling, we define

$$\delta_{REET} = \frac{E[X_m]\ Under\ PT-II\ CBR}{E[X_n{}^*]\quad under\quad complete\quad sampling}. \qquad (34)$$

It can be noted from δ_{REET} that important information is given in order to determine significantly the shortest experiment time if a much larger sample of n test units is used, the test is terminated, when m^{th} failures have been observed. But here we are interested in considering various values of n, m and p, numerically calculated under the expected experiment time of PT-II CBR and complete sample, which are derived in equations (31) and (32). Numerical results are obtained in Table 7 where for $n = 15, 12$ and 9 corresponding choices of m are given. From Table 7 we observed that, when n is fixed, the values of the δ_{REET} and expected termination time under PT-II CBR and complete test decrease as m decreases, while for fixed m, the value of the δ_{REET} and expected termination time under PT-II CBR and complete sampling increase as n decreases. Finally, for fixed values of m and n, we

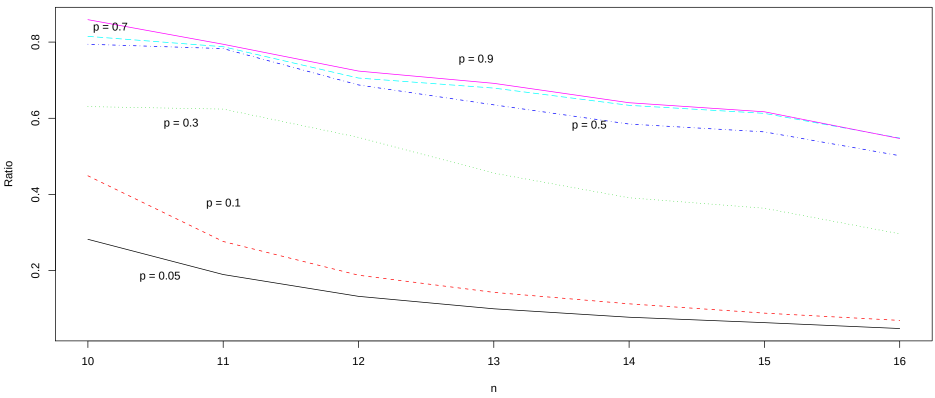


FIGURE 1. δ_{REET} under PT-II CBRs to δ_{REET} under complete sample

observed that the effect of variation of removal probability p with the values of the δ_{REET} and expected termination time of PT-II CBR increase as p increases.

Figure 1 shows the ratio of the expected test time under PT-II CBR to the expected test time under complete sample versus n for $m = 8$ and different values of removal probability p. We observed that, when the value of p is large, the ratio increases and approaches 1 quickly and the expected test time is not small in these cases. Hence, for small p, the expected test time is more significant than larger value of p. So, we have taken $p = 0.3$ from Figure 1, which was significant for further calculation.

7. Simulation studies

The estimators $\hat{\alpha}_M$ and $\hat{\theta}_M$ denote the ML estimators of the parameters α and θ respectively while $\hat{\alpha}_S$ and $\hat{\theta}_S$ are corresponding Bayes estimators under SELF and $\hat{\alpha}_E$ and $\hat{\theta}_E$ are the corresponding Bayes estimators under GELF. We compare the estimators obtained under GELF with corresponding ML estimators and Bayes estimators under SELF. The comparisons are based on the simulated risks (average loss over sample space) under GELF. It may be mentioned here that the exact expressions for the risks cannot be obtained because estimators are not in a nice closed form. Therefore, the risks of the estimators are estimated on the basis of Monte-carlo simulation study of 10000 samples. It may be noted that the risks of the estimators will depend on values of $n, m, \theta, \alpha, p, \lambda_1, \lambda_2, \nu_1, \nu_2$ and δ. In order to consider variation in the values of these parameters, we have obtained the simulated risks for $m = 9\,[3]\,15$, when $n = 15$, $\theta = 0.5$, $\alpha = 2$, $\delta = \pm 0.5$ and $p = 0.3$. For prior distribution we have used non-informative prior with $\lambda_1 = \lambda_2 = \nu_1 = \nu_2 = 0$, and informative prior and the hyper parameter are chosen in such a way that the prior mean became true value of the parameter and belief in prior mean strong or weak, i.e. the prior variance is small and large. Thus, the values of the hyper parameter of informative prior are $\lambda_1 = (0.5, 4), \lambda_2 = (0.125, 1), \nu_1 = (1, 8), \nu_2 = (0.0625, 0.5)$. Generating the progressive sample as mentioned in section 4, the simulated risks under SELF and GELF have been obtained for different values of m with selected values of the rest of the parameters $n, \theta, \alpha, p, \lambda_1, \lambda_2, \nu_1, \nu_2$ and δ have been taken. The results are given in tables Table 1-6. The entries in brackets in all the tables denote the risks of the estimators when δ is negative and the other non-bracket entries are the risks when δ is positive.

Discussion of the results

It is interesting to note that when effective sample size m increases, keeping n, fixed for fixed positive value of δ under both losses, the risks of the ML estimate of α, first increase then decrease slightly as m increases whereas the risks of Bayes estimators always increase with the increase in the value of m. This trend of the magnitude of the risks is also the same for fixed negative value of δ. It is observed when non-informative prior for α has been used (see, Table 1). While regarding the considered prior distribution, when we have smaller belief in considered prior distribution for α, i.e. prior variance is 1, then we observe that in over-estimation situation under both losses, the risks of estimator $\hat{\alpha}_M$ increase then slightly decrease as m increases but in under estimation situation under both losses, the risks of estimator of $\hat{\alpha}_M$ decrease then slightly increase as m increases. Finally, we observed that under both losses for positive and negative values of δ, the risks of estimator of $\hat{\alpha}_S$ and $\hat{\alpha}_E$ increase as m increases (see, Table 2). For larger prior variance of α, we observed that under both losses for $\delta < 0$, the risks of estimator $\hat{\alpha}_M$ decrease as m increases, and the rest of them for $\delta < 0$ and $\delta > 0$, the risks of estimators $\hat{\alpha}_S$, $\hat{\alpha}_E$ and for $\delta > 0$ $\hat{\alpha}_M$ increase as m increases (see, Table 3). The risk of estimators of θ under SELF and GELF, when priors for the parameter θ are non-informative types, the risks of estimator $\hat{\theta}_M$, decrease in case of both positive and negative values of δ, and the risks of Bayes estimators increase as m increases for both positive and negative values of δ, and under both losses namely SELF and GELF (see, Table 4). For smaller prior variance of θ, we observed under both losses that when $\delta > 0$, the risk of estimator $\hat{\theta}_M$ decreases as m increases but when $\delta < 0$, the risk of estimator $\hat{\theta}_M$ first increases then decreases as m increases and as in the previous table the risk of Bayes estimators as m increases for both positive and negative values of δ under both losses. The risk of estimators of θ under SELF and GELF, when prior for the parameter θ are non informative types, the risks of estimator $\hat{\theta}_M$ decrease in case of both positive and negative values of δ and the risks of Bayes estimators increase as m increases for both positive and negative values of δ and under both losses, namely SELF and GELF (see, Table 5). For larger prior variance of θ, under both losses, for $\delta > 0$ and $\delta < 0$, the risk of estimator $\hat{\theta}_M$ decreases as m increases and as in the previous cases, the trends for the risks are the same (see, Table 6).

8. Application

In this section we reanalyze the data extracted from Luz, Silva, Rodrigo, Bouruignon, Andrea and Gauss Coreiro (2012). For the purpose of real illustration, we have been discussed in presence of PT-II CBR. The data describe a study of remission time(in months) of a random sample of 128 bladder cancer patients reported in Lee and Wang (2003). The data are given as

0.08,2.09, 3.48, 4.87, 6.94, 8.66, 13.11, 23.63, 0.20, 2.23, 3.52, 4.98, 6.97, 9.02, 13.29,
0.40, 2.26,3.57, 5.06, 7.09, 9.22, 13.80, 25.74, 0.50, 2.46, 3.64, 5.09, 7.26, 9.47, 14.24,
25.82, 0.51, 2.54,3.70, 5.17, 7.28, 9.74, 14.76, 26.31, 0.81, 2.62, 3.82, 5.32, 7.32, 10.06,
14.77,32.15, 2.64, 3.88,5.32, 7.39, 10.34, 14.83, 34.26, 0.90, 2.69, 4.18, 5.34,7.59, 10.66,
15.96, 36.66, 1.05, 2.69, 4.23,5.41, 7.62, 10.75, 16.62, 43.01, 1.19, 2.75, 4.26, 5.41, 7.63,
17.12, 46.12, 1.26,2.83, 4.33, 5.49,7.66,11.25, 17.14, 79.05, 1.35, 2.87, 5.62, 7.87, 11.64,
17.36, 1.40, 3.02, 4.34, 5.71, 7.93, 11.79,18.10, 1.46, 4.40, 5.85, 8.26, 11.98, 19.13, 1.76,
3.25, 4.50, 6.25, 8.37, 12.02, 2.02, 3.31, 4.51, 6.54, 8.53, 12.03, 20.28, 2.02, 3.36, 6.76,
12.07, 21.73, 2.07, 3.36, 6.93, 8.65, 12.63, 22.69.

In order to identify the shape of lifetime data failure rate function, we shall consider, as a crude indicative, a graphical method based on TTT (Total time on test) plot Aarset (1985) Hence, in its empirical version the TTT plot is given as

$$T(\tfrac{n}{r}) = \frac{\sum_{i=1}^{r} y_{(i)} + (n-r) y_{(r)}}{\sum_{i=1}^{n} y_{(i)}}$$

where $r = 1, 2, \cdots, n$ and $y_{(r)}$ is the order statistics of the sample. On the basis of TTT plot, we identify that the failure rate function is increasing, decreasing and increasing then decreasing, i.e. when the TTT plot for considered data is concave, convex and concave then convex respectively. Figure 2 shows that TTT plot for considered data, which is concave then convex indicating an increasing then decreasing failure rate function, is properly accommodated by EPD with increasing then decreasing failure rate. According to Figure 3, we observed that this data is appropriate for EPD and Figure 4 shows estimated pdf, CDF and hazard functions. Also, we have obtained Kolmogrov-Smirnov (K-S) Statistics, Akaike's information criterion (AIC) and Bayesian information criterion (BIC) under sub model Pareto distribution for given data set and values summarized in Table 8. According to above considered criterion, we can say that EPD provide better fit than Pareto distribution. Therefore, we use this data to illustrate the proposed methodology. For this PT-II CBRs are generated from the given data under various schemes, which are summarized in Table 11. We have obtained the ML estimates, Bayes estimates (using non-informative prior), 95% CI and HPD intervals for the parameters α and θ respectively under SELF and GELF for $\delta = \pm 1.5$, and value of the hyper parameters α and θ are taken as $\nu_1 = 0.00001, \lambda_1 = 0.0001$ and $\nu_2 = 0.00001, \lambda_2 = 0.0001$ respectively. We have obtained the ML and the Bayes estimates of α and θ under SELF and GELF for $\delta = \pm 1.5$ presented in Table 9 and 10 respectively. When the degree of censoring decreases, the estimate of α and θ is closer to the estimates of without censoring. Under different censoring schemes, the length of HPD intervals is always less than

CI. The ML and Bayes estimates under SELF and GELF of α and θ always lies between HPD and CI.

9. Conclusion

In this paper, we consider a Bayesian estimation of EPD in presence of PT-II CBRs under the asymmetric loss function. We use independent gamma priors for the unknown parameters as the continuous conjugate priors do not exist. It is seen that the explicit expressions for the Bayes estimators are not possible. We obtain the approximate Bayes estimates of parameters using the MCMC via Gibbs sampling scheme. To observe the properties of the Bayes estimators based on the MCMC via Gibbs sampling, some numerical experiments are performed. In general most of cases, when the sample size increases the risk of the estimators decreases. The interesting points are observed regarding PT-II CBR, either prior belief of the model parameter is low or high, our proposed estimators $\hat{\alpha}_E$ and $\hat{\theta}_E$ perform well (in the sense of having smaller risk).

On the other hand, in context of the expected experiment time, we may also conclude that the removal probability p plays a great role in the expected test time. The increase in the removal probability p means more items are removed at the early stage of the experiment. Hence, for larger p, the collection of observations much closer to the tail of the life time distribution and the experiment under PT-II CBR increase as p increases.

Acknowledgement

The authors are thankful to referees and editor for their valuable comments and suggestions for improvement of the manuscript. The authors are also thankful to DST-CIMS, B. H. U., Varanasi, India for providing computational facilities.

REFERENCES

AARSET, M. W., (1985). The null distribution for a test of constant versus bathtub failure rate. *Scandinavian Journal of Statistics*, 12(1):55-68.

AFIFY, W. M., (2010). On estimation of the exponentiated Pareto distribution under different sample scheme. *Applied Mathematical Sciences*, 4(8):393–402.

ARNOLD, B. C., PRESS, S. J., (1983). Bayesian inference for Pareto populations. *J.Econom.*, 21:287-306.

BALAKRISHNAN, N., (2007). Progressive methodology: An appraisal (with discussion). *Test*, 16 (2):211–259.

BALAKRISHNAN, N., AGGARWALLA, R., (2000). *Progressive Censoring: Theory, Methods and Applications*. Birkhauser, Boston.

BALAKRISHNAN, N., KANNAN, N., (2001). Point and Interval Estimation for Parameters of the Logistic Distribution Based on Progressively Type-II Censored Samples, in Handbook of Statisticsm N. Balakrishnan and C. R. Rao, 20. Eds. Amsterdam, North-Holand.

CALABRIA, R., PULCINI, G., (1996). Point estimation under-asymmetric loss functions for life-truncated exponential samples. *Commun. statist. Theory meth.*, 25(3):585–600.

CHILDS, A., BALAKRISHNAN, N., (2000). Conditional inference procedures for the Laplace distribution when the observed samples are Progressively censored. *Metrika*, 52:253–265.

COHEN, A. C., (1963). Progressively censored samples in life testing. *Technometrics*, pages 327–339.

EISSA, F. H., NASSAR, M. M.,(2004). Bayesian estimation for the exponentiated Weibull model. *Communication in Statistics Theory and Methods*, 33:2343–2236.

GUPTA, R. C., GUPTA, R. D., GUPTA, P. L., (1998). Modeling failure time data by Lehman alternatives. *Commun. Statist. - Theory Meth.*, 27(4):887–904.

GUPTA, A., UPADHYAY, S. K., (2010). A Bayes analysis of modified Weibull distribution via Markov chain monte carlo simulation. *Journal of Statistical Computation and Simulation*, 80(3):241–254.

JAIN, M. K., IYENGAR, S. R. K., JAIN, R. K.,(1984). *Numerical Methods for Scientific and Engineering Computation*. New Age International (P) Limited, Publishers, New Delhi, fifth edition.

JOHANSON, N. L., KOTZ, S., BALAKRISHNAN, N., (1994). *Continuous Univariate Distributions*, volume 1. Wiley, New York, 2 edition.

JUNG, J., CHUNG, Y., KIM, C.,(2011). Bayesian estimation for the exponentiated Weibull model under type II progressive censoring. *Statistical Papers (accepted).*

LEE, E. T., WANG, J. W.,(2003). *Statistical Methods for Survival Data Analysis*. Wiley, New York, 3rd edition.

LUZ, M. ZEA, SILVA RODRIGO, B., BOURGUIGNON, M., ANDREA, S., GAUSS COREIRO, M., (2012). The Beta Exponentiated Pareto Distribution with Application to Bladder Cancer Susceptibility. *International Journal of Statistics and Probability*, 1(2):8–19.

MOUSA, M., JAHEEN, Z., (2002). Statistical inference for the burr model based on progressively censored data. *An International Computers and Mathematics with Applications,*, 43:1441–1449.

NG, K., CHAN, P. S.,BALAKRISHAN, N.,(2002). Estimation of parameters from progressively censored data using an algorithm. *Computational Statistics and Data Analysis*, 39:371–386.

SHAWKY, A. I., HANNA, H. ABU-ZINADAH.,(2009). Exponentiated Pareto distribution: Different method of estimations. *Int. J.Contemp. Math. Sciences*, 4(14): 677–693.

VASISHTA, N., SMITH, A. F. M., UPADHYAY, S. K., (2001). Bayes inference in life testing and reliability via Markov chain Monte Carlo simulation. *Sankhya*, A 63(1):15–20.

WU, S. J., CHANG, C. T. (2002). Parameter estimations based on exponential progressive type II censored with binomial removals. *International Journal of Information and Management Sciences*, 13:37–46.

WU, S. J., CHANG, C. T., (2003). Inference in the Pareto distribution based on progressive type II censoring with random removales. *Journal of Applied Statistics*, 30:163–172.

YANG, C., TSE, S. K., YUEN, H. K., (2000). Statistical analysis of Weibull distributed life time data under type II progressive censoring with binomial removals. *Jounal of Applied Statistics*, 27:1033–1043.

YUEN, H. K., TSE, S. K., (1996). Parameters estimation for Weibull distribution under progressive censoring with random removal. *Journal Statis. Comput. Simul*, 55:57–71.

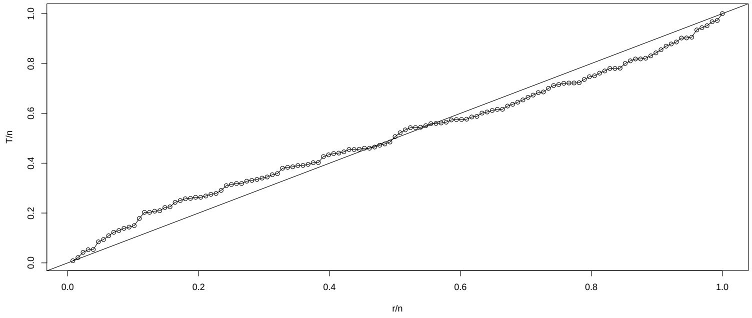


FIGURE 2. TTT plot for the remission times (in months) of 128 bladder cancer patients

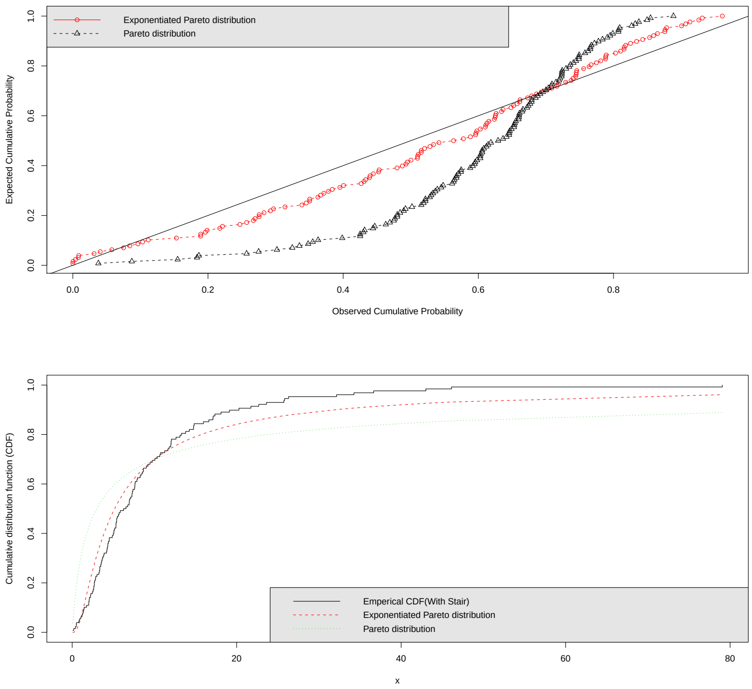


FIGURE 3. Upper graph represents the probability plot and lower graph shows CDF plot for the remission times (in months) of 128 bladder cancer patients.

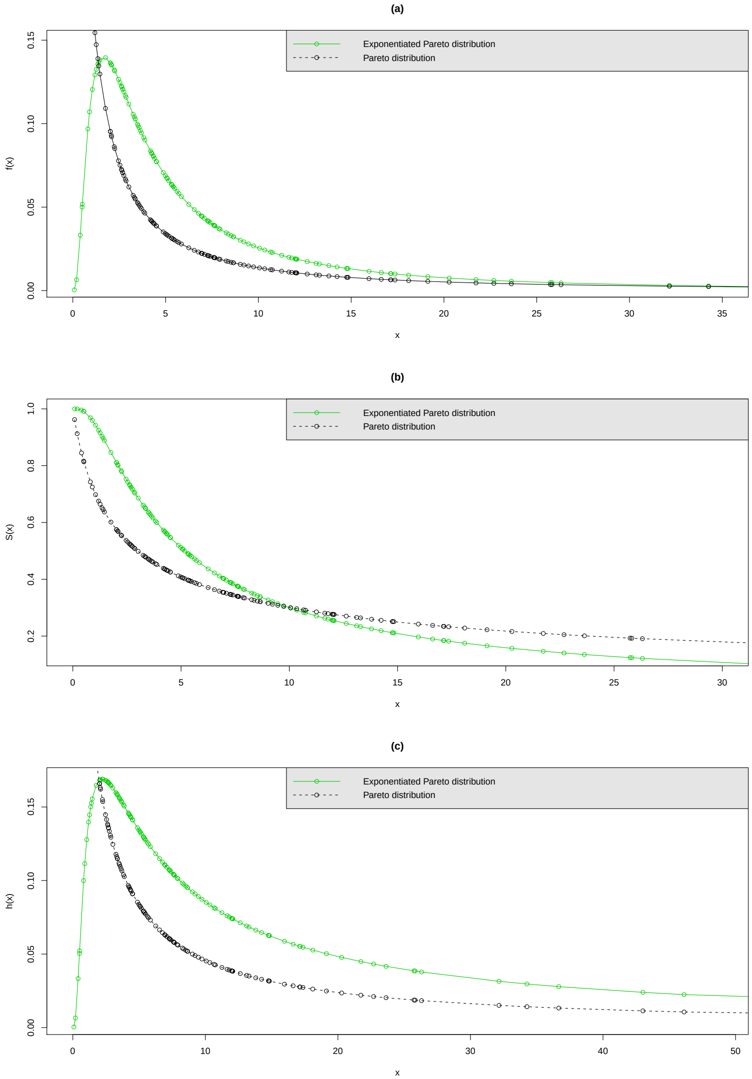


Figure 4. Estimated probability density, survival and hazard functions for the remission times (in months) of 128 bladder cancer patients.

