STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 403—404

FROM THE EDITOR

The present issue of *Statistics in Transition*, 46th since its inception in 1993, is a special one as the journal, renowned for its primary focus on transitional issues — especially, on progress in different areas of statistics pertinent to social and economic transformation world-wide — has reached the stage of its own transition. Both in organizational and substantive terms, including its main mission's objectives.

With regards to this, it is with great interest and honor that I assume the responsibility for the journal as its new Editor-in-Chief. The awareness of challenges ahead of the journal's new editorial team concurs with our optimistic view that they are not insurmountable. On the contrary, they will motivate us to better organize our efforts to run the journal along the successful pathway that was designed by its founders and former editors. On the other hand, we obtain a lot of encouragement and guidance from several sources. First of all, from the journal's new Editorial Board and its Chairman, Professor Józef Oleński, the President of the Central Statistical Office (the journal's principal sponsor), and from the Polish Statistical Association– in terms of logistics and scientific assistance. And, last but not least, from a large group of the journal's readers and collaborators, its contributors and associate editors, and referees — all those who have gathered around the journal and supported its mission over the past years.

Let me inform you that the above mentioned organizational events are the consequence of Professor Jan Kordos', the journal's founder and long-term editor, decision to retire from the post of Editor-in-Chief. Fortunately, he kindly accepted our invitation to stay with us in the capacity of honorary *Founder Editor*, and to continue to share with us his extraordinary editorial experience and knowledge of statistics and statistical experts round the world.

The key task for the new editorial team and its advisors is, therefore, to uphold the journal's well established position among international periodicals and to increase its significance and scope through making it more attractive to a growing number of readers and collaborators. To this purpose, we will strive to further enhance the journal's excellence and its relevance to policy issues through its continuous commitment to intellectual rigor and responsiveness to a wide array of questions nurturing both producers and users of statistics. Including topics that range from methodological aspects of statistical process (data gathering, analysis, dissemination and use) to issues of academic or policy interest in such vital areas of social concern as population's wellbeing, health and education, employment and migration, environment and natural resources, social services and local community development, etc. Combined with interest in improving the estimates

of key economic parameters providing a better view of the economy's performance, its productivity and growth, especially, from the international perspective. Also, through expanding the primary focus of the journal to embrace all forms of transformation, including those taking place within the system of public (national) statistics itself, its modernization or statistical capacity building activities.

On the technical side, let me inform you that the journal's website is under construction and that it will include an online manuscript processing system. It will shorten not only the distance between authors and referees, but also the publishing process by making the accepted articles available online prior to their appearance in the printed version of *Statistics in Transition*.

In this 'transitional' issue, I join my predecessor, Professor Kordos, in thanking all authors, referees and associate editors for their former and current contribution, and I sincerely invite all of them to continue their collaboration with the journal, and to join us in our attempts to assure a steady increase of its significance and usefulness among the international community of statisticians, data producers and users over years to come.

Włodzimierz Okrasa
Editor-in-Chief

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 405—408

FROM THE EDITOR OF THE VOLUME

This is last issue edited by me as an Editor-in-Chief. I worked for *Statistics in Transition* and later *Statistics in Transition –new series* for fifteen years. At the end of these introductory remarks I would like to say good-bye to our Associate Editors, authors and readers.

This issue is devoted mainly to selected papers presented at the *Second Baltic-Nordic Conference on Survey Sampling.* The Conference was held in Kuusamo, Finland, in June 2007. There are also in this issue four articles in the part ***Other Articles***, three ***reports*** from conferences and ***Acknowledgement*** for referees of volume 8.

The first part of this issue contains ten papers under the general title **Survey Sampling**. *Prof. Risto Lehtonen,* Helsinki University, Finland, agreed to serve as a Guest Editor for this part. I would like to express my highest gratitude and thanks for his efficient work. The papers have been reviewed by the following referees: Kari Djerf from Finland, Dan Hedlin from Sweden, Danutė Krapavickaitė, and Aleksandras Plikusas from Lithuania, Daniel Thorburn from Sweden, Imbi Traat from Estonia, Ari