Table 1. Risks of estimators of α under different losses for fixed $\alpha = 2$, $\theta = 0.5$, $\nu_1 = 0$, $\lambda_1 = 0$, $\nu_2 = 0$, $\lambda_2 = 0$, n= 15, $\delta = \pm 0.5$

m	$R_S(\hat{\alpha}_M)$	$R_S(\hat{\alpha}_S)$	$R_S(\hat{\alpha}_E)$	$R_E(\hat{\alpha}_M)$	$R_E(\hat{\alpha}_S)$	$R_E(\hat{\alpha}_E)$
9	0.70526	0.41143	.021448	0.01624	0.010061	0.000743
	(0.71638)	(0.42280)	(0.19942)	(0.01462)	(0.00937)	(0.00484)
12	0.72537	0.50554	0.06051	0.01665	0.01211	0.001691
	(0.72006)	(0.50812)	(0.33002)	(0.01468)	(0.01096)	(0.00757)
15	0.72402	0.54560	0.13901	0.01662	0.01296	0.00372
	(0.71659)	(0.53729)	(0.39171)	(0.01462)	(0.01150)	(0.00879)

Table 2. Risks of estimators of α under different losses for fixed $\alpha = 2$, $\theta = 0.5$, $\nu_1 = 4$, $\lambda_1 = 2$, n=15, $\nu_2 = 0.25$, $\lambda_2 = .5$, $\delta = \pm 0.5$

m	$R_S(\hat{\alpha}_M)$	$R_S(\hat{\alpha}_S)$	$R_S(\hat{\alpha}_E)$	$R_E(\hat{\alpha}_M)$	$R_E(\hat{\alpha}_S)$	$R_E(\hat{\alpha}_E)$
9	0.71061	0.07906	0.00245	0.01635	0.00220	.00074
	(0.70607)	(0.07960)	(0.04193)	(0.01445)	(0.00212)	(0.00116)
12	0.72158	0.13124	0.02286	0.01657	0.00355	0.00067
	(0.70378)	(0.12874)	(0.08569)	(0.01441)	(0.00330)	(0.00227)
15	0.70846	0.16820	0.04939	0.01631	0.00447	0.001406
	(0.71687)	(0.16936)	(0.12436)	(0.01463)	(0.00422)	(0.00319)

Table 3. Risks of estimators of α under different losses for fixed $\alpha = 2$, $\theta = 0.5$, $\nu_1 = 1$, $\lambda_1 = 0.5$, n=15, $\nu_2 = 0.0625$, $\lambda_2 = 0.125$, n=15, $\delta = \pm 0.5$

m	$R_S(\hat{\alpha}_M)$	$R_S(\hat{\alpha}_S)$	$R_S(\hat{\alpha}_E)$	$R_E(\hat{\alpha}_M)$	$R_E(\hat{\alpha}_S)$	$R_E(\hat{\alpha}_E)$
9	0.69846	0.23747	0.00534	0.01610	0.00613	0.00016
	(0.71844)	(0.24831)	(0.12497)	(0.01465)	(0.00592)	(0.00319)
12	0.71230	0.32581	0.04418	0.01638	0.00818	0.00125
	(0.71827)	(0.32775)	(0.21453)	(0.01465)	(0.00755)	(0.00520)
15	0.72212	0.38485	0.10290	0.01658	0.00951	0.00281
	(0.71713)	(0.38264)	(0.27947)	(0.01464)	(0.00864)	(0.00657)

Table 4. Risks of estimators of θ under different losses for fixed $\alpha = 2$, $\theta = 0.5$, $\nu_1 = 0$, $\lambda_1 = 0$, $\nu_2 = 0$, $\lambda_2 = 0$, n= 15, $\delta = \pm 0.5$

m	$R_S(\hat{\theta}_M)$	$R_S(\hat{\theta}_S)$	$R_S(\hat{\theta}_E)$	$R_E(\hat{\theta}_M)$	$R_E(\hat{\theta}_S)$	$R_E(\hat{\theta}_E)$
9	0.02434	0.01574	0.00451	0.00957	0.00642	0.00199
	(0.02350)	(0.01526)	(0.01082)	(0.008449)	(0.00577)	(0.00424)
12	0.02336	0.01687	0.00621	0.00921	0.00685	0.00269
	(0.02336)	(0.01717)	(0.01299)	(0.00834)	(0.00641)	(0.00500)
15	0.02256	0.01712	0.00699	0.00893	0.00693	0.00301
	(0.02305)	(0.01739)	(0.01351)	(0.00831)	(0.00649)	(0.00518)

Table 5. Risks of estimators of θ under different losses for fixed $\alpha = 2$, $\theta = 0.5$, $\nu_1 = 4$, $\lambda_1 = 2$, n=15, $\nu_2 = 0.25$, $\lambda_2 = .5$, $\delta = \pm 0.5$

m	$R_S(\hat{\theta}_M)$	$R_S(\hat{\theta}_S)$	$R_S(\hat{\theta}_E)$	$R_E(\hat{\theta}_M)$	$R_E(\hat{\theta}_S)$	$R_E(\hat{\theta}_E)$
9	0.02402	0.009106	0.00303	0.00945	0.00387	0.00136
	(0.02359)	(0.00911)	(0.00670)	(0.008478)	(0.00364)	(0.00275)
12	0.0230799	0.00939	0.00339	0.00912	0.00399	0.00151
	(0.02405)	(0.01004)	(0.00761)	(0.00861)	(0.00397)	(0.00309)
15	0.02300	0.00964	0.00375	0.00909	0.00409	0.00167
	(0.0224)	(0.00932)	(0.00712)	(0.00811)	(0.00371)	(0.00290)

Table 6. Risks of estimators of θ under different losses for fixed $\alpha = 2$, $\theta = 0.5$, $\nu_1 = 1$, $\lambda_1 = 0.5$, n=15, $\nu_2 = 0.0625$, $\lambda_2 = 0.125$, $\delta = \pm 0.5$

m	$R_S(\hat{\theta}_M)$	$R_S(\hat{\theta}_S)$	$R_S(\hat{\theta}_E)$	$R_E(\hat{\theta}_M)$	$R_E(\hat{\theta}_S)$	$R_E(\hat{\theta}_E)$
9	0.02446	0.01318	0.004151	0.00961	0.005469	0.00184
	(0.02434)	(0.0131)	(0.00953)	(0.008710)	(0.00505)	(0.0037)
12	0.02321	0.01385	0.00510	0.00916	0.00571	0.00223
	(0.02338)	(0.01392)	(0.0105)	(0.00841)	(0.00533)	(0.00414)
15	0.02298	0.01438	0.00582	0.009078	0.00591	0.00253
	(0.02272)	(0.01419)	(0.01095)	(0.008199)	(0.005418)	(0.00429)

Table 7. Expected experiment time $E(X_m)$ and δ_{REET} (in the brackets) for $(\alpha, \theta) = (2, 0.5)$ under PT-II CBR

(n, m)	p=0.05	p=0.1	p=0.3	p=0.5	p=0.7	p=0.9
15	4.0340	4.0340	4.0340	4.0340	4.0340	4.0340
(15,14)	2.5476	2.8857	3.4785	3.6652	3.7676	3.9679
	(0.6315)	(0.7154)	(0.8623)	(0.9086)	(0.9340)	(0.9836)
(15,13)	1.6483	2.1813	3.2342	3.5110	3.6181	3.7663
	(0.4086)	(0.5407)	(0.8017)	(0.8704)	(0.8969)	(0.9336)
(15,12)	0.7518	1.1819	2.6107	3.3734	3.3846	3.7347
	(0.1864)	(0.2930)	(0.6472)	(0.8362)	(0.8390)	(0.9258)
(15,11)	0.5354	0.8059	2.4429	2.6319	2.8624	3.6419
	(0.1327)	(0.1998)	(0.6056)	(0.6524)	(0.7096)	(0.9028)
(15,10)	0.5104	0.7854	2.2901	2.7439	3.0896	3.6191
	(0.1265)	(0.1947)	(0.5677)	(0.6802)	(0.7659)	(0.8971)
12	3.4955	3.4955	3.4955	3.4955	3.4955	3.4955
(12,11)	1.9313	2.3490	3.1130	3.1289	3.4512	3.4808
	(0.5525)	(0.6720)	(0.8906)	(0.8951)	(0.9873)	(0.9958)
(12,10)	1.2064	1.6751	2.7961	2.8736	2.8852	2.9184
	(0.3451)	(0.4792)	(0.7999)	(0.8221)	(0.8254)	(0.8349)
(12,9)	0.7102	1.1478	2.2872	2.6235	2.7740	3.0725
	(0.2032)	(0.3284)	(0.6543)	(0.7505)	(0.7936)	(0.8790)
(12,8)	0.4466	0.6991	1.8043	2.2604	2.5090	2.5323
	(0.1278)	(0.2000)	(0.5162)	(0.6467)	(0.7178)	(0.7244)
9	2.8353	2.8353	2.8353	2.8353	2.8353	2.8353
(9,8)	1.3739	1.7627	2.4083	2.5943	2.6157	2.7586
	(0.4845)	(0.6217)	(0.8494)	(0.9150)	(0.9225)	(0.9729)
(9,7)	0.7664	1.0396	1.8727	2.2745	2.3097	2.4699
	(0.2703)	(0.3667)	(0.6605)	(0.8022)	(0.8146)	(0.8711)
(9,6)	0.4260	0.5567	1.2967	1.8170	2.0244	2.0305
	(0.1503)	(0.1964)	(0.4573)	(0.6408)	(0.7140)	(0.7161)

Table 8. Goodness of fit for the remission times (months) of bladder cancer data

Distribution	AIC	BIC	K-S Statistics	P-value	Log-likelihood
EPD	856.6102	862.3142	0.1016	0.5239	-426.3051
Pareto	948.0433	953.7473	0.3125	7.45E-06	-472.0216

Table 9. Bayes and ML estimates, CI and HPD intervals for α with fixed $n = 128$ and $p = 0.3$ under PT-II CBR for the remission times (months) of bladder cancer data for different censoring schemes $(S_{n:m})$.

$S_{n:m}$	MLE	Bayes Estimates(MCMC)			Interval			
		SELF	GELF		95% CI		95% HPD	
			$\delta = 1.5$	$\delta = -1.5$	α_{L^c}	α_{U^c}	α_{L^h}	α_{U^h}
51	3.5109	2.9787	2.9787	2.9787	2.1359	4.8859	2.9635	2.9925
64	2.9321	2.9273	2.9273	2.9273	1.8982	3.9659	2.9141	2.9398
77	3.5733	3.5675	3.5674	3.5675	2.3763	4.7704	3.5528	3.5829
90	3.6080	3.6033	3.6033	3.6033	2.4537	4.7624	3.5882	3.6168
102	4.0606	4.0564	4.0563	4.0564	2.8117	5.3096	4.0416	4.0724
128	4.6574	4.6574	4.6573	4.6329	3.3135	6.0013	4.6409	4.6740

TABLE 10. Bayes and ML estimates, CI and HPD intervals for θ with fixed $n = 128$ and $p = 0.3$ under PT-II CBR for the remission times (months) of bladder cancer data for different censoring schemes $(S_{n:m})$.

$S_{n:m}$	MLE	Bayes Estimates(MCMC)			Interval			
		SELF	GELF		95% CI		95% HPD	
			$\delta = 1.5$	$\delta = -1.5$	θ_{L^c}	θ_{U^c}	θ_{L^h}	θ_{U^h}
51	0.6885	0.6895	0.6895	0.6895	0.4883	0.8885	0.68512	0.6937
64	0.7123	0.7112	0.7112	0.7112	0.5195	0.9051	0.7076	0.7152
77	0.8203	0.8193	0.8193	0.8193	0.6373	1.0034	0.8157	0.8227
90	0.8824	0.8816	0.8816	0.8816	0.7023	1.0624	0.8779	0.8849
102	0.9559	0.9554	0.9554	0.9554	0.7823	1.1296	0.9519	0.9586
128	1.0877	1.0877	1.0877	1.0845	0.9194	1.2559	1.0845	1.0912

TABLE 11. PT-II CBR under different censoring schemes $S_{n:m}$ for fixed $n = 128$ and $p = 0.3$ for the remission times (months) of bladder cancer data

$S_{m:n}$	i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$S_{51:128}$	y_i	0.08	0.2	0.4	0.5	0.81	0.9	1.05	1.26	1.35	1.4	1.46	1.76	2.02	2.02	2.09	2.23	2.46
	R_i	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0	1	0
	y_i	2.54	2.62	2.64	2.69	2.75	2.83	2.87	3.02	3.31	3.36	3.36	3.48	3.57	3.64	3.7	3.88	4.18
	R_i	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0
	y_i	4.23	4.26	4.33	4.4	4.5	4.51	4.87	4.98	5.06	5.09	5.17	5.32	5.32	5.34	5.41	5.49	5.62
	R_i	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	67
$S_{64:128}$	y_i	0.08	0.2	0.4	0.5	0.51	0.81	0.9	1.05	1.19	1.26	1.35	1.4	1.46	1.76	2.02	2.02	2.07
	R_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	y_i	2.09	2.23	2.26	2.46	2.54	2.62	2.64	2.69	2.75	3.02	3.31	3.36	3.48	3.52	3.57	3.64	3.7
	R_i	0	0	0	0	0	0	0	1	2	1	1	0	0	0	0	0	0
	y_i	3.82	3.88	4.18	4.26	4.33	4.34	4.4	4.5	4.51	4.87	5.06	5.09	5.32	5.32	5.34	5.41	5.49
	R_i	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	1	0
	y_i	5.62	5.71	5.85	6.25	6.54	6.76	6.93	7.09	7.26	7.28	7.32	7.39	7.62				
	R_i	0	0	0	0	0	0	2	0	0	0	0	1	52				
$S_{77:128}$	y	0.08	0.2	0.4	0.5	0.81	0.9	1.05	1.26	1.35	1.4	1.46	1.76	2.02	2.02	2.07	2.09	2.23
	R_i	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1
	y_i	2.46	2.62	2.64	2.69	2.83	2.87	3.02	3.25	3.31	3.36	3.36	3.52	3.64	3.7	3.82	3.88	4.18
	R_i	1	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0
	y_i	4.23	4.26	4.33	4.5	4.51	4.87	4.98	5.06	5.09	5.17	5.32	5.34	5.41	5.41	5.49	5.62	6.25
	R_i	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	2	1
	y_i	6.76	6.93	6.94	6.97	7.09	7.26	7.28	7.32	7.39	7.59	7.62	7.63	7.66	7.87	7.93	8.26	8.37
	R_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	y_i	8.65	8.66	9.02	9.47	9.74	10.06	10.34	10.66	10.75								
	R_i	0	0	1	0	0	0	0	0	35								
$S_{90:128}$	y_i	0.08	0.2	0.4	0.5	0.51	0.81	0.9	1.05	1.19	1.26	1.35	1.4	1.46	1.76	2.02	2.02	2.07
	R_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	y_i	2.23	2.26	2.46	2.54	2.62	2.64	2.69	2.69	2.75	2.83	2.87	3.02	3.25	3.31	3.36	3.36	3.48
	R_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
	y_i	3.57	3.64	3.7	3.82	3.88	4.18	4.23	4.26	4.33	4.34	4.4	4.5	4.51	4.87	4.98	5.06	5.09
	R_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	y_i	5.17	5.32	5.32	5.34	5.41	5.41	5.62	5.71	5.85	6.25	6.54	6.76	6.93	6.94	6.97	7.09	7.26
	R_i	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
	y_i	7.28	7.32	7.39	7.59	7.62	7.63	7.87	7.93	8.26	8.37	8.53	8.65	8.66	9.02	9.22	9.74	10.06
	R_i	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
	y_i	10.34	10.66	11.64	11.79	11.98												
	R_i	0	2	0	0	31												
$S_{102:128}$	y_i	0.08	0.2	0.4	0.5	0.51	0.81	0.9	1.19	1.26	1.35	1.4	1.46	1.76	2.02	2.07	2.09	2.23
	R_i	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
	y_i	2.26	2.46	2.54	2.62	2.64	2.69	2.69	2.75	2.87	3.02	3.25	3.31	3.36	3.36	3.48	3.52	3.57
	R_i	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
	y_i	3.7	3.82	3.88	4.18	4.23	4.26	4.33	4.34	4.4	4.5	4.51	4.87	4.98	5.06	5.09	5.17	5.32
	R_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	y_i	5.32	5.34	5.41	5.41	5.49	5.62	5.71	5.85	6.25	6.76	6.93	6.94	6.97	7.09	7.28	7.32	7.39
	R_i	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
	y_i	7.59	7.63	7.66	7.87	7.93	8.26	8.37	8.53	8.65	8.66	9.02	9.22	9.47	9.74	10.06	10.34	10.66
	R_i	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	y_i	10.75	11.25	11.64	11.79	11.98	12.02	12.07	12.63	13.11	13.29	13.8	14.24	14.76	14.77	14.83	15.96	16.62
	R_i	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	18

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 427–436

THE POSITION OF THE WIG INDEX IN COMPARISON WITH SELECTED MARKET INDICES IN BOOM AND BUST PERIODS

Anna Czapkiewicz[1], Beata Basiura[2]

ABSTRACT

The main aim of this work is to discover the differences between the rank of Polish stock market in the boom and the bust cycles. The data of the daily stock market returns for the twenty three major international indices from Europe, America and Asia are used in the research. Two boom and two bust periods are considered. The correlation coefficient obtained from Copula-GARCH model is a similarity measure between the considered indices returns. The cluster analysis carried on for these series in the boom and bust the cycles allows us to find the differences in the market behaviour.

Key words: clustering stock indices, dependence parameter, Copula-GARCH model.

1. Introduction

Finding similarities between world financial markets has been one of the primary intention amongst investigations. Practitioners are interested in identifying these similarities to assess investment risk. Knowledge about market relationships enables us to diversify this risk. To gain insight into the internal relationship between financial time series the cluster analysis has proved useful, producing a set of markets grouped according to a given measure of similarity in their behaviour. However, in the case of the clustering time series we encounter difficulties with the choice of an appropriate measure which could be used as a measure of similarity between the indices returns and takes into account the character of the considered series. It is known that the function of the Pearson correlation coefficient as a measure of similarity between pairs of stock returns (Mantegna, 1999; Bonanno et al., 2001) is not a satisfactory measure of

[1] Faculty of Management, AGH University of Science and Technology, Krakow, Poland, A. Mickiewicza 30 Ave., 30-059. E-mail: gzrembie@cyf-kr.edu.pl.

[2] Faculty of Management, AGH University of Science and Technology, Krakow, Poland, A. Mickiewicza 30 Ave., 30-059. E-mail: bbasiura@zarz.agh.edu.pl.

dependence. Without the multivariate normality assumption, two pairs of markets can have equal linear correlation coefficient while they can still differ in terms of dependence structure.

A useful tool to describe dependence between time series is the application of copula function to model the multivariate distribution (Embrechts et al. 2001, 2003). Copulas are useful to apply because they allow us to separate the dependence properties of the data from their marginal properties and to construct multivariate models with marginal distributions of an arbitrary form. Some of them are appropriate for financial markets. The most popular are t-Student and Joe-Clayton copulas. The first one is recommended, for example, by Mashal, Zevi (2002) and Breymann et al. (2003). The AR(1)-GARCH(1,1) model with skewed t-Student conditional distribution is quite satisfactory for describing the indices returns behaviour and may be applied as the marginal. The elements of the correlation matrix obtained from t-Student copula may be considered as the similarity measure between data. Having the matrix of distances based on these the similarity measure the Ward algorithm (Ward 1968) may be used to cluster the indices into the similar groups. The applicability of such a methodology for grouping global markets was presented in the work by Czapkiewicz, Basiura (Czapkiewicz, Basiura 2010).

This study concerns the determination of the Poland's position in comparison with the selected market indices. Empirical study covers two boom periods and two bust periods from June 2003 to March 2012. The possibility of difference in grouping of the markets in the boom and the bust periods is taken into consideration. It is anticipated that the negative moods in the stock markets strongly influence the other markets than the positive ones. The aim of this empirical work is to search for the differences in the relation strength of Polish market with other markets in the boom and bust periods. Furthermore, the clustering of the twenty three markets is carried on in these periods. These periods are defined according to WIG and WIG 20 indices behaviour.

2. The model

2.1. The distributions of returns

The advantage of using copulas, as mentioned in the introduction, stems from the fact that marginal distributions can be separated from the underlying dependency structure. Many models have been proposed to describe the dynamics of return. In this paper we consider the univariate AR(1)-GARCH(1,1) model. It is defined as follows:

$$y_t = \mu + \alpha y_{t-1} + \varepsilon_t, \quad \varepsilon_t = \sqrt{h_t}\eta_t$$
$$h_t = a_0 + a_1\varepsilon_{t-1}^2 + a_2 h_{t-1}, \quad \eta_t \sim iid(0,1)$$

In the above equation y_t denotes the daily return of stock market index. Scrutiny of daily returns led to the introduction of fat-tailed distributions for this residuals. Fat-tails are not the only problem in the context of conditional distribution, the skewness can also be noticed. That is why the skewed distribution was considered as the conditional distribution in the above process.

2.2. The copula function

The copula function is a multivariate distribution defined on the unit cube $[0,1]^d$, with uniformly distributed margins. The function $C: [0,1]^d \to [0,1]$ is a d-dimensional copula if it satisfies the following properties:

1. For all $u_i \in [0,1]$, $C(1,\ldots,1,u_i,1,\ldots,1) = u_i$.
2. For all $u \in [0,1]^d, C(u_1,\ldots,u_d) = 0$, if at least one coordinate $u_i = 0$.
3. C is d-increasing.

The importance of the copula function stems from the fact that it captures the dependence structure of the multivariate distribution. According to Sklar's theorem (Sklar, 1959) a given d-dimensional distribution function F with margins $F_1,\ldots,F_d$ can be presented as:

$$F(x) = C\big(F_1(x_1),\ldots,F_d(x_d)\big).$$

When $F(x)$ is a multivariate continuous distribution function of a random vector $X \in R^d$ and $F_i(x_i)$ are continuous margins the copula is uniquely determined.

The copula used in the empirical part is the t-Student copula:

$$C(u_1,\ldots,u_d) = t_{\Sigma,\eta}\left(t_\eta^{-1}(u_1),\ldots t_\eta^{-1}(u_d)\right)$$

where t_η is the t-Student's cumulative distribution with η degrees of freedom and $t_{\Sigma,\eta}$ is the t-Student's cumulative distribution with η degrees of freedom and the correlation matrix Σ. The bivariate case of t-Student copula is given by:

$$C(u_i,u_j;\rho_{ij}) = \int_{-\infty}^{t_\eta^{-1}(u_i)} \int_{-\infty}^{t_\eta^{-1}(u_j)} \frac{1}{2\pi\sqrt{1-\rho^2}}\left(1+\frac{s^2-2\rho_{ij}st+t^2}{\eta(1-\rho_{ij}{}^2)}\right)^{-\frac{\eta+2}{2}} ds\,dt$$

where ρ_{ij} is the correlation ratio.

The Copula-GARCH model may be estimated by maximum likelihood method. The IFM strategy (Shih, Louis, 1995; Joe, Xu, 1996) is used for this purpose in the empirical work. IFM proceeds in the two steps. Firstly, the parameter estimates of margins distribution are obtained, secondly, the estimate of copula dependence parameter[1] is calculated. Under some regularity conditions, Patton (2006) shows that the IFM procedure yields consistent and asymptotically normal estimates.

[1] It is the commonly used name of the parameter ρ_{ij}, because it measures dependence between the marginals (cf. P. K. Trivedi and D. M. Zimmer (2005)).

3. Empirical study

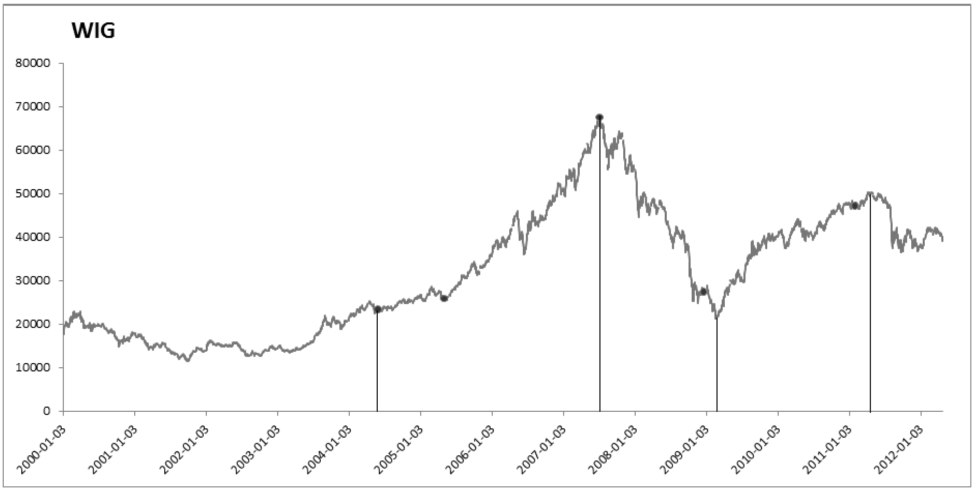
3.1. The data

The study covers the period from June 2003 to March 2012. During this period four sub-periods are extracted. The choice of these sub-periods is related to WIG and WIG 20 trading. The first - from July 2003 to June 2007 is a boom sub-period; the second - from August 2007 to February 2009 - the bust one; the third - a post crisis sub-period, where both indices WIG and WIG20 rise again (from February 2009 to July 2011). The last sub-period (from July 2011 to January 2012) is determined as the bust one. Figure 1 and Figure 2 present the trading WIG index and the trading WIG20 index in the searching sub-periods.

The relationships between some selected stock indices are investigated separately in each sub-period. We will compare the stock markets in the following countries on the basis of the indices given in brackets:

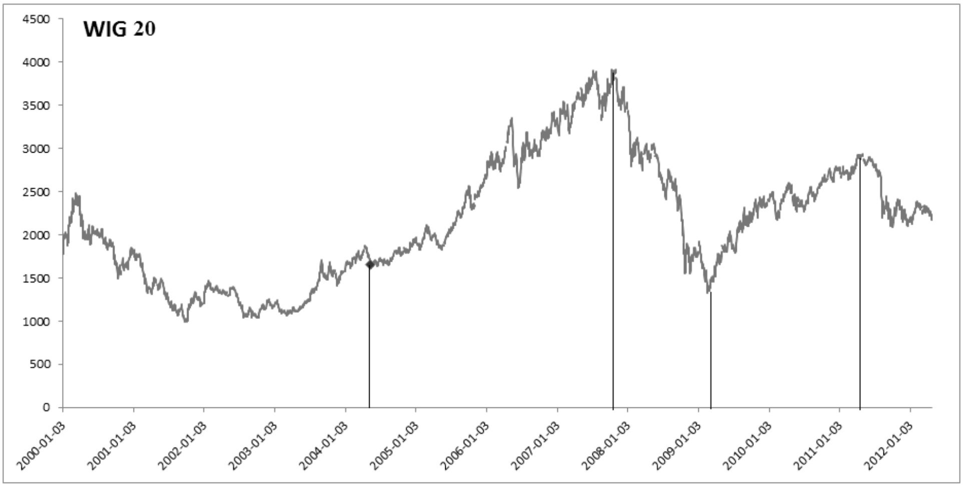
Poland (WIG and WIG20), Austria (ATX), Euronext Brussels (BEL20), Bulgaria (SOFIX), Canada (TSX), China (HSI), Czech Republic (PX), Finland (HEX), France (CAC40), Germany (DAX), Hungary (BUX), Japan (NIKKEI), Norway (OSE), Romania (BET), Russia (RTS), Slovakia (SAX), South Korea (KOSPI), Spain (IBEX), Switzerland (SMI), the Netherlands (AEX), the UK (FTM), the USA (DJIA), Turkey (ISE).

Figure 1. The trading WIG index from January 2002 to March 2012.



Source: gpw.pl, April 2012.

Figure 2. The trading WIG20 index from January 2002 to March 2012.



Source: gpw.pl, April 2012.

The investigation covers returns of market indices from the world. It includes the markets deemed as emerging as well as those already developed of North America, Europe and Asia. In the case of the USA, the DJIA index is taken into consideration. In addition, the BEL20, the benchmark stock market index of Euronext Brussels, is included in the research.

It should be noted that the indices selected for testing represent wide or narrow market. So, to meet these constrains two indices: WIG and WIG20 are chosen as representatives of Polish market.

The daily frequency data are taken into study. Missing data are filled by linear interpolation from the preceding to the following missing quotations. The return of indices is defined as $r_t = \ln(P_t / P_{t-1})$ where P_t is an adjusted index value at period *t*.

Some tests for conditional heteroskedasticity and autocorrelation are performed. The results of Engle test led us to assume that the choice of the GARCH model is justified while the Ljung-Box test results indicate the possibility of the autocorrelation presence. For all considered cases GARCH effect and the autocorrelation exist. These results are the reason for the introduction of AR(1)-GARCH(1,1) model to describe the indices returns behaviour.

3.2. Estimation of the multivariate model

In a preliminary step of our empirical work, we investigate the structure of the univariate marginal returns. The AR(1)-GARCH(1,1) model with the skewed *t*-Student's conditional distribution is considered to describe returns modelling.

Thus, the procedure of testing the goodness-of-fit is carried out. For the testing purposes, we follow the procedure described in Diebold et al. (1998). If a marginal distribution is correctly specified, the margins denoting the transformed standardized AR(1) - GARCH(1,1) residuals should be iid Uniform $(0,1)$. For the most of the analyzed time series the test results confirm the correctness of the chosen model.

Prior to the main study, a preliminary analysis of relationship between WIG and WIG20 is carried out. The study is conducted using data from the whole sample. The estimated parameter of t-Student copula ($\rho = 0.98$) indicates a very strong correlation between these two indexes.

This strong relationship makes very small differences between the parameters defining the dependence between the Polish market and other markets if we consider the WIG20 index instead of the WIG index. So, taking pairs of the Polish index with an index representing the market of another country, the parameters of the bivariate t-Student copula are estimated. If an index represents a narrow market the WIG20 index is a representative of the Polish market.

Table 1 presents the correlation coefficients obtained from t-Student copula for Polish index with other indices considered in the boom and bust sub-periods. According to intuition, one would expect the dependencies with other markets should be greater during the boom sub-periods than during the bust sub-periods.

Table 1. The correlation coefficients obtained from t-Student copula for Polish index with other indices considered in the boom and bust periods

Country	1^{th} boom	1^{th} bust	2^{nd} boom	2^{nd} bust
Austria	0.39	**0.67**	**0.61**	**0.72**
Belgium	0.43	**0.66**	**0.63**	**0.76**
Bulgaria	0.05	0.19	0.26	0.27
Canada	0.27	0.36	0.43	**0.57**
China	0.08	0.15	0.21	0.15
Czech Republic	**0.46**	**0.68**	**0.66**	**0.67**
Finland	0.43	**0.60**	**0.63**	**0.77**
France	**0.44**	**0.67**	**0.68**	**0.77**
Germany	0.40	**0.65**	**0.62**	**0.77**
Hungary	**0.54**	**0.64**	**0.64**	**0.64**
Japan	0.28	0.35	0.27	0.34
Norway	0.43	**0.55**	**0.65**	**0.73**
Romania	0.06	0.41	0.44	**0.49**
Russia	0.42	**0.56**	**0.64**	**0.68**
Slovakia	0.04	-0.05	-0.03	0.04
South Korea	0.33	0.34	0.35	0.43
Spain	0.43	**0.63**	**0.61**	**0.68**
Switzerland	0.40	**0.62**	**0.61**	**0.70**
the Netherlands	**0.46**	**0.66**	**0.68**	**0.76**
the UK	**0.48**	**0.67**	**0.66**	**0.75**
the USA	0.21	0.34	0.48	**0.60**
Turkey	0.36	**0.66**	0.55	**0.68**

The results confirm our prediction. The analysis of the estimated correlation coefficients indicates that the relationship of Polish market with other markets is stronger during the bust sub-periods than during the boom ones. The beginning of the first considered sub-period (as a reminder it is from June 2003 to July 2007) precedes the date of Poland's entry into the European Union. Under this sub-period study the strongest correlation is observed only in the case of the Hungarian index. The relations with markets of the Eurozone are relatively weak. The correlation between Polish and Russian indices is similar to the correlations between Polish index and indices of Western Europe. It is noted that no significant relations of the Polish market with markets of Slovakia, Romania, Bulgaria or China exist.

The stronger dependence between Poland and other markets is observed in the first bust sub-period. The Polish index is strongly correlated with indices of Czech Republic, Austria, Belgium, France, Germany, the Netherlands, the UK and Turkey (coefficients are greater than 0.65).

In the second boom sub-period, quite strong correlation is observed between the Polish index and the indices of the Netherlands, France, Czech Republic, the UK, Norway, Hungary, Russia, Belgium, Finland, Germany, Austria, Spain and Switzerland (correlation coefficients are greater than 0.60).

In the second bust sub-period the strength of relationship of Poland with other markets increases. In this sub-period there are the strongest dependences between the Polish and other studied markets. At that time Romania and Bulgaria are in the European Union, so a stronger correlation with those markets indices is found than in the previous sub-periods. At the same time Poland consolidates its position in the European Union so much stronger relationship of Poland with the Eurozone countries is observed.

It is worth noting that in the second boom sub-period the correlations between Poland and other markets are not significantly lower than under the previous boom sub-period study. So, one might conclude that the strength of relationships between markets increases, regardless of the economic situation.

For all studied periods the weakest correlation is noted between Polish and Slovakian indices. It seems that Slovakian market is not linked with Polish market at all.

A more complete picture of links between the markets is obtained using the clustering method based on these determined parameters. Although the correlation coefficient might be relatively high, the index might belong to another cluster. In the following part of the empirical study the results of grouping of indices in the considered periods are presented. The Ward algorithm is adopted for this purpose with the dissimilarity measure $d_{ij} = 1 - \rho_{ij}$. Figure 3 presents the dendrograms for two boom periods. Figure 3a shows the results of grouping of indices for the first boom period, while Figure 3b presents clustering results for the second boom one. A similar analysis is performed for data from the two bust periods. Figure 4a shows the clustering results for the period of global crisis, while Figure 4b presents the results for the data of the last bust period.

Figure 3. The dendrograms for market indices in the two periods of boom; a) from July 2003 to June 2007; b) from February 2009 to July 2011

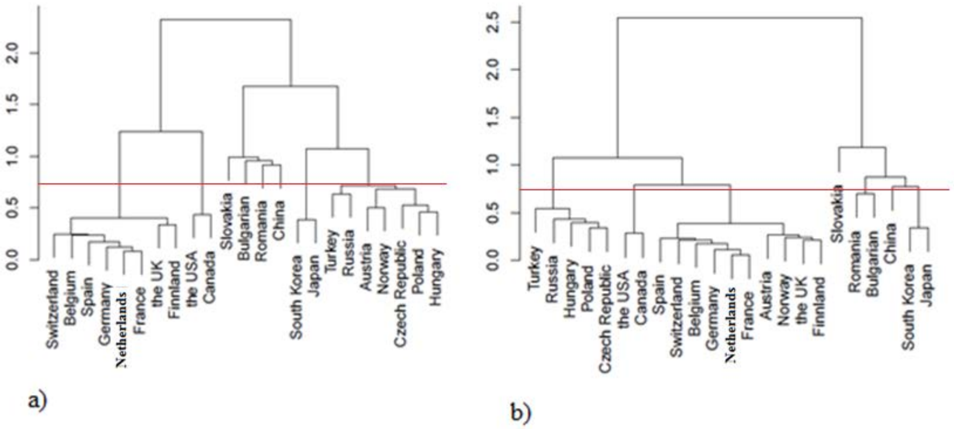
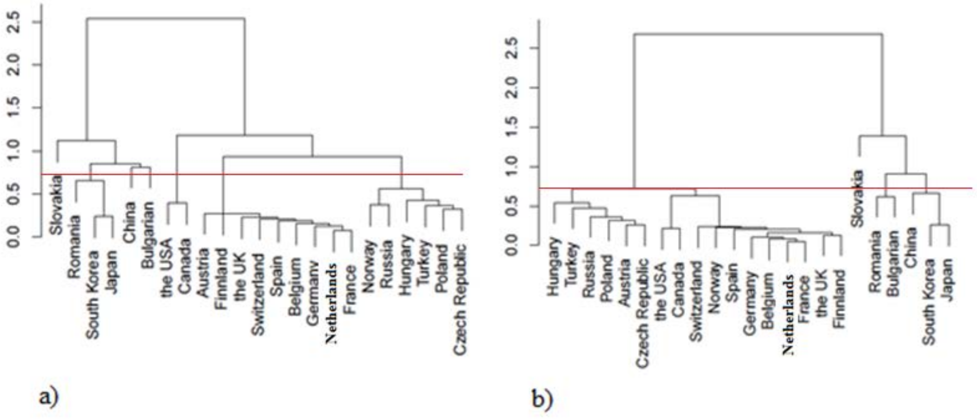


Figure 4. The dendrograms for market indices in the two periods of bust; a) from August 2007 to February 2009; b) from July 2011 to January 2012



For this research it is very important that the group included data obtained with a strong dependence (ρ_{ij} close to one). So, it is assumed that the tree should be cut for $d = 0{,}75$. It can be noticed that clusters are being varied during the considered periods. In the first boom period there are four groups. First group concentrate markets from the Eurozone: Switzerland, Belgium, Spain, Germany, the Netherlands and France. The markets of Finland and the UK stand out against the background of European markets. The USA and Canada markets belong to one group. The Japan market is grouped with South Korea one. A separate group consists of markets from North, Central and Eastern Europe. There are three abstracted subgroups: first - Turkey and Russia, second - Austria and Norway -

and last - Poland, Hungary and Czech Republic. The isolated markets are: Slovakia, Bulgaria, Romania and China. When the dendrogram from the second boom period is analyzed similar grouping can be noticed.

In the first bust sub-period, the indices of Austria, Finland, the UK, Switzerland, Spain, Belgium, Germany, the Netherlands and France are in the one subgroup, whereas indices of the Norway, Russia, Hungary, Turkey, Poland and Czech Republic are in the other. The rest of the analyzed indices form separate clusters. Similar grouping is observed in the second bust period.

The Polish index is in the same subset as Hungarian, Czech Republic, Turkish and Russian indices regardless of the studying sub-periods.

4. Conclusions

The purpose of this paper is to investigate the relationships of Polish market with some European markets and main markets of America and Asia. As a measure of the relationship between the markets the correlation coefficient obtained from the *t*-Student copula is used. Returns are modelled by AR(1)-GARCH(1,1) process. The study is conducted for the four sub-periods: two boom periods and two bust periods. These sub-periods are defined on the basis the WIG and WIG20 indices trading.

The empirical results indicate that the relationship of Polish index with other indices is stronger during the bust sub-periods than during the boom ones. Furthermore, it is noted that the strength of the relationship between Polish market and others increased, regardless of the situation on the stock markets. The relationships between the Polish market and other markets may be affected by many factors, of which by Poland's entry into the European Union.

As the results show the clustering methods yield different groupings depending on the considered sub-periods. The groupings in the boom sub-periods seem to be similar to each other although in the case of the second boom period, the binding to other markets took place on the lower levels (as evidenced by stronger correlation coefficients). In the bust period, relatively large number of markets seemed to be in one class. Polish index occurs in one subset with Hungarian, Czech Republic, Turkish and Russian indices, regardless of the studied sub-periods.

REFERENCES

BASIURA, B., CZAPKIEWICZ, A., (2010). *Clustering Financial Data Using Copula-GARCH Model In an Application for Main Market Stock Returns*, Statistics in Transition (New Series), Poland, Vol. 11, No. 1, pp. 25–45.

BONANNO, G., LILLO, F., MANTEGNA, R. N., (2001). *High-frequency cross-correlation in a set of stocks*, *Quantitative Finance* 1, pp. 96–104.

BREYMANN, W., DIAS ,A., EMBRECHTS, P., (2003). *Dependence Structures for Multivariate High-Frequency Data in Finance*, Quantitative Finance, 3, pp. 1–14.

DIEBOLD, F. X., GUNTHER, T. A., TAY, A. S., (1998). *Evaluating Density Forecasts with Applications to Financial Risk Management*, International Economic Review, 39(4): pp. 863–883.

EMBREECHT, P., MCNEIL, A. J., STRAUMANN, D., (2001). *Correlation and dependency in risk management: properties and pitfalls*, In: M. Dempster, H. Moffant, *Risk Management*, Cambridge University Press, New York, pp. 176–223.

EMBRECHTS, P., LINDSKOG, F., MCNEIL, A., (2003). *Modeling Dependence with Copulas and Applications to Risk Management*, In: Rachev, S.T. (Ed.), *Handbook of Heavy Tailed Distributions in Finance*, Elsevier/North-Holland, Amsterdam.

JOE, H., XU, J. J., (1996). *The estimation method of inference function for margins for multivariate models*, Technical Report, Departments of Statistics, University of British Columbia.

MASHAL, R., ZEEVI, A., (2002). *Beyond Correlation: Extreme co-movements Between Financial Assets*, Mimeo, Columbia Graduate School of Business.

MANTEGNA, R. N., (1999). *Hierarchical structure in financial markets*, *European Physical Journal* B 11, pp. 193–197.

MIRKIN, B., (2005). *Clustering for Data Mining: A Data Recovery Approach*, Boca Raton Fl., Chapman and Hall/CRC.

PATTON, A. J., (2006). Estimation of multivariate models for time series of possibly different lengths, *Journal of Applied Econometrics*, John Wiley & Sons. Ltd., 21(2): pp. 147–173.

ROSENBLATT, M., (1952). Remarks on a Mu ltivariate Transformation, *The Annals of Mathematical statistics*, 23, pp. 470–472.

R Development Core Team, (2004). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN is 3-900051-07-0 URL http://www.Rproject.org.

SHIH, J., LOUIS, T. A., (1995). Inference on the Association Parameter in Copula Models for Bivariate Survival Data, *Biometric*s, 51: pp. 1384-1399.

SKLAR, A., (1959). Fonction de Repartition a n D imension et Leur Marges, *Publications de L'Institut de Statistiques de L'Universite de Paris*, 8, pp. 229–231.

TRIVEDI, P. K., ZIMMER, D. M., (2005). *Copula Modeling: An Introduction for Practitioners*, Foundations and Trends in Econometrics, Vol. 1, No 1, pp. 1–111, ed. Now, the Essence of Knowledge.

WARD, J. H., (1963). Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association*, 58, pp. 236–244.

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 437–452

JOINT LONGITUDINAL AND SURVIVAL DATA MODELLING: AN APPLICATION IN ANTI-DIABETES DRUG THERAPEUTIC EFFECT

Atanu Bhattacharjee[1], **Dilip C. Nath**[2]

ABSTRACT

The longitudinal and survival analyses are useful tools in the exploration of drug trial data. In both cases the challenge is to deal with correlated repeated observations. Here, the joint modelling for longitudinal and survival data has been carried out via Markov chain Monte Carlo (MCMC) method in type 2 diabetes clinical trials to compare different combinations of drugs, viz. Metformin plus Pioglitazone and Gliclazide plus Pioglitazone. Despite the complexity of the model it has been found relatively easier to implement with WinBugs software. The results have been computed and compared with software R. In both types of the analyses it has been found that no estimates of treatment appear to have significant effect on the evolution of the matter of HBA1c, neither on the longitudinal part nor on the survival one. The Bayesian approach has been considered as an extended tool with classical approach for estimation of clinical trial data analysis.

Key words: random effects, semi-parametric survival model, Weibull distribution, linked sub-models.

1. Introduction

The longitudinal and survival analyses are useful tools in exploring the drug trial data. In type diabetes drug trials, the level of HBA1c is a widely used biomarker for diabetes while studying the efficacy of the drugs in patients. In drug effect comparison the level of HBA1c is used to measure over follow-up periods in clinical trials. The repeated measurements of HBA1c on the same patients give the scope to application of longitudinal and survival data analysis. The level of HBA1c is an important indicator for measuring the endogenous glucose over a period of 2-3 months by recommendation of The International Expert Committee

[1] Lecturer (Biostatistics), Division of Clinical Research and Biostatistics, Department of Cancer Registry and Epidemiology, Malabar Cancer Centre, Kerala-670103, India.
E-mail: atanustat@gmail.com.

[2] Professor (Statistics), Department of Statistics, Gauhati University, Guwahati 781014, India.

Report (2009). HBA1c is the important diagnostic parameter for type 2 diabetes by the report of American Diabetes (2010). The mean HBA1c is a powerful predictive tool to determine the diabetes complications are concluded by Lind et al. (2008) and Stratton et al. (2000). The HBA1c is positively associated with blood sugar level has been concluded by DCCT Study Group, 1995. The Bayesian approach in autoregressive longitudinal data analysis in type 2 diabetes patients of India has been explained by Nath and Bhattacharjee (2011). The Bayesian approach has been found the best choice in model variable selection by Nath and Bhattacharjee (2011). The joint model is associated with sub-models by the longitudinal and survival process measurement model concluded by Henderson et al. (2000). In the last two decades, the field of longitudinal and survival data analysis was enriched through adjusting statistical inferences on longitudinal measurements by Carlin et al. (2000), Celeux et al. (2006), Chen (2006), Schluchter (1992), DeGruttola and Tu (1994), Elashoff and Li (2008), Little (1995), Henderson et al. (2000), Hogan and Laird (1997), and many others.

In this context, the linear or random effect model is found more effective by Tsiatis et al. (1995). Li et al. (2009) proposed the joint model for longitudinal and survival data in the correlated repeated observations. Deslandes et al. (2010) concluded that the proportional cause-specific hazard model is the standard regression model of choice to compare the competing risks. However, the Cox analysis is a widely used method for the cause-specific hazard model. In this work, the joint longitudinal and survival models are applied to compare the updated mean value of HBA1c as the effect of different drug treatment.

2. Objective

The aim of this work is to compare the drug treatment effect with the result of HBA1c value during different visits in type 2 diabetes patients. The longitudinal and survival analysis is applied with prior assumption. The performance of a combined drug therapy, i.e., “Metformin with Pioglitazone” and “Gliclazide with Pioglitazone” is compared in reducing the HBA1c level. The Bayesian approach in the separate and joint modelling procedure is applied and compared to drug treatment effect in type 2 diabetes patients.

3. Methods

The linear model presented by Tsitaes et al. (1995) is

$$R_{1i}(g)=Z_{1i}+Z_{2i}(g). \qquad (1)$$

The parameter $R_{1i}(g)$ can be obtained by U_{1i} and U_{2i}, where (U_{1i}, U_{2i}) are subject-specific bivariate normal distributions with σ_1^2,σ_2^2 standard deviation. The next term R_{2i} can be segregated to

$$R_{2i}(g)=\lambda_1 Z_{1i}+\lambda_2 Z_{2i}+\lambda_3(Z_{1i}+Z_{2i})+Z_{3i}, \text{ where } Z_3\sim N(0,\sigma_3^2) \text{ and} \qquad (2)$$

where λ_1 can be taken as a coefficient.

The sequence of the response variables $Y_{i1},Y_{i2},\ldots\ldots..Y_{in}$ at times $g_{11},g_{21},\ldots.g_{n1}$ can be obtained from

$$Y_{ij}=\mu_i(g_{ij}) \tag{3}$$

where μ_i is the link function for $g_{ij}\sim N(0,\sigma^2{}_2)$ which is a sequence of mutually independent measurement errors. It has also been assumed that $\mu_i(g)=x_{1i}(g)'\beta$, in which the vectors $x_{1i}(g)$ and β give the time-varying explanatory variable and their corresponding regression coefficient.

In the case of survival modelling for the time t, the semi-parametric multiplicative model is extended into

$$\tau_i(g)=\tau_0(g)\alpha_0(g)\exp\{x_{2i}(g)'\beta+R_{2i}(g)\}, \tag{4}$$

where $\alpha_0(g)$ is unspecified and X for the covariate information. The term R_{2i} is useful as a latent process. The parameter $\tau_0(g)$ is the baseline hazard function.

3.1. Longitudinal data models

To deal with longitudinal data with continuous outcome the widely used method is the linear mixed effects model. The linear mixed effect longitudinal models have had a long history in biostatistical theory and practice since the first published paper of Laird and Ware (1982). If $Y_{i1},Y_{i2},\ldots\ Y_{ini}$ is ith subject observations for the g_{i1}, g_{i2}, . . . , g_{ini} times then the model can be formulated to

$$Y_{ij}=\mu_i(g_{ij})+R_{1i}(g_{ij})+\varepsilon_{ij} \tag{5}$$

where $\mu_i(g_{ij}) = x^{T}{}_{1i}(g)\beta_1$ is the mean response, $R_{1i}(g_{ij})=d^T{}_{1i}(g_{ij})Z_i$ is applied to explain the subject-specific random effects, and $\varepsilon_{ij} \sim N(0, \sigma^2{}_\varepsilon)$ is for random error. The terms $R_{1i}(g)$ is applied for subject specific HBA1c observations. The time-varying covariates are explained by the vectors $x_{1i}(g)$ and β_1. The term U_i is used to represent the random factor of the covariates $d_{1i}(s)$ (as compartment of $x_{1i}(g)$) and assumed distributed as $N(0,\Sigma)$.

3.2. Survival data models

The semi-parametric survival model is becoming an attractive tool for the survival analysis. However, the parametric model is more attractive due its simplicity in the survival analysis. The widely applied statistical methods for the survival analysis are Weibull and Cox proportional hazard models.

In the case of the parametric model the ith subject is assumed to follow the Weibull distribution by $g_i\sim$Weibull $(r, r_i(g))$.

where $$\log(r_i(g)) = x^T{}_{2i}(g)\beta_2+R_{2i}(g) \text{ and } r > 0. \tag{6}$$

The $x_{2i}(g)$ and β_2 are the covariates of interest and corresponding regression coefficients. The object $R_{2i}(g)$ is applied for the subject specific covariate and intercepts.

However, the event history can be formulated for time g by

$$\tau_i(g)= \tau_0(g)t^{r-1}r_i(g) = \tau_0(g)t^{r-1}\exp(x^T_{2i}(g)\beta_2+R_{2i}(g)), \quad (7)$$

Guo et al. (2004) applied the semi-parametric proportional hazard model in clinical trial by

$$\tau_i(g)= \tau_0(g)\exp(x^T_{2i}(g)\beta_2+R_{2i}(g)), \quad (8)$$

where $\tau_0(g)$ is used for the baseline hazard function. The fundamental properties of the model were discussed by Cox and Oakes (1984).

3.3. Joint model

The joint model has been linked to sub-models by the measurement model for the longitudinal process and the intensity model for the survival process. The connection between longitudinal and survival analysis can be established by stochastic dependence between R_{1i} and R_{2i}. Henderson et al. (2000) discussed the joint modelling via latent zero-mean bivariate. The joint model can be classified into two linked sub-models, (i) the measurement model for the longitudinal process and (ii) the intensity model for the survival process. The joint model becomes applicable to the sub-model.

The joint model in equations (3) and (4) can be formed by

$$R_{1i}(g)=Z_{1i}+Z_{2i}(g), \quad (9)$$

and

$$R_{2i}(g)=\lambda_1Z_{1i}+\lambda_2Z_{2i}+\lambda_3(Z_{1i}+Z_{2i}g)+Z_{3i} \quad (10)$$

Equation (3) used the random intercept model as a link function to the longitudinal data.

In equation (9) $(Z_{1i}, Z_{2i})^T$ follows the bivariate normal distribution with $N(0,\Sigma)$, Z_{3i} is independent and assumed to follow $N(0,\sigma^2)$. The parameters λ_1,λ_2 and λ_3 in the survival model (9) measure the association between the two sub-model indicated by the random intercept, slopes and fitted longitudinal value at the even time $R_{1i}(g)$.

The dependence between R_{1i} andR_{2i} is useful to describe the relation between longitudinal and survival processes.

The longitudinal model (3) is basically the random effect model introduced by Laird and Ware (1982). In equation (6), the parameters λ_1, λ_2, and λ_3 are functional to describe the association between two sub-models through random intercepts, via event $R_{1i}(g)$at time t. It is assumed that the latent variables $(Z_{1i},Z_{2i})^T$ have bivariate Normal distribution $N(\mathbf{0},\mathbf{\Sigma})$. More specially, Z_{3i} is assumed with $N(0, \sigma^2_\varepsilon)$. The term U_{3i} is assumed to be not dependent on $(Z_{1i}, Z_{2i})^T$.

4. Analysis of Metformin with Pioglitazone or Gliclazide with Pioglitazone data

4.1. Sources of data

The data set obtained as a secondary source has been taken from the clinical trial conducted in 2008. The patients are taken from the randomized, double blind and a parallel group study conducted in Menakshi Mission Hospital, Tamil Nadu. A total of 65 patients has been selected to participate in the study, 32 in (1) A combination of Metformin with pioglitazone, and 33 in the group of (2) A combination of Pioglitazone with Gliclazide.

4.2. Description of data set

The drug effectiveness is compared through longitudinal and survival data.

Table 1. Description of HBA1c according to different treatment groups and visits

Treatment	Visits	Min	Max	Mean	SD	Missing observation	Available number of observations
Metformin with pioglitazone	HBA1c1st	7.0	12.6	9.52	0.23	0	32
	HBA1c 2nd	6.8	11.7	8.31	0.35	4	28
	HBA1c 3rd	6.3	10.8	7.52	0.29	10	22
Gliclazide with pioglitazone	HBA1c 1st	6.8	12.9	9.51	0.28	0	33
	HBA1c 2nd	6.7	11.9	8.62	0.33	4	29
	HBA1c 3rd	6.3	11	8.03	0.30	13	20

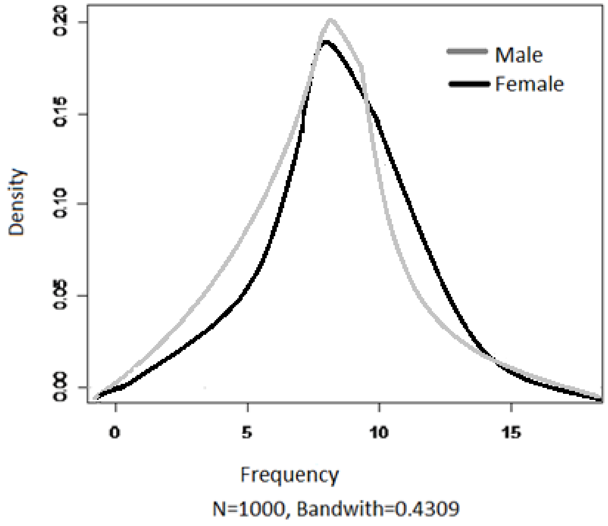


Figure1. Estimated sex wise posterior density of the patient from joint analysis

In this trial, n =65, type 2 d iabetes who met the entry conditions were included and randomly allocated to receive either Metformin with Pioglitazone or Gliclazide with Pioglitazone.

The HBA1c levels have been recorded at the study entry, 3and 12 m onth visits. The death of the subject has also been recorded. However, it is to be noted that the reason of death cannot be specify due to the drug effect. The recorded sample sizes for the drug group (Metformin plus Pioglitazone) in three visits are (32, 28 and 22) and (33, 29 and 20) for the (Pioglitazone plus Gliclazide) group. The estimated posterior density observed adjusted through male and females are given in Figure1. The data is highly affected by drop-out and missing data over time due to the occurrence of death. The Kaplan-Meier curve has been used to show the comparative figure of death between the two drug groups over the follow-up visits. It shows that the survival rate among both groups were same up to the initial 100 d ays after the randomization. Afterwards, survival in the Pioglitazone with Gliclazide group has been found to be better than Metformin with Pioglitazone group. The level of HBA1c is represented through Y_{ij} for i^{th} observations of the j^{th} individual. The considered dichotomous covariates are Sex (female=0, male=1), value of ECO and ECG (Normal level=0, otherwise 1), and Drug (Pioglitazone with Gliclazide=0 and Metformin with Pioglitazone = 1). The covariates value levelled with “0” is considered as reference value in the analysis.

The objective of the study is to observe the effect of the drug on HBA1c and survival time in type 2 diabetes individuals.

5. Analysis

The analyses for the longitudinal and survival data in type 2 diabetes trial are compared with the Bayesian approach. The linear random effects model for HBA1c is specified as

$$Y_{ij}=\beta_{11}+\beta_{12}*Drug_i+\beta_{14}{}^{*}Sex_i+\beta_{15}\,{}^{*}ECO_i+\beta_{16}{}^{*}ECGi+R_{1i}(g_{ij})+\varepsilon_{ij} \qquad (11)$$

where $R_{1i}(g_{ij})=Z_{1i}+Z_{2i}g_{ij}$. The term $R_{1i}(g_{ij})$ is induced as a random factor for the intercept and slopes over the duration of study, where the $Z_i=(Z_{1i},Z_{2i})^T\sim N(0,\Sigma)$. It gives the scope to assume that different individuals have different observations before as well during the study of HBA1c.

The estimated regression coefficients have been obtained by R programming. In the case of longitudinal analysis, the rlm (http://cran.r-project.org/web/packages/) function has been applied in R, whereas in the case of survival analysis surv(http://cran.r-project.org/web/packages/) function has been used in survival library. The summarized results are given in the Table 2. As a results the estimated average mean for the Metformin with Pioglitazone is obtained with -0.42 with 95% confidence interval of (-0.67, 0.17), proposing significant increment of HBA1c in the Metformin with Pioglitazone group as compared to Pioglitazone with Gliclazide group.

The comparative changes of HBA1c level are provided in the Figure 3. The estimated regression coefficient value for ECO with 95% confidence interval is observed with 0.23 (-0.01, 0.47). Hence, a patient who has randomized with Drug 1 is found to be more effective to reduce the level of HBA1c in comparison with Drug 0. Other variables ECG and sex are observed with insignificant contribution. Similarly, ECG and ECO are found statistically not significant in survival analysis.

Table 2. Classical analysis for type 2 diabetes drug treatment effect data

Parameters	Point estimate	95% Confidence Interval
Longitudinal Data Analysis		
(Linear Mixed Effect Model)		
Intercept	9.58	(9.32, 9.06)
SEX(reference=female)	-0.29	(-0.39, -0.19)
ECO(reference=Normal)	0.23	(-0.01,0.47)
ECG(reference= Normal)	0.12	(-0.13,0.37)
DRUG(reference=Pioglitazone with Gliclazide)	-0.42	(-0.67,0.17)
DRUGXTIME	-0.18	(-0.42,0.06)
Survival Analysis		
Intercept	9.55	
SEX	-0.12	(9.12,9.98)
ECO	0.14	(-0.22,-0.03)
ECG	-0.18	(-0.04,0.33)
DRUG	-0.33	(-0.45,0.09)
DRUGXTIME	-0.11	(-0.74,-0.08)
		(-0.24,-0.02)

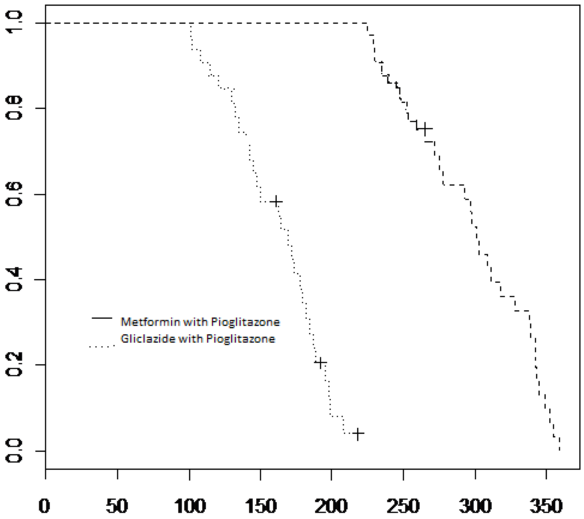


Figure 2. Kaplan-Meier Curve for the drug effect comparison in the type 2 diabetes patients.

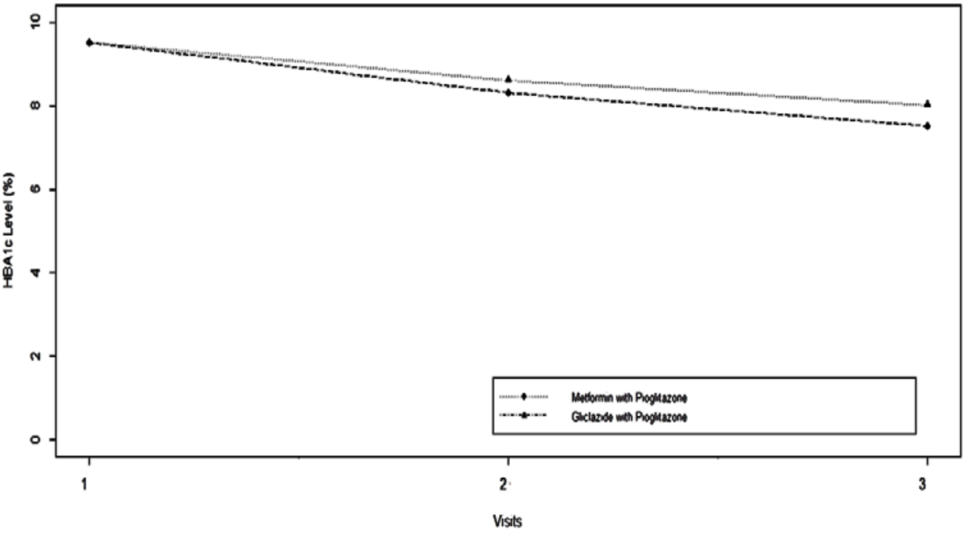


Figure 3. The comparative changes of HBA1c level throughout the study period in drug treatment group.

Henderson et al. (2000) proposed to use the Bayesian approach to fit the joint longitudinal model. The Bayesian approach with the vague prior (Uniform (-1,1)) has been applied and compared with the classical approach. The vague prior has been used to make the possible comparison between the classical approach and the Bayesian approach in WINBUGS. The hyperparameter has been chosen for the minimum impact on t he relative data. In the longitudinal sub-model, the multivariate normal and inverse gamma priors have been assumed for the main effect $\boldsymbol{\beta}_1$ and the error variance σ^2_ε, respectively. In the same way the multivariate normal and inverse gamma priors have been assumed for the effects $\boldsymbol{\beta}_2$ and σ^2_ε in the survival sub-model. The vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have been expressed by $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{15}, \beta_{16})^T$ and $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}, \beta_{25})^T$. The parameters γ_1 and γ_2 have been assumed to follow the normal distribution. Priors are selected to reflect the appearance of likelihood.

6. Model selection

The models under consideration are:

Model 1:- $Y_{ij}=\mu_i(g_{ij})+R_{1i}(g_{ij})+\varepsilon_{ij}$

Model 2- $\tau_i(g)=rt^{r-1}\mu_i(g) = rt^{r-1}\exp(x^T_{2i}(g)\beta_2+R_{2i}(g))$,

Model 3:- $Y_{ij}=\mu_i(g_{ij})+(Z_{1i}+Z_{2i}(g))(g_{ij})+\varepsilon_{ij}$

Model 4:- $\tau_i(g)=rt^{r-1}\mu_i(g) = rt^{r-1}\exp(x^T{}_{2i}(g)\beta_2+ (\lambda_1U_{1i}+\lambda_2Z_{2i}+\lambda_3(Z_{1i}+Z_{2i}g)+ Z_{3i})$

The comparison between different models is an important issue in the statistical inference. In the case of the Bayesian approach, the widely applied tools are AIC, DIC, BIC and Bayes factor for model comparison. In this work, we have used the DIC. Our priors are selected to make less influence on the likelihood. The model selection can be performed through AIC, BIC, Bayes factor and DIC. Like other selection methods, DIC also gives the model summary to single parameters, through a specific Bayesian inference. Let θ and y be the parameters of interest and the response variable is defined as

$$p_D = E_{\theta/y}[D(\theta)] - D\big(E_{\theta/y}[\theta]\big) = \bar{D} - D(\bar{\theta}) \tag{12}$$

The notation $D(\theta)$ is the deviance function and $D(\theta) = -2\log f(y/\theta) + 2\log g(y)$, where $f(y/\theta)$ is the likelihood function is and $g(y)$ is the standard function of the data. Further, $D(\theta)$ can be formed through $D(\theta) \approx D(\bar{\theta}) + \chi_p^2$. It is formulated through Bayesian Central limit Theorem and details are available in Carlin and Louis (2000). The model selection is obtained through

$$DIC = \overline{D} + p_D \tag{13}$$

Here, p_D is the number of parameters. The posterior expectation of equation (12) is $\overline{D} = E_{\theta/y}[D(\theta)]$, small value of p_D and corresponding minimum value of the DIC gives maximum effective model. The details about the DIC can be seen in the highly cited papers of Spiegelhalter et al. (2002). In WINBUG the parameters are obtained through MCMC technique. There are several versions of DIC available for model selection in recent articles, namely Celeux et al. (2006) and Chen (2006).

The Table 3 gives the DIC values for different models of drug trial comparison data in type 2 diabetes patients. The results are obtained by the two parallel chains of MCMC sampling through 10,000 iterations. We st art with simple model of equation (1). As an extension, the term R_{1i} has been added in the equation (1) and in both cases the DIC values have been obtained. The DIC value for the Model (1) is 2345 and for the Model (2) is 2356. In the case of survival analysis, the DIC values for the Model (3) and the Model (4) are found to be 2424 and 2452, respectively. The minimum DIC value of the specific model can be considered as the best fitted model. The details about DIC can be cited with Spiegelhalter et al. (2002). Here, the minimum DIC value for the Model 1 has been found. So, it can be concluded that the Model 1 is the best.

7. Comparison of separate and joint models

The Table 2 gives the point estimates of regression coefficients for covariate of interest by the linear mixed effect model. The linear mixed effect model has been computed with respect to sex (female=0), Drug (Metformin with Pioglitazone=1). The regression coefficient -0.29 in the case of sex showed that the male type 2 diabetes patients are reduced to lower amount of HBA1c in comparison to the female ones. The Table 3 gives the Highest Posterior Density (HPD) interval for the covariates in different models. The results for the longitudinal model have been obtained from the model in equation (5). The term $R_{1i}(g_{ij})$ is applied as an extension of the joint model to the separate longitudinal model. The term $R_{1i}(g_{ij})$ has also been separated to Z_{1i}, Z_{2i} by equation (9). The 95% credible intervals for the survival model have been obtained from the equation (3) and equation (5) for the survival model, where $R_{2i}(g)$ is used as an extension over the separate model. The performances of both models are found similar. The regression coefficient 0.01 obtained through longitudinal sub-model with separate analysis confirmed that the male type 2 diabetes patients are reduced to higher amount of HBA1c in comparison to the female ones. On the other hand, in the case of sex regression coefficients, the longitudinal sub-model in joint analysis, survival sub-model with separate analysis and joint analysis follow the same pattern as observed in separate analysis in the longitudinal sub-model. The covariates of interest, ECG and time, are observed with considerable extension, while only ECG is found significant in the case of the survival sub-model. The regression coefficients from the classical approach are observed with 0.12(-0.13, 0.37) and 0.23(-0.01,0.47) for ECG and ECO, respectively. In the case of joint modelling applied through the Bayesian approach the posterior means of the regression coefficients are obtained with 0.03(-0.39, 0.42) and -0.02(-0.45, 0.42) for ECG and ECO, respectively.

Table 3. Posterior estimates of the parameters observed through different models

Parameter	Separate analysis			Joint analysis		
	Posterior mean	DIC	95% Credible interval	Posterior mean	DIC	95% Credible interval
Longitudinal Sub-model						
Intercept(β_{11})	8.70	2345	(8.23,9.15)	8.72	2356	(8.21,9.17)
Time(β_{11})	-0.17		(-0.97,0.09)	-0.19		(-0.93,0.07)
Time*Drug(β_{12})	0.19		(-0.14, 0.53)	0.15		(-0.12,0.49)
Sex(β_{13})	0.01		(-0.33,0.37)	0.03		(-0.35,0.39)
ECG (β_{14})	-.02		(-0.41,0.46)	-.03		(-0.39,0.42)
ECO (β_{15})	-0.01		(-0.48,0.45)	-0.02		(-0.45,0.42)
Σ_{11}	1.97		(1.41,2.67)	1.95		(1.38,2.63)
Σ_{22}	0.98		(0.73,1.5)	0.95		(0.70,1.2)
P	-0.12		(-0.30,0.07)	-0.10		(-0.27,0.05)
σ^2	0.95		(0.73,1.21)	0.93		(0.71,1.18)
Survival Sub-model						
Intercept(β_{21})	-15.31	2424	(-19.31,-8.75)	-15.31	2452	(-19.36, -8.76)
Drug(β_{22})	-0.47		(-4.94,4.24)	-0.49		(-4.99,4.28)
Sex(β_{23})	2.81		(-0.75,5.85)	2.80		(-0.79,5.83)
ECG (β_{24})	0.00		(-2.49,2.99)	0.01		(-2.53,3.01)

The separate analysis in longitudinal setup reveals the regression coefficients with -0.02(-0.41, 0.46) and -0.01(-0.48, 0.45) for ECG and ECO. The separated and joint survival analysis is computed with regression coefficients by 0.00(-2.49, 2.99), 0.01(-2.53, 3.01) for ECG and 0.19(-2.72, 2.51), -0.17 (-2.12, 2.15) for ECO, respectively. It is concluded that the regression coefficients obtained through classical approach for ECG are higher in joint and separate approach in longitudinal setup and further followed by survival setup through prior assumption. The same pattern is obtained in the case of ECO. The highest value of the regression coefficient is found with frequency approach. In both types of analysis it is found that no estimates of the treatment appear to have significant effect on the evolution of the matter HBA1c either on the longitudinal part or on the survival. The rate reduction of HBA1c over the follow-up period is found higher in the Metformin with Pioglitazone group.

8. Discussion

In this paper, the Bayesian approach with the longitudinal and survival analysis is applied in the type 2 diabetes drug comparison. This type of the model is important in clinical trial. The models are also useful with other biochemical parameters. It is important to investigate how the biomarker of interest changes over time and its correlation with the treatment under study to better explore the therapeutic effect as pointed by Deslandes and Chevret (2010). The results are obtained through the freely available software and compared with R and WINBUGS. Due to intention-to-treat and other logistical reasons, the whole data set has not been provided to the authors for analysis. The work is carried out only on fully observed but partially data set. Therefore, the whole information about mortality of the patients could not be provided. The aim of this paper is to compare two effects of drug treatment through HBA1c level among type 2 diabetes. Nathan et al. (2009), Holman et al. (2007), Holman et al. (2009) and Meneghini et al. (2007) recommended the level of HBA1c as thresholds for starting insulin. Kilpatrick et al. (2008) discussed broadly the limitation of HBA1c for the screening test. Ginde el al. (2008) and Anand et al. (2003) examined the variation of HBA1c with different demographic characters in the US population. Mirzazadeh et al. (2009) found that the HBA1c can be affected by age distribution. Zahra et al. (2010) concluded that the low HBA1c is a strong evidence to rule out diabetes. However, we acknowledge the deficiency in not including the glucose tolerance test in this work. In addition, as another limitation in this study, we have not used the life style parameters of the type 2 diabetes patients since some patients cannot be followed or died due to other reason. The analysis becomes complicated due to the presence of dropouts in the data. Thus, the analysis of such type of data by separate analysis may generate biased and inappropriate results whereas the application of joint analysis is useful to deal with dropout observations. Recently, Chi (2006), Williamson et al. (2008) and Li et al. (2009) discussed joint modelling in the longitudinal and survival data analysis. Actually, Guo (2004) has motivated our work to apply the Bayesian approach in longitudinal data analysis to obtain the posterior inference for any parameter. Thus, we have developed a fully Bayesian approach, implemented via MCMC in WINBUGS software. Recently, such a Bayesian approach for joint longitudinal and survival analysis has also been implemented by Li et al. (2009). This work illustrates how the joint model strategy may affect the results. Here, the joint analysis is found inferior in comparison with the separate analysis. It may be due to the presence of other complicated issues in the data set. Lind et al. (2008) concluded that the latent mixed effect is appropriate in the hazard model. In this work, it is found by joint longitudinal and separate analysis that Metformin plus Pioglitazone is equally effective to reduce the HBA1c level as compared to Gliclazide plus Pioglitazone.

9. Conclusions

Here, the HBA1c observations by longitudinal and survival analysis tools are compared with type 2 diabetes patients. The results confirm that the joint modelling approach is a useful tool for longitudinal data analysis, survival analysis and, consequently, for the actual application to the drug effect comparison in clinical trials. The Markov Chain Monte Carlo method is employed to effectively estimate HBA1c values for different visits in type 2 diabetes patients. The applied models can be useful in different fields like oncology, endocrinology and other specific drug research. It is confirmed that the combination of Metformin plus Pioglitazone is equally beneficial to reduce HBA1c level, hence the risk of type 2 diabetes. The Bayesian approach is considered as extending over the Frequency approach on longitudinal and survival data analysis.

Acknowledgements

The authors thank the two anonymous referees for their cautious reading and constructive suggestions which have led to improvement on earlier versions of the manuscript. They are also very grateful to R. Saroj for providing data for analysis and to the International Biometric Conference, Japan 2012, for providing a platform to present the initial version of this work.

REFERENCES

AMERICAN DIABETES ASSOCIATION (2010).Diagnosis and classification of diabetes mellitus. Diabetes Care, 33 (Suppl 1), S62–S69.

ANAND, S. S., RAZAK, F., VUKSAN, V., GERSTEIN, H. C., MALMBERG, K., YI., Q., TEO, K. K., YUSUF, S., (2003). Diagnostic strategies to detect glucose intolerance in a multiethnic population, Diabetes Care, 26(2), 290–296.

CARLIN, B. P., LOUIS, T. A., (2000). Bayes and Empirical Bayes Methods for Data Analysis (2nd ed.), Boca Raton, FL: Chapman and Hall/CRC Press.

CELEUX, G., FORBES, F., ROBERT, C. P., TITTERINGTON, D. M., (2006). Deviance information criteria for missing data models, Bayesian Analysis, 4, 651–674.

CHEN, M. H., (2006). Comments on article by celeux et al., Bayesian Analysis, 4, 677–680.

CHI, Y. Y., IBRAHIM, J. G., (2006). Joint models for multivariate longitudinal and multivariate survival data. Biometrics. 62(2), 432–445.

COX, D. R., D. OAKES, (1984). Analysis of Survival Data, London: Chapman and Hall.

DCCT STUDY GROUP, (1995). The relationship of glycemic exposure (HbA1c) to the risk of development and progression of retinopathy in the diabetes control and complications trial, Diabetes, 44(8), 968–983.

DEGRUTTOLA, V., TU, X. M., (1994). Modeling progression of cd4 lymphocyte count and its relationship to survival time, Biometrics, 50, 1003–1014.

DESLANDES, E., CHEVRET, S., (2010). Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: Application to ICU data, BMC Med Res Methodology, 2010, 10, 69.

ELASHOFF, R., LI, G., LI, N., (2008). A joint model for longitudinal measurements and survival data in the presence of multiple failure types, Biometrics, 64, 762–771.

GHAZANFARI, Z., HAGHDOOST, A. K., ALIZADEH, S. M., ATAPOUR, J., ZOLALA, F., (2010). Comparison of HbA1c and Fasting Blood Sugar Tests in General Population. International Journal of Preventive Medicine, 1(3), 187–194.

GINDE, A. A., CAGLIERO, E., NATHAN D. M., CAMARGO, C. A., J. R., (2008). Value of risk stratification to increase the predictive validity of HbA1c in screening for undiagnosed diabetes in the US population. J Gen Intern Med, 23(9), 1346–1353.

GUO, X. U., CARIN BRADLEY, P., (2004). Separate and Joint Modeling of Longitudinal and Event Time Data Using Standard Computer Packages, The American Statistician, February 2004, Vol. 58, No. 1, 1–9.

HENDERSON, R., DIGGLE, P. J., DOBSON, A., (2000). Joint Modeling of Longitudinal Measurements and Event Time Data, Biostatistics, 1, 465–480.

HOGAN, J. W., LAIRD, N. M., (1997). Model-based approaches to analysing incomplete longitudinal and failure time data, Statistics in Medicine, 16, 259–272.

HOLMAN, R. R., THORNE, K. I., FARMER, A. J., DAVIES, M. J., KEENAN, J. F., PAUL, S., LEVY, J. C., (2007). Addition of biphasic, prandial, or basal insulin to oral therapy in type 2 diabetes. N Engl J Med, 357, 1716–1730.

HOLMAN, R. R., FARMER, A. J., DAVIES, M. J., LEVY, J. C., DARBYSHIRE, J. L., KEENAN, J. F., PAUL, S. K., (2009). Three-year efficacy of complex insulin regimens in type 2 diabetes. N Engl J Med, 361, 1736–1747.

KILPATRICK, E. S., (2008). Haemoglobin A1c in the diagnosis and monitoring of diabetes mellitus, J Clin Pathol, 61(9), 977–82.

LITTLE, R. J. A., (1995). “Modeling the drop out mechanism in repeated measures studies,” Journal of the American Statistical Association, 90, 1112–1121.

LIND, M, ODEN, A., FAHLÉN, M., ELIASSON, B., (2008). A Systematic Review of HbA1c Variables Used in the Study of Diabetic Complications, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 282–293.

LAIRD, N. M., WARE, J. H., (1982). Random-Effects Models for Longitudinal Data, Biometrics, 38, 963–974.

LI, L., HU, B., GREENE, T., (2009). A semi-parametric joint model for longitudinal and survival data with application to hemodialysis study. Biometrics. 65(3), 972–986.

MENEGHINI, L. F., ROSENBERG, K. H., KOENEN, C., MERILAINEN, M. J., LÜDDEKE H. J., (2007). Insulin detemir improves glycaemic control with less hypoglycaemia and no weight gain in patients with type 2 diabetes who were insulin naive or treated with NPH or insulin glargine: clinical practice experience from a German subgroup of the predictive study. Diabetes Obes Metab. 9, 418–427.

MIRZAZADEH, A., BARADARAN, H. R., HAGHDOOST, A. A., SALARI, P., (2009). Related factors to disparity of diabetes care in Iran, Med Science Monit, 15(5), H32–H36.

NATHAN, D. M., BUSE, J. B., DAVIDSON, M. B., FERRANNINI, E., HOLMAN, R. R., SHERWIN, R., ZINMAN, B.,(2009). Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes, Diabetes Care. 32, 193–203.

NATH, D. C., BHATTACHARJEE, A., (2011). A Bayesian Approach in Autoregressive models in Longitudinal Data Analysis: An Application to Type 2 Diabetes Drug Comparison, Asian Journal of Applied Science, 4(6), 640–648.

SCHLUCHTER, M. D., (1992). Methods for the analysis of informatively censored longitudinal data, Statistics in Medicine, 11, 1861–1870.

SPIEGELHALTER, D. J., NICOLA, G. BEST, CARLIN, B. P., LINDE, A. V. D., (2002). Bayesian measures of model complexity and fit .Journal of the Royal Statistical Society: Series B (Statistical Methodology. 64, 4, 583–639.

STRATTON, I. M., ADLER, A. I., NEIL, H. A., MATTHEWS, D. R., MANLEY, S. E., CULL, C. A., HADDEN, D., TURNER R. C., HOLMAN, R. R., (2000). Association of glycaemia with macrovascular and microvascular complications of type 2 di abetes (UKPDS 35): prospective observational study, British Medical Journal, 2000; Aug 12; 321(7258), 405–412.

The International Expert Committee (2009).The International Expert Committee Report on the Role of the A1C Assay in the Diagnosis of Diabetes. Diabetes Care, 32(7), 1327–1334.

TSIATIS, A. A., DEGRUTTOLA, V., ANDWULFSOHN, M. S., (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error, Applications to Survival and CD4 Counts in Patients with AIDS, Journal of the American Statistical Association, 90, 27–37.

WILLIAMSON, P. R., Kolamunnage-Dona, R., Philipson, P., Marson, A. G., (2008). Joint modelling of longitudinal and competing risks data. Statistics in Medicine. 27(30), 6426–6438.

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 453–466

THE IMPACT OF ALTERATIONS IN THE LOCAL INSOLVENCY LEGISLATION ON BUSINESS BANKRUPTCY RATES IN POLAND

Henryk Gurgul[1], **Paweł Zając**[2]

ABSTRACT

The purpose of this paper is to analyze the number of insolvencies in Poland before and after the major bankruptcy code novelization in the second quarter of 2009. Authors check whether the novelization had its intended effect of reducing bankruptcy rates. Therefore, econometric models have been implemented to investigate changes in bankruptcy rates using quarterly data from the period 2003-2013. While controlling the variety of macroeconomic factors that have influenced insolvency rates, we found that after implementation of the novelization the aggregate bankruptcy rates significantly increased.

Key words: small and medium firms, aggregate bankruptcy rates, amendments to the law.

1. Introduction

Insolvency, which is the result of either law regulations or the court judgement (i.e. bankruptcy), can happen despite the economic reasons, leading to the cessation of the company. Company’s liquidation regardless of its size, sphere of interest, territory or trade partners is the source of confusion and, more importantly, distress on the market (Zdyb, 2009). Economists agree that in a short period of time bankruptcy is harmful to the business market. However, in the long run the positive results of closing an ineffective company can be visible (Schumpeter, 1934).

The market itself is often unable to eliminate the unsuccessful entrepreneurs, e.g. it cannot eliminate from the market the companies functioning on the verge of cost-effectiveness. The protection procedures against disastrous outcomes of their activity are regulated by the bankruptcy law. It helps to avoid or minimize the

[1] Department of Applications of Mathematics in Economics, AGH University of Science and Technology, Cracow, Poland. E-mail: henryk.gurgul@gmail.com.

[2] Department of Applications of Mathematics in Economics, AGH University of Science and Technology, Cracow, Poland. E-mail: pawel.zajac@agh.edu.pl.

negative consequences for the surrounding as a result of insolvency of the debtor. The regulations are necessary because no restrictions describe the conditions of setting up any business activity. The candidates to become entrepreneurs are not always suitable as well. Therefore, some of them fail due to the lack of competence, management errors and changes in the surrounding of a company.

In order to protect firms and investors from some consequences of bankruptcy in recent years, the number of countries in the European Union reformed their bankruptcy legislations (e.g. France, Germany, the UK, Spain, Finland, Italy, Belgium), while several other EU members are going to introduce similar regulations in the future. The main aim of these reforms was changing traditional, old framework, which was solely focused on liquidation, into a modern framework which should combine reorganization and liquidation. The effort is concentrated on creating more transparent and efficient system. The legal regulation in the new system should encourage more reorganization instead of liquidation only. The respective guidelines were outlined by the European Commission and the World Bank in order to suggest the best practice bankruptcy procedures (EC 2003; World Bank 2001). Economists assume that applied reforms in the bankruptcy system could lower aggregate bankruptcy rates[3]. In the EU the majority of failure cases concerns SMEs which essentially contributes to GDP in EU countries (cf. Hudson, 1986). Therefore, the reduction of the bankruptcy rates of small businesses should be an important topic in a bankruptcy reform. In addition, not all suggested legal regulations in the EU seem to be beneficial or useful for smaller firms. This problem occurs across different business sectors. Some industries are more represented in a country's economy than others and this fact can have significant implications. However, the mentioned problems have received not much attention in the economic literature. There is only little empirical evidence on the impact of bankruptcy legislation reform on a ggregate bankruptcy rates in the EU besides UK. In the EU the financial system is based on banks. In Anglo-Saxon countries it is based on the market. Therefore, the impact of reforms in the Continental Europe is not always similar to the one in the Anglo-Saxon countries. In this context the question arises how the bankruptcy rates can be influenced by the implementation of recommendations of international best practices in the Continental Europe.

At the turn of the century there still existed antiquated bankruptcy law from 1934 in Poland. Therefore, in the nineties of the twentieth century Poland underwent long-anticipated, immense economical change. In the year 1990 there were about 1.2 million registered enterprises in Poland, whereas in the year 2000 the number increased up to 3 m illion. The growing number of companies

[3] The aggregate bankruptcy rates are measured as the percentage of liquidation type bankruptcies to the total company population.

influenced the bankruptcy issue which applied to a wider group of entrepreneurs and, due to the new market situation, the law had to be readjusted. The new Bankruptcy and Remedial Act came into force in February 2003. The Act had been worked out according to the guidance of the European Commission and the World Bank (Zedler, 2003). The aim of remedial proceedings was to repay creditors with simultaneous attempt to preserve the existence of a company. In the following years the application of this new act was analyzed and the necessity of further changes was alleged. As a result, the Act of 6 March 2009 amending Bankruptcy and Remedial Act, Bank Guarantee Fund Act and National Court Register Act was signed on 12 March 2009 (the Amendment). The Amendment introduced more than 150 changes in Act of February 2003. The new solutions aimed at accelerating bankruptcy procedures and satisfying the creditor's claims to the debtor. The Amendment significantly reformed the remedial proceeding, which previously was rarely instigated. Under the Amendment the entrepreneur is allowed to fill the declaration of bankruptcy along with a demand of permission to initiate a remedial proceeding. Additionally, present recovery proceedings may concern the restructuring of not only monetary liabilities, but all liabilities which can be subject to an arrangement (Kallaur, 2009).[4]

Authors of this paper will examine if the introduced changes in the Polish legal system had significant statistical influence on aggregate bankruptcy rates in Poland. In particular, it will be tested if the aggregate bankruptcy rates after the novelization of 2009 were lower or higher than expected due to the existing macroeconomic conditions[5]. Visual inspection of the data suggests that the novelization in bankruptcy law in the second quarter of 2009 had positive effects on aggregate business bankruptcy rates in Poland. However, this first impression will be checked by means of quantitative tools.

The remainder of the paper is organized as follows: Section 2 overviews the existing literature. Data and methodology are presented in Section 3. In Section 4 the empirical results are discussed and Section 5 concludes.

2. Literature overview

Researchers of the insolvency are concentrated either on the risk of bankruptcy for specific firms or on modelling aggregate bankruptcy rates. The research stream concerning the bankruptcy of individual firms is reflected in the contribution by Greiner and Schein (1988). They argued that flexibility of the company depends mostly on t he abilities and creativity of the owner. The

[4] In the second quarter of 2009 the Act of 5 December 2008 was also adopted. Since then natural person not engaged in economic activity earned the possibility to declare the so-called consumer bankruptcy. The consumer bankruptcy statistics are not included in our calculations.

[5] One may assume that the aggregate bankruptcy rates should have increased during the Global Financial Crisis.

companies adherent to old, uncompetitive solutions are swiftly eliminated from the market. In times of uncertainty and rapidly changing environment the lack of alterations inside the company often leads to its closure.

In the late sixties of the twentieth century quantitative methods (e.g. discrimination methods) became commonly used to predict the risk of bankruptcy. Altman (1968) was a pioneer in this type of research. The grouping into firms vulnerable to bankruptcy and those not was done on t he basis of individual financial indicators of companies. Those ratios included: working capital to total assets ratio, retained earnings to total assets ratio (retained earnings - profits which were not paid out in dividends and which can be re-invested in the business), EBIT (Earnings Before Interest and Taxes) to the total assets ratio, the market value of equity to the book value of total liabilities and sales to total assets. Altman correctly classified up to 95 percent of companies the year before their bankruptcy and 83 percent two years before. The results of that research on the insolvency which uses discrimination methods can significantly differ from one another according to the country and time period. The reason is different propensity to bankruptcy in various periods. In addition, diverse indicators concerning situation of the company are applied. There is widely accepted point of view that researchers using the same set of variables but for companies from different countries or different time periods may obtain quite different results. Hence, some new attempts to form models are applied in order to allow to predict the bankruptcy.

The second stream of research aims at modelling of aggregate bankruptcy rates. It refers to macroeconomic factors. In the literature there is no doubt that macroeconomic factors play an important role in respect of bankruptcy. In particular, Hudson (1986), Ilmakunnas and Topi (1999) and Liu (2004) found that GDP growth or business cycle indicators are negatively correlated with aggregate failure rates.

Various authors argued that aggregate corporate birth rates (Hudson 1986, 1997; Johnson and Parker, 1994) and inflation rate (Altman, 1983; Wadhwani, 1986) are likely to have an impact on bankruptcy rates. However, the direction of the relationship is not clear as it may be either positive or negative. In the literature there is little evidence concerning the effect of macroeconomic environment on the bankruptcy of small enterprises.

The competitiveness of the market has a significant impact on the number of bankruptcies. According to Foster and Kaplan (2001) the failure of the company is a consequence of two parallel processes, namely destruction and creation. The researchers claim that initiation, directing and controlling of the creative destruction is conducted by the financial markets. The bankruptcy is mainly a tool of control and protection of the market.

Chen and Williams (1999) found that the US government assistance programmes lowered bankruptcy rates in high-technology industries. Australian economists Everett and Watson (1998) established that in the case of small Australian retail enterprises between 30% and 50% of failures were caused by macroeconomic factors. The similar results with respect to small business failure in case of the US were derived by Peterson et al. (1983). On the basis on survey data they documented that economy wide factors are the reasons number two and three for small enterprises failure in the US (number one is a lack of management expertise). In line with these results Sullivan et al. (1999) proved that conditions in business surrounding are the most important reasons for bankruptcy.

Besides the macroeconomic environment, the effects of institutional factors on the aggregate bankruptcy rate attracted attention of economists. Claessens and Klapper (2005) demonstrated that formal bankruptcy in the economy depends on the bankruptcy legislation. In addition, the research on the effects of institutional change supported the important impact of legal reform on aggregate failure rates.

Fisher and Martel (2003) demonstrated that the 1992 Canadian bankruptcy reform had an influence on the number of filled corporate reorganization proposals. Cuthbertson and Hudson (1996) and Vlieghe (2001) based on the UK's 1986 Insolvency Act established a significant decrease in the corporate liquidation rate after the implementation of this reform. Liu and Wilson (2002) and Liu (2004) suggested that the beneficial effect became lower after about 4 years.

Dewaelheyns and van Hulle (2008) pointed that continental European countries recently reformed their bankruptcy legislations to stimulate reorganization and firm survival. They argued that the Belgian 1997 bankruptcy code reform, which implemented several international best practice recommendations, significantly reduced aggregate small and micro business bankruptcy rates. The contributors supplied evidence that the beneficial effect of the reform is similar among small firms (i.e. stock corporations) and micro firms (i.e. partnerships). Therefore, it was only significant in some industries (manufacturing and trade). Their results showed that especially the measures taken to limit domino bankruptcy effects were likely to have a substantial impact.

To summarize, from the literature discussed above we can learn that the corporate bankruptcy rates are related to numerous macroeconomic variables. The main goal of this paper is to examine if the legal reform also impacts aggregate bankruptcy rates in Poland. To fulfil this task the macroeconomic determinants of aggregate bankruptcy rates discussed in the literature above will be used in further calculations. These variables should be as complete and accurate as possible to describe the state of Polish economy before and after the legal reform. In particular, it should be remembered how difficult those years for the global economy were. There are strong links between the economic situation in Poland and other European countries, and because of that additional variables describing the relationship between the Polish and the EU economy will also be included in the further study.

3. Data and methodology

3.1. Presentation of dataset

The Polish law imposes on the Ministry of Justice a duty to report to public opinion information about bankruptcies pronounced in Polish courts. Acting on those reports authors calculated the aggregate bankruptcy rates from the period of 2003-2013. The data released by the Central Statistical Office of Poland (*GUS*) is used to estimate the percentage of bankruptcies out of the total number of enterprises[6]. With the use of previously mentioned literature we have distinguished a group of macro variables which, according to the theory, have influenced failures. Among those variables we may find aggregated number of start-ups, real *GDP* growth, the *OECD* composite leading indicator (*CLI*) for the euro zone countries and for Poland, return on the Warsaw Stock Exchange Index (*WIG*), current economic condition indicator (*BOSE*), inflation and average exchange rates *EURO/PLN*. All of the collected data come from various sources such as Polish Ministry of Justice, the Central Statistical Office of Poland, the National Bank of Poland and the *OECD*.

Table 1. Description of variables

Variable	Description (Δ stands for *1-Year change in*)
ΔBR	*quarterly business bankruptcy rate[7] (%)*
ΔNEW	*quarterly corporate birth rate[8] (%)*
ΔGDP	*growth in real GDP (%)*
ΔCLIeu	*OECD euro zone composite leading indicator (%)*
ΔCLIpl	*OECD Polish composite leading indicator (%)*
ΔBOSE	*current economic condition indicator (%)*
ΔWIG	*return on the Warsaw Stock Exchange Index (%)*
ΔINFL	*inflation (%, based on consumer price index)*
ΔEUR/PLN	*average exchange rates (%)*

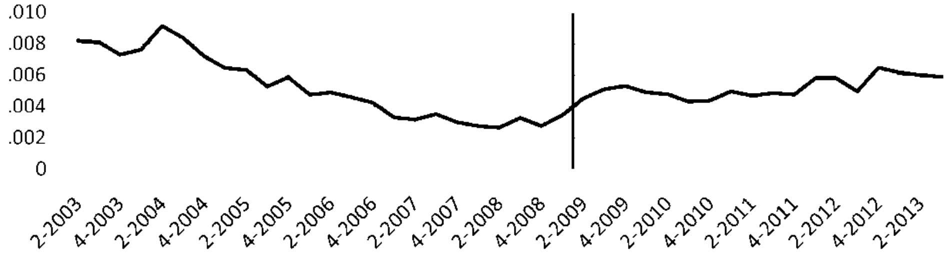
All available statistical data apply to the period between the introduction of the act and the third quarter of 2013. They were divided in to pre- and post-novelization periods (2003Q2-2009Q1; 2009Q2-2013Q3). Figure 1 shows quarterly bankruptcy rates in the concerned period.

[6] It is important to realize that companies which managed to survive the bankruptcy procedure are not included in those statistics.

[7] Quarterly business bankruptcy rate was calculated as the number of bankruptcies divided by the number of companies in existence at end of previous quarter.

[8] Quarterly corporate birth rate was calculated as the number of new companies divided by the number of companies in existence at end of previous quarter.

Figure 1. Quarterly bankruptcy rates in Poland



Source: Polish Ministry of Justice (iMSiG.pl)

The first conclusion derived from the visual examination of the picture is that overall bankruptcy rates significantly increased after the novelization. Previously, after the introduction of a new law in 2003, the indicator had been on a really high level, whereas later it started to decrease until reaching the lowest point in 2008. Immediately after the novelization in 2009 the number of failures considerably rose. In subsequent years quarterly bankruptcy rates were on the similar level reaching the values approximately twice higher than the ones in 2008. During the examined period 8134 companies bankrupted – before the novelization the mean was 190 bankruptcies per quarter, whereas after the novelization it increased to the level of 202. Of course, the rise of bankruptcy rates in Poland since 2009 reform may be, to some extent, not only the effect of novelization of 2009, but also of the essential slowdown of Polish economy as result of the world financial crisis in subsequent years. Table 2 shows the summary statistics and basic equality tests for selected variables in each selected period. It presents mean, median as well as the Kruskal-Wallis test for median (χ^2 distributed) and t-test for equality of means.

Table 2. Summary statistics and p-value for equality tests

Variable	Full period		Period I		Period II		p-value for equality tests	
	median	mean	median	mean	median	mean	Kruskal-Wallis test	*t*-test
ΔBR	0.000	0.004	-0.001	-0.001	0.000	0.009	0.000***	0.000 ***
ΔNEW	0.121	0.067	0.118	0.086	0.148	0.044	0.977	0.622
ΔGDP	4.275	4.082	5.443	5.125	3.097	2.854	0.001 ***	0.000 ***
ΔCLIeu	0.302	0.101	0.302	-0.227	0.248	0.466	0.807	0.430
ΔCLIpl	-0.156	-0.207	-0.968	-0.686	0.057	0.356	0.044**	0.017**
ΔBOSE	0.100	0.870	3.600	2.025	-1.100	-0.488	0.306	0.500
ΔWIG	0.185	0.144	0.280	0.184	0.150	0.097	0.128	0.388
ΔINFL	-0.100	0.111	1.000	0.585	-0.450	-0.417	0.085 *	0.098*
ΔEUR/PLN	-0.018	-0.004	-0.047	-0.036	-0.047	0.032	0.026**	0.052*

*** denotes significance at the 1% level; ** denotes significance at the 5% level; * denotes significance at the 10% level

The performed tests prove that both mean and median failure rates are significantly different across sub-periods. Finally, the data on m acro variables indicate significant differences in mean (median) for real *GDP* growth, change in the *OECD* Polish composite leading indicator, inflation and the average *EUR/PLN* exchange rate. It suggests that these conditions may have an important impact on bankruptcy rates.

As it has already been mentioned for the proper analysis of changes in the aggregate bankruptcy rates it is necessary to take into account the effects of global economic crisis. It is usually considered to start in the years 2007-2008, but many economists relate it to the bankruptcy of the Lehman Brothers Holdings Inc. in September 2008. Therefore, a simple analysis concerning bankruptcy rates needs to be extended by statistical models containing previously chosen macroeconomic data.

3.2. Outline of methodology

In general, we assume that the aggregate failure rate depends on current and historical values of certain variables (Altman, 1983). It may be written in the form of finite distributed lag model (*FDL*):

$$\mathrm{BR_t} = \alpha_0 + \textstyle\sum_{i=1}^{k} \sum_{j=0}^{T} \alpha_{i,j} X_{i,t-j} + \varepsilon_t, \tag{1}$$

where X_i is a macroeconomic variable, k is the number of variables and T is the maximum lag length.

Estimation of parameters from *FDL* model leads to certain difficulties. In our case the number of periods *T*, covered by lag function is so large that the individual coefficients cannot be estimated with sufficient accuracy. In addition, our variables are highly autocorrelated. High levels of correlation among the regressors imply multicollinearity, which leads to unreliable coefficient estimates with large variances and standard errors. In both scenarios the Almon polynomial distributed lag (*PDL*) specification could be helpful. The method assumes that any $\alpha_{i,j}$ can be approximated by a polynomial of order *p*:

$$\alpha_{i,j} = \beta_{i,0} + \beta_{i,1} j + \beta_{i,2} j^2 + \cdots + \beta_{i,p} j^p. \tag{2}$$

We begin with substituting equation (2) into (1):

$$\mathrm{BR_t} = \alpha_0 + \sum_{i=1}^{k} \sum_{j=0}^{T} \left(\beta_{i,0} + \beta_{i,1} j + \beta_{i,2} j^2 + \cdots + \beta_{i,p} j^p\right) X_{i,t-j} + \varepsilon_t$$

$$\mathrm{BR_t} = \alpha_0 + \textstyle\sum_{i=1}^{k} \sum_{n=0}^{p} \beta_{i,n} \left(\sum_{j=0}^{T} j^n X_{i,t-j}\right) + \varepsilon_t \tag{3}$$

By defining new variables as follows:

$$Z_{i,n,t} = \sum_{j=0}^{T} j^n X_{i,t-j} \tag{4}$$

we have a linear model of an ordinary form:

$$BR_t = \alpha_0 + \sum_{i=1}^{k} \sum_{n=0}^{p} \beta_{i,n} Z_{i,n,t} + \varepsilon_t \tag{5}$$

The method of polynomial approximation allows us to use all estimation methods which are appropriate for linear equations. In order to fulfil standard least squares assumptions, during our investigation we work on 1-year changes of macroeconomic variables. It helps to eliminate problems with stationarity and seasonality from time-series used in our research.

4. The empirical results

The problem of the autocorrelation is one of the basic problems for researchers of statistical relations between macroeconomic variables in time, and yet application of *PDL* model allows one to avoid the autocorrelation between historical values of the specific variable. Therefore, we start with estimation of *PDL* models containing only single variables. In order to correctly fit more complicated models, variables will be selected in a way to limit correlations between them.

4.1. Models with single macroeconomic variable

It is important to realize that before the approximation of coefficients begins, the decision about the lag length (*T*) and the order of polynomials (*p*) must be taken. In our case we used three criteria: the Schwartz information criterion, the Akaike information criterion and the Hannan–Quinn information criterion[9]. With the intention of examining whether amendments to the bankruptcy law influenced significantly the number of adjudicated bankruptcies we chose to use Chow Breakpoint Test. We decided to split the sample period (2003Q2 – 2013Q3) into pre- and post-reform period. The results of conducted estimations are reported in Table 3. On the left side parameters for models without break adjustment are presented and the right side of the table reports statistics for models including jump dummy to control for law change (*SPLITDUM*). Because of limited space available in Table 3 we decided to report only the sum of the lag term's coefficients. The importance of the sum was verified using t-statistics based on Newey-West HAC standard errors.

[9] In general, a lag length of 4 periods and polynomials of order 2 result in the best fit.

Table 3. Bankruptcy rates estimation: models with single macroeconomic variable

Macroeconomic variable	no regime changes		regime change 2009Q2			
	lag_SUM	Adj. R^2	lag_SUM	Chow TEST	SPLITDUM	Adj. R^2
ΔNEW	0.0011	0.0032	-0.0001	6.5633 ***	0.0006 ***	0.4048
ΔGDP	-0.0001 ***	0.1750	-0.0006	6.2085 ***	0.0013 *	0.5562
ΔCLIeu	-0.0001 ***	0.3117	-0.0011	14.2499 ***	0.0005 ***	0.5539
ΔCLIpl	-0.0004	0.0229	-0.0008	12.5859 ***	0.0007 ***	0.2070
ΔBOSE	-0.0000 ***	0.5446	0.0005	5.9641 ***	0.0005 ***	0.5925
ΔWIG	-0.0003 ***	0.5467	-0.0032	16.1954 ***	0.0007 ***	0.7165
ΔINFL	0.0000	-0.0526	0.0005	3.1774 **	0.0006 **	0.0328
ΔEUR/PLN	0.0047 ***	0.6237	-0.0005	2.4314 *	0.0004 **	0.6805

Notes: PDL models with yearly change in quarterly business bankruptcy rates as dependent variable; only cumulative lag coefficients reported; t-statistics based on Newey-West HAC standard errors used; Chow Breakpoint Test statistics for break in 2009Q2.

*** denotes significance at the 1% level; ** denotes significance at the 5% level; * denotes significance at the 10% level

The presented results confirm the existence of statistically significant relationship between selected macroeconomic variables and changes in aggregate business bankruptcy rates. According to calculations based on Newey-West HAC standard errors, the sum of the lag term coefficients proves to be significant for models with real *GDP* growth, the *OECD* composite leading indicator (*CLI*) for the euro zone countries, the *BOSE* indicator, return on the Warsaw Stock Exchange Index (*WIG*) and the average *EUR/PLN* exchange rates. Presented values for long run relationship between all proxies are in line with the literature and reasonable expectations. The bankruptcy rates increase as the birth rates for new companies, the inflation and *EUR/PLN* exchange rates decline. According to the adjusted R^2 the model fit is the lowest for the inflation and the highest for changes in exchange rates. In all estimated models the Chow breakpoint test points to possible structural breaks in the second quarter of 2009 – when the novelization was introduced. The *SPLITDUM*, which has a value of 1 after the new legislation came into effect is significant for all models. It is worth noticing that models with the new dummy got considerably higher adjusted R^2. The presented results suggest that aggregate bankruptcy rates after the novelization are higher than expected according to the macroeconomic conditions.

4.2. Model with multiple macroeconomic variables

In the last part of this chapter we decided to use two stage least square (*2SLS*) for proper estimation of coefficients. This approach allows us to control expected endogeneity issues which are caused by multicollinearity of macroeconomic variables. While selecting variables to the model we used backward regression to eliminate insignificant and correlated variables. Therefore, as a r esult we estimated models having only four explanatory variables and using all remaining variables defined in Table 1 as instruments. The first two selected variables are year-to-year change in average exchange rates *EUR/PLN* and yearly return on WIG, as both had the highest adjusted R^2 among models with single macroeconomic variables and both could be approximated by chosen instruments. Furthermore, only one general macroeconomic indicator is taken, as current economic condition indicator (*BOSE*) was shown to lead to the best fit. The jump dummy for change in bankruptcy law is also included. Table 4 reports the results of aggregate bankruptcy rates estimation using *2SLS* method for selected variables[10].

Table 4. Bankruptcy rates estimation: model with multiple macroeconomic variables

Variable	ΔBOSE	ΔWIG	ΔEUR/PLN	SPLITDUM
Cumulative lag coefficient	-0.00006**	-0.00381***	-0.00521***	0.00075***
Adjusted R^2	0.81141			

Notes: PDL model estimated by 2SLS with yearly change in quarterly business bankruptcy rates as dependent variable; instruments are: ΔNEW, ΔGDP, ΔCLIeu, ΔCLIpl, ΔINFL; only cumulative lag coefficients reported; results of t-test based on Newey-West HAC standard errors reported.

*** denotes significance at the 1% level; ** denotes significance at the 5% level; * denotes significance at the 10% level

Table 4 confirms the findings from our previous research based on *PDL* models with single macroeconomic variable. Each of the chosen macroeconomic conditions has proven to be significant[11]. Calculated adjusted R^2 is high and leads to conclusion that bankruptcy rates are well explained by the applied model. Estimated coefficients indicate that when *BOSE* indicator, Warsaw stock (*WIG*) and average *EUR/PLN* exchange rates increase, then in response the business bankruptcy rate (*BR*) declines. Nevertheless, the *SPLITDUM* dummy continues to be statistically significant with positive sign. It is worth to mention the correspondence between current results and those received from the previous

[10] As a robustness check models with different sets of explanatory variables have been estimated. In all of these models, findings remain similar with those presented.

[11] Moreover we performed the Sargan test for testing over-identifying restrictions and the null hypothesis that the over-identifying restrictions are valid was rejected with p-value 0.26.

subsection. Most importantly the result is in line with the main thesis of this paper which is the positive impact of the novelization in the bankruptcy law in the second quarter of 2009 on aggregate business bankruptcy rates in Poland.

5. Conclusions

The economic growth in Poland, as recorded since the nineties, resulted in the necessity of novelization in the bankruptcy legislation. The new bankruptcy law was introduced in 2003 and was originally designed in favour of the reorganization type of bankruptcy rather than liquidation. An efficient bankruptcy code should have allowed enterprises to be restructured. It included several international best practice recommendations adopted to the legal system. After six years it became clear that some of available formal procedures were not in use (i.e. they are superfluid) and the novelization of law was necessary. According to lawmakers new solutions improved insolvency procedures and made the whole bankruptcy process faster.

Our findings prove that after the Amendment bankruptcy rates in Poland increased more significantly than expected due to the existing macroeconomic conditions. This conclusion is confirmed by analyses based on t wo general concepts of measuring differences in aggregate bankruptcy rates. The novelization extended variety of insolvency rules. We suggest that the further effort should be applied to create even more transparent and efficient legal system. The Amendment not only encouraged to conduct reorganization, but also broadened the range of companies that could benefit from such reorganization. However, it is still impossible to start a r ecovery proceedings without an application for bankruptcy. Entrepreneurs often do not use this option, because the word 'bankruptcy' has a negative connotation in the colloquial language and in their opinion it could harm their businesses. As a r esult entrepreneurs often start bankruptcy proceedings when it is too late to save their businesses.

We recommend that future research should be concerned with the impact of differences in size and business sectors of firms on their failure rates.

Acknowledgement

We would like to thank the two anonymous referees for valuable comments and suggestions on earlier versions of this paper.

REFERENCES

ALTMAN, E. I., (1968). Financial Ratios, Discriminant Analysis and the Prediction of the Corporate Bankruptcy, *The Journal of Finance*, Vol. 23, pp. 589–609.

ALTMAN, E. I., (1983). Why business fail, *Journal of Business Strategy*, Vol. 3(4), pp. 15–21.

CHEN, J. H., WILLIAMS, M., (1999). The determinants of business failures in the US low-technology and high-technology industries, *Applied Economics*, Vol. 31, pp. 1551–1563.

CLAESSENS, S., KLAPPER, L. F., (2005). Bankruptcy around the World: Explanations of its relative use, *American Law and Economics Review*, Vol. 7, No. 1, pp. 253–283.

CUTHBERTSON, K., HUDSON, J., (1996). The determinants of compulsory liquidations in the UK, *The Manchester School of Economic and Social Studies*, Vol. 64, No. 3, pp. 298–308.

DEWAELHEYNS, N., VAN HULLE, C., (2008). Legal reform and aggregate small and micro business bankruptcy rates: evidence from the 1997 Belgian bankruptcy code, *Small Bus Econ*, Vol. 31, pp. 409–424.

EUROPEAN COMMISSION, (2003). BEST project on restructuring, bankruptcy and a fresh start: Final report of the expert, *Enterprise Directorate-General.*

EVERETT, J., WATSON, J., (1998). Small business failure and external risk factors, *Small Business Economics*, Vol. 11, No. 4, pp. 371–390.

FISHER, T. C. G., MARTEL, J., (2003). The effect of bankruptcy reform on the number of corporate reorganization proposals, *Canadian Public Policy*, Vol. 29, No. 3, pp. 339–350.

FOSTER, R., KAPLAN, S., (2001). Creative Destruction, *Financial Times.*

GREINER, L. E., SCHEIN, V. E., (1988). Power and Organization Development, *Addison-Wesley.*

HUDSON, J., (1986). An analysis of company liquidations, *Applied Economics*, Vol. 18, pp. 219–235.

HUDSON, J., (1997). Company bankruptcies and births matter, *Applied Economics*, Vol. 29, pp. 647–654.

ILMAKUNNAS, P., TOPI, J., (1999). Microeconomic and macroeconomic influences on entry and exit of firms, *Review of Industrial Organization*, Vol. 15, pp. 283–301.

JOHNSON, P., PARKER, S., (1994). The interrelationships between births and deaths, *Small Business Economics*, Vol. 6, No. 4, pp. 283–290.

KALLAUR, K., (2009). Bankruptcy and Remedial Act after recent changes, *Polish Construction Review*, Vol. 4(97), pp. 16–17.

LIU, J., WILSON, N., (2002). Corporate failure rates and the impact of the 1986 insolvency act: An econometric analysis, *Managerial Finance*, Vol. 28, No. 6, pp. 61–71.

LIU, J., (2004). Macroeconomic determinants of corporate failures: Evidence from the UK, *Applied Economics*, Vol. 36, pp. 939–945.

PETERSON, R. A., KOZMETSKY, G., RIDGWAY, N. M., (1983). Perceived causes of small business failures:A research note, *American Journal of Small Business*, Vol. 8, No. 1, pp. 15–19.

SCHUMPETER, J. (1934). The Theory of Economic Development: An inquiry into profits, capital, credit, interest and the business cycle, *Transaction Publisher*.

SULLIVAN, T. A., WARREN, E., WESTBROOK, J., (1999). Financial difficulties for small businesses and reasons for their failure, *US Small Business Administration Office of Advocacy Research Study*, No. 188.

VLIEGHE, G. W., (2001). Indicators of fragility in the UK corporate sector, *Bank of England Working Paper* 146.

WADHWANI, S. B. (1986). Inflation, default premia and the stock market, *The Economic Journal*, Vol. 96, pp. 120–138.

WORLD BANK, (2001). Principles and guidelines for effective insolvency and creditor rights systems.

ZDYB, M. (2008), Jakie czynniki generują upadłości przedsiębiorstw w Polsce? Przyczyny upadłości przedsiębiorstw w Polsce, *Biuletyn E-rachunkowość*.

ZEDLER, F., (2003). Prawo upadłościowe i naprawcze–wprowadzenie, *Zakamycze*.

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 467—476

REPORT

International Conference on Small Area Estimation (SAE 2014)

The international conference on Small Area Estimation (SAE 2014) was held from 3rd to 5th September 2014 in Poznan, and was devoted to the methodology of small area statistics, which, following arrangements made by the European Working Group on Small Area Estimation, was organized by the Department of Statistics at the Poznan University of Economics (PUE). The conference was co-organized by the Central Statistical Office (CSO) in Warsaw and the Statistical Office in Poznan. Professor Janusz Witkowski, President of CSO and Professor Marian Gorynia, Rector of the Poznan University of Economics took the Honorary Patronage over the conference. The conference was preceded by a special workshop devoted to using R in SAE conducted by Li-Chun Zhang from University of Southampton and Statistics Norway.

The conference was partially financed by R Revolution Analytics with the support of the National Bank of Poland granted under the program of economic education.

The Chairperson of the Organizing Committee of the Conference SAE 2014 was Marcin Szymkowiak from the Department of Statistics, PUE. Other members of the committee included: Wojciech Adamczewski and Katarzyna Cichońska from the Central Statistical Office in Warsaw, Tomasz Józefowski, Tomasz Klimanek and Jacek Kowalewski from the Statistical Office in Poznan. The Programme Committee of the conference was headed by Professor Domingo Morales (Universidad Miguel Hernández de Elche). Other members of the programme committee included Professors: Ray Chambers (University of Wollongong), Grażyna Dehnel (University of Economics in Poznan), Elżbieta Gołata (University of Economics in Poznan), Malay Gosh (University of Florida), Jan Kordos (CSO), Partha Lahiri (University of Maryland), Risto Lehtonen (University of Helsinki), Isabela Molina (Universidad Carlos III de Madrid), Ralf Münnich (University of Trier), Jan Paradysz (University of Economics in Poznan), Danny Pfeffermann (Hebrew University of Jerusalem), J.N.K. Rao (Carleton University) and Li-Chun Zhang (University of Southampton). The Steering Board of the SAE 2014 Conference was chaired by Domingo Morales and included Ray Chambers, Elżbieta Gołata, Partha Lahiri and Danny Pfeffermann.

The idea behind the SAE 2014 conference was to provide a platform for the exchange of ideas and experiences between statisticians, scientists and experts from universities, statistical institutes, research centers as well as other governmental agencies, local government and private companies involved in developing and applying the methodology of regional surveys, in particular, small area estimation. The Poznan conference was another one in the series of conferences (Jyväskylä 2005, Piza 2007, Elche 2009, Trier 2011), intended to combine theoretical considerations and practical applications of SAE in public statistics.

The SAE 2014 conference was focused on applications of SAE in censuses, model-based estimation and its evaluation, the use of spatio-temporal models, robust methods, non-response, issues in sample selection, poverty estimation, teaching SAE and its applications in public statistics. The conference featured a discussion panel and a specialist workshop devoted to the theory and practice of indirect estimation methodology.

The SAE 2014 conference was attended by 140 researchers and practitioners from 22 countries (Albania 2, Australia 6, China 1, Czech Republic 1, Finland 3, Spain 9, The Netherlands 3, Israel 1, Japan 2, Canada 2, Kuwait 1, Lithuania 3, Luxembourg 1, Malta 1, Germany 9, Norway 1, New Zealand 1, Poland 64, Thailand 1, Turkey 1, USA 15, United Kingdom 6, Italy 6). The participants represented both domestic and foreign research centers. The SAE 2014 conference was attended by scientists from 46 universities from around the world, representatives of statistical offices from nearly 30 countries, as well as representatives of international scientific organizations, the World Bank, the Central Statistical Office and statistical offices from Poland.

The conference brought together some eminent experts in this field of statistics including: J.N.K. Rao, Malay Ghosh, Ray Chambers, Li-Chun Zhang, Partha Lahiri, Danny Pfeffermann, Risto Lehtonen, Ralf Münnich, Domingo Morales, Graham Kalton (Westat), Wayne Fuller (Iowa State University) and Isabel Molina.

The conference featured two plenary sessions and special lectures given by Professors J.N.K. Rao and Malay Gosh. During the lecture entitled **Inferential Issues in Model-Based Small Area Estimation: Some New Developments** Professor J.N.K. Rao discussed developments in the theory and applications of small area estimation which have taken place especially over the past 15 years in response to a growing demand for reliable small area statistics. In particular, the lecture addressed some recent important developments concerning area level and unit level models, mainly addressing issues related to assumed models. During his talk Professor Rao specifically focused on the bootstrap methods for mean squared error (MSE) estimation and confidence interval construction. Other subjects raised during the talk included recent work on robust estimation of small area means, informative sampling, new developments in model selection and checking, methods for the estimation of complex parameters such as small area poverty measures and the role of ’big data’ in small area estimation.

Professor Malay Ghosh’s lecture, entitled **Benchmarked Empirical Bayes Estimators for Multiplicative Area Level Models,** was devoted to empirical Bayes and benchmarked empirical Bayes estimators of positive small area means under multiplicative models. In his presentation, he discussed the transformed Fay-Herriot model as a multiplicative model for estimating positive small area means and a weighted Kullback-Leibler divergence as a loss function. Professor Malay Ghosh demonstrated that the resulting Bayes estimator is the posterior mean and that the corresponding benchmarked Bayes and empirical Bayes estimators retain the positivity constraint. The prediction errors of the suggested empirical Bayes estimators were investigated asymptotically, and their second-order unbiased estimators were provided. In addition, bootstrapped estimators of these prediction errors were also provided. The performance of the considered procedures was investigated by the author through simulation and in an empirical study.

The SAE 2014 conference included ten invited sessions organized by top specialists:

- SAE: robust and nonparametric methods (Professor Ray Chambers, University of Wollongong),
- Small Area Methods for Repeated Survey (Professor Partha Lahiri, University of Maryland),
- SAE in poverty mapping (Professor Isabela Molina, Universidad Carlos III de Madrid),
- SAE models: selection and checking (Professor Danny Pfeffermann, Hebrew University of Jerusalem),
- SAE in official statistics (Professor Jan Kordos, CSO),
- Teaching SAE (Professor Risto Lehtonen, University of Helsinki),
- SAE applications (Professor Ralf Münnich, University of Trier),
- Benchmarking, design issues and nonresponse in SAE (Professor Stefano Falorsi, ISTAT),
- Population Census and SAE (Professor Li-Chun Zhang, University of Southampton),
- Other topics related to SAE (Professor Domingo Morales, Universidad Miguel Hernández de Elche).

During the session **SAE: robust and nonparametric methods** organized by Ray Chambers, and chaired by Graham Kalton, four papers were presented:

- Raymond Chambers − Two Recent Developments in Robust and Semiparametric Small Area Estimation,
- Beate Weidenhammer, Nikos Tzavidis, Timo Schmid, Nicola Salvati − Domain Prediction for Counts using Microsimulation via Quantiles,

- Payam Mokhtarian − On Outlier Robust Small Area Prediction of the Empirical Distribution Function,
- Forough Karlberg − Small Area Prediction for Skewed Data in the Presence of Zeroes.

In particular, Ray Chambers presented new developments in modeling for small area estimation including the spatial extension of recently published results on robust bias correction when asymmetric unit level and area level outliers in the survey data are used to predict a small area mean. Professor Chambers specifically focused on the extension of M-quantile modelling for small area estimation for count data rather than realizations of continuously distributed variables. Three other presentations were devoted to different aspects of robust and nonparametric methods in small area estimation and included the problem of estimation when outliers occur, the problem of estimation for asymmetric distribution with zeroes and using quantiles for the purpose of prediction of counts in the context of microsimulation.

The invited session on **Small Area Methods for Repeated Survey,** organized by Partha Lahiri and chaired by Wayne Fuller, consisted of four presentations:

- Partha Lahiri − An Overview of Small Area Estimation with Repeated Survey Data,
- Jan A. van den Brakel, Sabine Krieg − Small area estimation with state-space common factor models for rotating panels,
- Enrico Fabrizi, Maria Rosaria Ferrante, Carlo Trivisano − Estimation of value added for firms cross-classified by region, industry and size using repeated survey data,
- Carolina Franco, William R. Bell − Alternative Approaches to Borrowing Information Over Time in Small Area Estimation with Application to Data from the Census Bureau's American Community Survey.

This session provided an overview of different small area estimation methods for repeated surveys. In particular, the main presentation, given by Partha Lahiri, highlighted the fact that repeated surveys not only offer opportunities for improving small area statistics that are usually produced in cross-sectional surveys, but they may also deliver reliable estimates of changes over time, which may be more important than estimating current time. Professor Lahiri also pointed out that repeated surveys could conveniently help statisticians explain the benefits of small area statistics to public policy makers. Three other presentations addressed different aspects of small area estimation methods for repeated surveys and covered such issues as modeling for rotating panels, estimation of value added in business statistics and the problem of borrowing strength over time.

The third invited session on **SAE in poverty mapping**, organized by Isabel Molinaand chaired by Monica Pratesi, was devoted to issues connected with poverty in the context of small area estimation methodology. This session consisted of four presentations:

- Isabel Molina, J.N.K. Rao − An overview of small area estimation methods for poverty mapping,
- Gauri Datta, Abhyuday Mandal − Small area estimation with uncertain random effects,
- Domingo Morales − Partitioned area-level time models for estimating poverty indicators,
- Roy Van der Weide − Estimation of Normal Mixtures in a Nested Error Model with an Application to Small Area Estimation of Poverty and Inequality.

The main purpose of this session was to show how different techniques offered by small area estimation can be used in the field of poverty. This is especially very important for many institutions which have to conduct more effective and efficient policy at the regional level. During this session the main approaches for small area estimation techniques for poverty mapping were reviewed and their advantages and disadvantages were discussed. In particular, special attention was given to recent variants of the basic methods in the field of poverty mapping and inequality.

The invited session on **SAE models: selection and checking**, organized by Danny Pfeffermann and chaired by Elżbieta Gołata, consisted of four presentations:

- Danny Pfeffermann − Model Selection and Checking for Small Area Estimation, Graham Kalton – discussant,
- Jay Breidt, Daniel Hernandez-Stumpfhauser, Jean D. Opsomer − Variational Approximations for Selecting Hierarchical Models of Circular Data in a Small Area Estimation Application,
- Jiraphan Suntornchost, Partha Lahiri − Variable selection for Linear Mixed Models with Applications in Small Area Estimation,
- Yahia El Horbaty − A Simple Score Test for Random Effects with Application to Small Area Models.

The main aim of this session was to present recent developments in the field of modeling in small area estimation methodology. In his paper, Professor Danny Pfeffermann gave an overview of some methods proposed in the literature for small area model selection and checking, distinguishing between frequentist methods and Bayesian methods. He also discussed some issues related to the theoretical foundation of small area estimation models and in particular, the interpretation and role of the random effects. Three other presentations in this session were devoted to practical aspects of using proper chosen models in different surveys.

The invited session on **SAE in official statistics**, organized by Jan Kordos and chaired by William Bell, featured four presentations:

- Jan Kordos − Small area estimation in official statistics and statistical thinking,
- Danute Krapavickaite, Tomas Rudys − Application of small area estimation methods for Lithuanian Labour force survey data,
- Jan Paradysz, Karolina Paradysz − Indirect estimation of disability on the base of Polish National Census 2011,
- Jan Kubacki, Alina Jędrzejczak − Small area estimation under spatial SAR model.

This session was a response to the growing role of small area estimation methodology in official statistics. In the main presentation, Professor Jan Kordos outlined the general mission of national statistics institutes to produce high quality statistical information on the state and evolution of the population, the economy, the society and the environment. Professor Kordos paid special attention to the so called statistical thinking in the context of small area statistics and Total Quality Management. He also presented selected applications of Small Area Estimation procedures in official statistics in the context of an increasing demand for information. Other presentations in this session were more practical and focused on applications of SAE methodology, with particular emphasis on issues related to labour market and disability.

The invited session on **Teaching SAE**, organized by Risto Lehtonen and chaired by Gauri Datta, was devoted to different aspects of teaching small area estimation methods and consisted of four presentations:

- Risto Lehtonen − Experiences and challenges in teaching small area estimation,
- Jan Pablo Burgard, Ralf Münnich − SAE teaching using simulations,
- Elżbieta Gołata, Tomasz Klimanek − Challenges faced by academics and the NSI in SAE education,
- Esther Lopez Vizcaino Lombardía Cortiña, M. José, Domingo Morales − mme: An R package for small area estimation with multinomial mixed models.

This session was a response to the problems and issues related to the basic and fundamental question of how to teach small area estimation methodology at universities and within statistical offices. In the main presentation, Professor Risto Lehtonen argued that SAE teaching should be treated as one of the main components of the 'ecosystem', which consists of scientific conferences devoted to SAE, textbooks related to SAE, SAE chapters in edited books and hundreds of journal articles, active research groups, large-scale international research projects and programs, geo-coded and spatio-temporal databases, 'big data' sources and a variety of software tools for computing and graphical illustration. In this context some selected aspects of teaching SAE including problems, challenges and experiences were discussed in detail. Three other presentations in this session looked at different aspects of teaching SAE at universities and, in particular, raised issues related to using simulations while teaching SAE, using selected R packages in this field and challenges faced by the system of education in terms of the needs of both academics and statistical offices.

The next invited session on **SAE applications** was organized by Ralf Münnich and chaired by Roy van der Weide from the World Bank. This session consisted of four presentations:

- Ralf Münnich − Small area applications: some remarks from a design-based view,
- Ugarte, MD, Adín, A., Goicoa, T., Militino, A.F., López-Abente, G. − Space-time analysis of young people brain cancer mortality in Spanish provinces,
- Rebecca C. Steorts − Constrained Smooth Bayesian Estimation,
- William R. Bell, Mark Seiss − A Modeling Approach to Estimating the Mean Squared Error of Synthetic Small Area Estimators.

This session was mainly devoted to different SAE applications using real data. From one point of view, there are only few National Statistical Institutes which use the SAE methodology in the production of statistical data. The reason is the difficulty of using model-based techniques in the production of small area estimates. On the other hand, statistical offices are increasingly responsible for delivering estimates at a lower level of spatial aggregation. This calls for applications using real data and taking into account practical situations which are faced by statistical offices.

In the main talk, Professor Ralf Münnich highlighted the impact of sampling designs on small area estimation methods. He also presented real applications of using small area estimation methods in the context of household and business data. In addition to sampling designs, Professor Münnich also considered methods of benchmarking in order to provide coherent results between design-based and model-based estimates. Three other presentations in this session were devoted to different SAE applications and included: analysis of young people brain cancer mortality in Spanish provinces, analysis using data coming from U.S. Census's Small Area Income and Poverty Estimates program and application of the modeling approach to a real application involving synthetic estimation of correct enumerations in the 2010 U.S. census using data from a post-enumeration follow-up survey.

The next invited session on **Benchmarking, design issues and nonresponse in SAE** was organized by Stefano Falorsi and chaired by Michel Hidiroglou. This session consisted of four presentations:

- Andrea Fasulo, Michele D'Alo', Lorenzo Di Biagio, Stefano Falorsi, Fabrizio Solari − Benchmark constraints for space and time unit level EBLUP estimators,
- Li-Chun Zhang, Alison Whitworth − Benchmarked synthetic small area estimation,
- Serena Arima, Gauri S. Datta, Brunero Liseo − Multivariate Fay-Herriot model with structural measurement error,
- Janusz Wywiał − On sampling design proportional to function of auxiliary variable order statistics.

Some very important topics were raised during this session, which related to the negative impact of nonresponse in the process of estimation, benchmarking and design issues. In the main presentation delivered by Fasulo *et al*., the authors focused on small area estimators based on unit level nonparametric mixed models with area random effects. They also considered the benchmark problem for SAE estimates, which was consistently extended to the case of space and time benchmark constraints. The presenters demonstrated practical applications of the issues raised in their presentation by reviewing two empirical studies and presenting their conclusions. The three other presentations were devoted to the problem of benchmarking, which is very crucial in production of statistical information as estimates for lower level of aggregations should add up to estimates at higher level, the problem of modeling using multivariate Fay-Herriot approach with structural measurement error and issues to do with basic properties of sampling strategies based on the sampling designs dependent on quintiles.

The invited session on **Population Census and SAE,** organized by Li-Chun Zhang and chaired by Stephen Haslett, included four presentations:

- Li-Chun Zhang − Census and SAE: Population size estimation,
- Ralf Münnich − Small area estimation in the German Census 2011,
- Paul Williamson, Karyn Morrissey, Ferran Espuny-Pujol − Survey reweighting as a means to SAE,
- Angela Luna-Hernandez, Li-Chun Zhang − Multivariate Generalized Structure Preserving Estimation.

The main aim of organizing this session was to present how small area estimation methodology can be used in the field of modern censuses, in which data are often collected using the mixed approach. In the main presentation delivered by Li-Chun Zhang, the author focused on the topic of census or census-like population size estimation. The presentation reviewed common traditional direct estimation methods, as well as some new developments in the treatment and modeling of enumeration coverage errors. Prof. Li-Chun Zhang also discussed some perceived challenges to the indirect estimation of local population size as well as the question of how design-based and model-based estimation can be used in the context of modern censuses. In the next talk, Ralf Münnich discussed the use of small area estimation methodology in the German census

2011. Some aspects of survey reweighting and SPREE estimation were also discussed as part of this session.

The last invited session, entitled **Other topics related to SAE**, which was organized by Domingo Morales and chaired by María Dolores Ugarte, included four talks:

- Wayne Fuller, Andreea Erciulescu − Small Area Prediction under Alternative Model Specifications,
- Domingo Morales, Miguel Boubeta, María José Lombardía − Empirical best prediction in Poisson mixed models,
- M.A. Hidiroglou, Victor Estevao − A comparison of small area and direct estimators via simulation,
- Monica Pratesi, Fosca Giannotti, Caterina Giusti, Stefano Marchetti, Dino Pedreschi, Nicola Salvati − Area level SAE models with measurement errors in covariates: an application to sample surveys and Big data sources.

This session mainly covered issues related to small area estimation methodology and not discussed in detail in the others. During this session special attention was paid to issues concerning modeling in the field of SAE. That was the topic of the main presentation given by Wayne Fuller and Andreea Erciulescu. In their talk, they discussed the construction of small area predictors and estimation of the prediction mean squared error, given different types of auxiliary information and for different population models and illustrated the problem with a study of soil erosion. The three other presentations also dealt with modeling and addressed such topics as the use of Poisson mixed models, area level SAE models with measurement errors in covariates and the comparison between SAE and direct estimators using a simulation approach.

There were also six sessions of contributed papers organized during the SAE 2014 conference, which covered different issues related to small area estimation methodology. Some of the topics covered included modeling in SAE, SAE applications, poverty mapping, SAE in business statistics and the use of Big Data in the context of SAE. In total, 72 talks and 10 poster presentations were delivered during the conference.

One of the highlights of the SAE 2014 conference was the panel discussion organized and chaired by Professor Elżbieta Gołata from the Department of Statistics at the University of Economics in Poznan, which addressed the newest achievements in SAE both in the theoretical and practical field. The panel brought together nearly all of the organizers and chairs of the invited sessions, who gave an account of the most important issues raised within the ten invited sessions. The panel constituted a scientific summary of the whole conference and was a great opportunity to review recent developments both in the field of theoretical and practical use of SAE.

Detailed information about the SAE 2014 conference is available on the conference website at www.sae2014.ue.poznan.pl.

The plan to organize future SAE conferences is an expression of the growing role of small area estimation methodology in the modern statistical world.

It is worth noting that the next SAE 2016 conference in the series of conferences under the auspices of the EWORSAE group is planned to be held in the Netherlands.

Moreover, the First Latin American International Statistical Institute Satellite Meeting on Small Area Estimation will be held on August 3-5, 2015, in Santiago, Chile. More details about this conference can be found at www.encuestasuc.cl/sae2015.

Prepared by:
Marcin Szymkowiak
Poznan University of Economics

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 477–486

BOOK REVIEW

***Sampling Elusive Populations: Applications to studies of child labour,* by Vijay Verma.** International Labour Organization, Geneva. 2013. XIX + 821 pp. ISBN 978-92-2-128321-8.

Reviewed by Włodzimierz Okrasa

Surveying populations difficult to find or to reach *(hard-to-reach),* elusive or invisible, mobile or intentionally avoiding statistical observation for some reasons, presents not only additional problems to those being faced by a survey maker in the context of researching regular populations. The problems involved in the former are often qualitatively different from those in the latter, at practically each stage of the process, from sampling populations with non-existing or imperfect frame, through data collecting and processing, to estimating the looked-up parameters. Comparison of methodologies focused on researching basically the same units – such as child workers – in both types of contexts can serve as a laboratory case for demonstrating differences between the problems specific to surveying regular and irregular populations. We are now in a perfect position for making such type of methodological comparisons as Vijay Verma – the author of the widely acclaimed book (cf. Kordos' review in this journal, 2008) on *Sampling for household-based surveys of child labour* prepared within the framework of the ILO Statistical Information and Monitoring Programme on Child Labour (SIMPOC) – has elaborated a complementary volume on *Sampling Elusive Populations: Applications to studies of child labour,* also published by the International Labour Organization.

This book offers the comprehensive treatment of methods specifically designed to meet challenges posed by the irregular type of populations, in general, with extensive presentation of their applications to formal and practical issues in researching children's work. As such, the book fills a key gap in the survey methods literature, on the one hand, while promoting the approach especially suited to dealing with this complex research objects. In a consequence, its usefulness goes far beyond studying child labour within the public statistics, but contributes also to economic and social analysis of labour market and work activities. With respect to the former, the book addresses also several problems fundamental to the methodology of surveying hard-to-survey populations, a "common but not universal characteristic of (which) is that they are rare populations for which no separate sampling frames exist" (Kalton, 2014, p. 401). Such populations are becoming of growing interest to statisticians inspired also by economists, demographers and other social scientists in view of the fact that in addition to the problems with forming a list of potential respondents – hard-to-sample populations – they are more and more frequently being faced with insurmountable sometimes difficulties to persuade members of such populations to take part, or to be interviewed. A good indication of the rising demand for operating knowledge, i.e., for manuals on dealing with such a type of challenges and problems – which are covered jointly by the term hard-to-survey populations – might be a book j ust under the title *Hard-to-survey populations*, edited by Roger Tourangeau et al., 2014, which was published recently. In addition to the first of the two ILO-issued books by Vijay Verma, this seems to complement its contents and the expected audience from the side of general survey methodology, including the post-sampling survey making procedures.

At a glance, the enormous scope and richness of the issues addressed in Vijay Verma's book (*Sampling Elusive Populations* – for convenience an abbreviation SEP will stand for it in this text) can best be characterized by indicating its size – 821 pages – and composition: fourteen chapters, written as self-contained thematic sub-manuals. With his extraordinary experience and knowledge of methods of surveying child labour – including awareness of the fact that "collecting comprehensive data on child labour is a challenging task, and no single survey method can satisfy all data needs " (SEP, p. 1) – the Author begins with useful introduction to the context of the questions covered by his manual. Starting with sketching main differences between household-based child labour survey – which misses working children not living at home – and the specificity of such an elusive population as presented by labouring children (ill-defined population of a great heterogeneity, rare and highly mobile, partly hidden and reclusive, staying away from participation in a survey) the Author is constructing – not simply applying – a methodology of sampling especially suited to accounting for these characteristics. His understanding of the basic concept – elusive population – is given explicitly: "By elusive populations we mean *populations for which – by virtue of their characteristics, or of the lack of suitable sampling frames, or difficulties in obtaining the required information – adequate samples cannot be defined, drawn or implemented using the normal procedures of general population sampling."* (*ibid.,* p. 4). Since the problems of under-coverage or of non-response may occur in a survey of populations considered 'normal', they also may be added to the *most elusive* population groups given that they are the ones of the greatest policy concern. The typology of populations in terms of characteristics important from sampling standpoint (following Kish, 1991), along with the specification of the nature of the methodological problem involved, organize the book contents of this book.

Accordingly, this well-organized (reader-friendly) manual – which is conceived as a complementary to the previous one – concentrates on various groups of working children not living at home. Their key characteristic is a degree of connection to their household as the starting point for being identified or traced in a way. Also, some categories of children living at home and working away are included due to insufficient information available to their parents/guardians on the location or type of work they are doing. All the possible to imagine situations which can appear in a variety of configurations, given different circumstances of work and relations to the household, make the survey (sample) designer open to any of the theoretically envisioned approaches, including the need to enumerate working children at the place of their work based on a sample of such places.

An overview of the book's contents

After providing the reader with preliminary ideas and concepts of working children as an example of elusive population – an object of chapter 1 – the Author concentrates on systematic presentation of available sampling methods, in a way facilitating also the proper choice of the one, most effective in the concrete problem context.

In chapter 2, ***Child labour situations, data needs and sources,*** diverse and specific to the particular situation child labour problems are discussed as the understanding of the situation to be studied is considered by the Author essential for choosing an appropriate survey methodology and sampling strategy. The chapter illustrates the variety of situations and types of child labour in order to provide the necessary background for the diverse sampling techniques discussed in subsequent chapters. The variety of forms of child labour covers diverse sectors: child domestic work; agriculture including commercial crops; fishing and aquaculture; mining and quarrying; manufacturing including handicrafts; construction; street work and the informal sector; and also various 'unconditionally worst forms of child labour' including child trafficking, commercial sexual exploitation, forced or bonded labour, engaging or living in armed conflict, and children’s involvement in illicit activities, in particular in drug trafficking.

It is followed by an overview of different data sources for different types of child labour: household-based surveys; supplementary sources or surveys (school-based surveys, community-level inquiries, general national household surveys, censuses, other secondary sources); employers' surveys; establishment surveys; baseline surveys and studies; and rapid assessments. Two major strategies of generating data on child labour – household-based surveys and rapid assessment studies – are characterized briefly. The former as a large probability sample of the general population; the latter as a small-scale but intensive survey. The two form two ends of the range of application of the various sampling techniques addressed in the book.

Chapter 3, ***Basic sampling and estimation procedures***, is devoted to providing a reminder of some basic principles concerning sample design and selection which underlie the more specialised techniques discussed in this book, such as: principles of probability sampling, common departures from simple random sampling (stratification, clustering, unequal selection probabilities), probability proportional to size (PPS) sampling, and systematic sampling. Also, the chapter reviews basic principles concerning weighting of sample data and estimation from a sample, along with sources of information for weighting and presentation of a step-by-step procedure for weighting (computation of design weights, adjustment for non-response, calibration against external standards, and trimming and scaling of the weights).

The problem of sampling from imperfect frames is discussed in chapter 4, ***The sampling frame***. It starts with reviewing shortcomings of sampling frames and basic concepts (the survey population, the sampling frame for single-stage and for multi-stage sampling). The surveys cover selected sectors; most of the establishments in the sectors are small and a high proportion employ child labour. Several common aspects concerning sampling frames are presented using an illustrative material – the problems such as surveying a population in the absence of an existing sampling frame; including the cost and quality implications of the quantity of information to be collected for each unit during the operation; economising research by sharing the costs between different surveys; using the listing operation for making substantive estimates; and special problems related to the type of units in the frame (e.g. establishments versus other locations where working children are found). Basic requirements and desirable quality, efficiency and cost-related properties of *area frames* are also discussed. The problems of *list frames* are considered from a p ractical perspective, focusing on the correspondence between listing, sampling and analysis units.

The second theme of the chapter is developing and explaining such important the concepts as: (i) correspondence between sampling and analysis units (any analysis unit is associated with *at most one* sampling unit in *direct sampling,* otherwise it would be 'indirect sampling'); (ii) *sampling with multiplicity*, as the multiplicity estimator links many of the sampling techniques discussed in this book. Situations when the sample has to be obtained by exploiting links between analysis units themselves, *link trace sampling,* are also addressed (for being discussed in later chapters).

Chapter 5, ***Sampling establishments employing children,*** discusses sampling aspects which apply equally to both small and informal sector establishments and to larger establishments, as they differ significantly in terms of sampling considerations and procedures. The main difference concerns the selection of establishments – for large and medium-sized establishments samples are often selected directly from lists, and the chapter describes sampling procedures for this type of selection procedures. However, the second type of design concerns samples for surveys of small and informal sector establishments which, like households, are small-scale, numerous and widely dispersed in the population. The commonly used samples for small and informal sector establishments are area-based and involve two (or sometimes more) stages of sampling. The technical issues discussed include: (a) characteristic features of small and informal sector establishments and their consequences for survey design; (b) the choice between integrated multi-sectoral and separate single-sector surveys; (c) stand-alone versus surveys attached as m odules to other surveys; (d) the construction and use of 'strata of concentration' of different types and sectors of establishments to control distribution of the sample; (e) procedures for selecting establishments within sample areas; and (f) issues in survey implementation.

The rare populations are discussed in chapter 6: ***Sampling rare populations,*** the characteristic feature of w hich is that sampling the whole population with normal procedures does not yield a representative sample of adequate size for the subpopulations of interest because of their small size. In surveying different types of child labour, the rare populations of interest – working children – are generally unevenly distributed among the general population of children. The Author discusses five aspects of the strategy: (1) locating concentrations of the rare population using existing large-scale sources; (2) partitioning the frame according to the degree of concentration of the rare population (using different techniques); (3) oversampling strata of concentration, making use of the patterns of concentration identified; (4) listing, screening and two-phase sampling, aimed at the identification and sampling of the final elements (households, children); (5) employing special procedures to increase selection probabilities of units in the rare population and thereby increase the achieved sample size. There are a number of other procedures discussed in subsequent chapters (such as multiplicity sampling, multiframe sampling and adaptive cluster sampling) and the common link between them is that they involve *sampling with multiplicity.*

Multiplicity sampling is discussed in chapter 7, focused on situations where the approach may be useful in surveying the rare populations of labouring children. Since the basis of multiplicity sampling is the relationship between sampling units and analysis units, sampling with multiplicity arises when an analysis unit is linked to *more than one sampling unit.* The Author discusses potential advantages and uses of multiplicity sampling, identifying situations where multiplicity sampling may be useful, but also addresses its limitations and the problems of the method. For instance, reporting biases are often larger for multiplicity counting rules than for ordinary unitary counting rules. Another concern is the increased complexity. There can also be serious ethical, confidentiality and privacy concerns in using the method. Procedures for estimation with multiplicity sampling are also explained in this chapter. The standard *multiplicity estimator* takes the weights as inversely proportional to the unit's multiplicity.

The next, 8th chapter, is devoted to ***Multi-frame sampling,*** discussed in the context of child labour surveys, generally, in order to reduce coverage errors when no single sampling frame can provide a complete representation of the target population. Typically, the multiple sampling frames overlap and procedures such as constructing a new single frame without duplicates or by accounting for the duplicates in the estimation procedure need to be used. Therefore, the Author presents the main methods of removing the duplicates and constructing non-overlapping frames, as well as the main procedures for accounting for duplications and estimation from overlapping frames. Both types of situation when multiple frames can involve multiplicity in the selection of units are considered: either a unit may appear in more than one frame, or within any of those frames the unit may appear more than once. The chapter considers practical aspects of implementing this procedure in the context of a child labour survey.

Adaptive cluster sampling discussed in chapter 9 is a technique designed to obtain more adequate and efficient samples for a population which is rare and very unevenly distributed. The technique specifically involves selecting an additional sample in the neighbourhood of points where a concentration of the population of interest is found during implementation of the initial sample. It is presented as being most effective when the population of interest tends to be concentrated in relatively few and large clusters, but little information is available on the extent, location and patterns of its concentration – e.g., such populations include street children, children engaged in street trades and child beggars. Several technical aspects are discussed as well, such as: unequal unit selection probabilities; stratification with adaptive sampling; multistage sampling; multivariate criteria for adaptive sampling; adaptive sampling using 'order statistics'; arbitrary rules for stopping the adaptive process; problem of imperfect detectability; and aspects of the estimation procedures with adaptive sampling. In addition to discussing issues involved in its implementation, the procedure is illustrated in detail on the basis of an artificially constructed small population. The illustration demonstrates how adaptive sampling can help in locating large concentrations of the population of interest by increasing the chance of their appearance in the sample, and hence also in obtaining a larger number of elements of interest (such as children working in a particular sector).

Sampling mobile populations is described in chapter 10, i ncluding special problems and issues which arise in this approach, while stressing that the concept of 'mobile population' is more general than simply not having a fixed place of residence or work – sometimes it is *necessary* or *preferable* to sample and enumerate units through their *mobility* (movement).

The following questions are involved in difficulties of enumerating such populations: (i) who are the eligible respondents for the survey, and (ii) where and (iii) when to find them; also (iv) what information concerning their mobility to ask them for, and (v) how to obtain the information; (vi) how to use sample data to produce valid estimates for the population, and (vii) how to assess variances and biases to which those estimates are subject. The Author develops a framework to organise the variety of circumstances, problems and solutions encountered in sampling mobile populations – four important concepts in the framework are: sampling locations, observation points, time segments, and 'time-location primary sampling units'. Also, procedures for estimating the *probability of selection and sample weight* of a mobile individual are developed, along with quantitative expressions for variations in individual selection probabilities in a number of commonly encountered situations.

The approach discussed in chapter 11, ***Capture-recapture sampling***, is devoted to sampling techniques which involve taking two (or more) independent samples from the same population and using the overlap found between the samples to estimate the selection probabilities applied to obtain those samples and the total population size. Capture-recapture applications in the social field are usually based on a combination of sample surveys and administrative sources. The Author provides instructive illustrations of application of this technique, stressing its usefulness (and robustness) even in the situation of departures from the assumed statistical model, and the fact that statistical procedures have been developed to control the effect of certain departures from the original simple model. In the Author's view, a major technical contribution of the chapter concerns the development of procedures for the estimation of sample weights in a more general situation, along with explanation of procedures for putting together all these effects.

Controlled selection and balanced sampling discussed in Chapter 12 is a procedure to control the structure of the sample beyond what is possible with ordinary independent selection within strata. Surveys, in particular of mobile and other difficult-to-access populations, often have to be restricted to a limited area and to a small number of primary units. The Author provides several arguments for using controlled selection, especially when one has to select a small sample of primary units, but at the same time ensures that it is 'balanced' and 'representative' of the population in terms of many characteristics (or control variables). He also discusses this procedure in the context of the modern theory of balanced sampling, thus providing the possibility of dealing with a wider range of issues and more efficient sampling algorithms. These control variables may include one or more stratification variables, which correspond to controlled selection. The formal considerations are complemented by illustrations useful for a reader interested in practical applications.

The reclusive populations are discussed in chapter 13, ***Snowball sampling***, which is meant as an approach to surveying reclusive populations of labouring children. In particular, the term snowball sampling refers to a convenience sampling mechanism in settings characterised by the lack of a serviceable sampling frame. A unit of the target population can enter the sample through direct selection into the initial sample, or by being identified ('named') for inclusion by someone already in the sample. There are a number of parameters which define the design of a snowball sample: the number of waves, number of contacts to request, and criteria for including a participant in the sample. The Author considers it especially useful in the context of an exploratory study, but accents its recent development and advantages within a more advanced analysis. The primary advantage of the method is its success in identifying individuals from unknown populations and from small, hidden groups dispersed within a large population; also, it provides a means of accessing social groupings which are vulnerable, etc.

Noting also its deficiencies – in particular selection bias which limits the validity of the sample – the Author outlines a simplified procedure for estimating size of a hidden population. A number of illustrations from surveys in diverse settings present both positive and negative experiences in applying the method.

Respondent-driven sampling (RDS), discussed in the last chapter (14th), is considered to be an improved variant of the usual snowball sampling, as both procedures are types of *chain-referral* sampling. As with snowball sampling, a unit of the target population can enter the sample through direct selection into the initial sample, or by being identified for inclusion by someone already in the sample. In both cases, the process starts with a small number of peers, usually chosen non-randomly. However, as an improvement over ordinary snowball sampling, RDS is designed to produce a closer approximation to probability sampling. It incorporates features such as *the direct recruitment of peers by their peers,* a *dual system of incentives* (for participation and for recruiting), and *recruitment quotas* (e.g. a maximum of three recruits per respondent). As regards the assessment of the RDS procedure, the chapter analyses the experience with two sets of studies. The first set involves performance comparisons of the RDS and alternative sampling approaches. In the second set of examples, studies concerning assessment of RDS examine how well its procedures are implemented in terms of the theoretical assumptions of the model. For both of them numerous examples of studies undertaken for assessing comparative performance and validity of the RDS techniques are provided.

A set of three annexes and references – bibliography, author index and subject index – concludes impressive contents of this very professionally prepared book, both by the Author and the ILO editors. This book is for anyone interested in researching the labouring children, and also for all interested in surveying elusive populations, in general – students and academicians, as well as policy makers and practitioners. As mentioned earlier, it is a self-contained manual, providing the reader/user with a great piece of the subject-matter knowledge and an advanced methodology at work.

REFERENCES

KALTON, G., (2014). Probability sampling methods for hard-to-sample populations. Chapt. 19 [in] Roger Tourangeau, Brad Edwards, Timothy P. Johnson, Kirk M. Wolter, and Nancy Bates (Eds.) Hard-to-Survey Populations. Cambridge University Press, Cambridge, UK.

http://www.amazon.com/gp/reader/1107031354/ref=sib_dp_pt#reader-link.

KISH, L., (1991). Taxonomy of elusive populations, in Journal of Official Statistics, 7(3), 339–347.

KORDOS, J., (2008). Book review. Sampling for household-based surveys of child labour, by Vijay Verma, Statistics in Transition new series, Vol. 9, No. 3, pp. 587–590.

TOURANGEAU, R., EDWARDS, B., JOHNSON, T. P., WOLTER, K. M., BATES, N. (Eds.), (2014). Hard-to-Survey Populations. Cambridge University Press, Cambridge, UK.

VERMA, V., (2008). Sampling for household-based surveys of child labour prepared within the framework of the ILO Statistical Information and Monitoring Programme on Child Labour (SIMPOC).

STATISTICS IN TRANSITION new series, Summer 2014
Vol. 15, No. 3, pp. 487–490

ABOUT THE AUTHORS

Basiura Beata graduated with master of science degree in mathematics from the Jagiellonian University of Cracow. In 2007 s he defended her PhD thesis in economics at the Faculty of Management Cracow University of Economics. She works in the Faculty of Management AGH. She is interested in the development of statistical methods for high-dimensional and correlated data. In particular, the issues of taxonomy and classification of time-series are the main topics of her research.

Bhattacharjee Atanu was born in Durgapur, India, to Swapan Bhattacharjee and Sandhya Bhattacharjee. He attended A-Zone M.P School and T.D.B. College before joining the Institute of Medical Science, Banaras Hindu University, where he obtained master's degrees in health statistics in 2008. He obtained his PhD in statistics in 2012 at the Gauahati University under Dilip C Nath (statistics). His dissertation was titled "Bayesian Analysis for Longitudinal Data on Type 2 Diabetes Patients". After that, he joined the Ward of Clinical Research and Biostatistics at Malabar Cancer Centre (MCC) as a L ecturer. Over the last few years, A. Bhattacharjee has published more than 35 articles, in various peer-reviewed reputed journals.

Czapkiewicz Anna works at the Faculty of Management at AGH. She defended her PhD thesis in economics at the University of Lodz. She investigates the usefulness of Copula-GARCH models in modelling the relationship between returns of main indices of the stock markets. In these studies, she uses both static and dynamic models. The time-varying dependencies between time series may be explained by the regime switching process.
The author studies the regime switching copula model with a Markov switching mechanism for modelling financial time series.

Dihidar Kajal is an Assistant Professor at Indian Statistical Institute since February 01, 20 13. She graduated from Calcutta University with Mathematics Honours in B.Sc. in 1984. She post-graduated from Statistics in M.Stat. at Indian Statistical Institute in 1987. She holds Ph.D. in Statistics (in the area of Survey Sampling) under the supervision of Prof. Mausumi Bose of Applied Statistics Unit of Indian Statistical Institute (2010).

The main fields of current research interest: Analytical study of complex surveys, Randomized response, Small area estimation.
She published about 17 research papers in standard Indian and foreign journals and 2 more are to appear.

Gurgul Henryk is a Full Professor of economics and the head of the Department of Applications of Mathematics in Economics, at AGH University of Science and Technology in Cracow, Poland. He is also a visiting professor at several foreign universities. His major research interests focus on financial econometrics, analysis of financial markets, bankruptcy predictions, input–output models, growth models, including CEE economies in transition. He has been a referee for several distinguished international academic journals. In 2007 he won the Bank Handlowy (City Bank) Award, which is believed to be one of the most prestigious tokens of recognition in the Polish economic community.

Khare B. B. is a Professor in the Department of Statistics, Banaras Hindu University (BHU), Varanasi, India and obtained his Ph.D. degree at Banaras Hindu University, Varanasi, India in 1984. He is a life-long member of National Academy of Sciences, India and a Honorable Member of Research Board of Advisors, ABI (U.S.A.). He has established an active research group in BHU, making significant contributions in the field of sampling theory, and guided five Ph.D. candidates. Prof. Khare has published more than 85 research papers in international/national journals and conferences and presented more than 67 research papers at international/national conferences. He has also published two books/monographs and written five chapters in different recognized books. His area of specialization is Sampling Theory, Inference and Population Studies.

Kumar Manoj is an Assistant Professor at the Department of School of Basic Science and Research in Sharda University, Plot 32, 34, Knowledge Park-III, Greater Noida-201306, U. P. and also at the Department of Statistics, Banaras Hindu University, Varanasi-221005. His main research area is Bayesian inference in the presence of censoring and survival analysis, and computations.

Nath Dilip C. is a Professor & Head at the Department of Statistics in Gauhati University, Assam, India. His research interests are Biostatistics, Medical Statistics/Demographics and Actuarial Statistics. He has guided successfully 20 Ph.D. scholars so far to get their degree in Statistics on different topics. He has successfully collaborated with more than 45 national and international scholars and contributed more that hundred (130) research papers in reputed national and international journals. He was awarded the prestigious (i) Rockefeller Foundation Fellowship (1991-93) and (ii) Andrew W. Mellon Foundation Fellowship (1997-98). He is an active member of twenty scientific professional bodies. He has organized 14 Conferences, Seminars, Workshop and Training Programs as Principal Coordinator/Director/ Organizer.

Raghunathan Trivellore is a Professor of Biostatistics and a Research Professor at the Institute for Social Research at the University of Michigan. He is a Research Professor in the Joint Program in Survey Methodology. His primary research interests include Bayesian methods, small area estimation, combining information from multiple data sources, measurement error models, and statistical methods for epidemiological studies.

Sakshaug Joseph is a Senior Researcher at the Institute for Employment Research in Nuremberg, Germany and a F aculty Associate in the Program in Survey Methodology at the University of Michigan. His primary research interests include statistical disclosure control, applications of multiple imputation, small area estimation, record linkage techniques, and survey data quality.

Singh S. K. is a Professor at the Department of Statistics, Banaras Hindu University. His main research area is Classical and Bayesian Inference and Survival Analysis. He is also an expert in numerical computation as well as programming, e.g. MCMC and Gibbs sampling, etc. He has also published more than 100 articles.

Singh Umesh is a Professor Coordinator DST-CIMS & Department of Statistics, Banaras Hindu University. His research area is Classical and Bayesian Inference, Loss function and Survival Analysis. He has published more than 200 articles in reputed journals and also organizes numerous international conferences.

Sinha R. R. is an Assistant Professor at the Department of Mathematics, Dr. B. R. Ambedkar National Institute of Technology, Jalandhar, India and obtained his Ph.D. Degree in “Sampling Techniques” from the Department of Statistics, Banaras Hindu University, Varanasi, India in 2001. He has guided one Ph.D. and three M.Phil. candidates. He is a life-long member of Indian Statistical Association and International Indian Statistical Association. Dr. Sinha has published more than 18 research papers in international/national journals and conferences and presented more than 20 research papers at international/national conferences. His area of specialization is Sampling Theory, Data Analysis and Inference.

Zajac Pawel is a R esearch Assistant at the Department of Applications of Mathematics in Economics, at AGH University of Science and Technology in Cracow, Poland. His major research interests cover business demography and the application of statistical models for bankruptcy prediction. He has been an author, co-author and referee for several distinguished international academic journals.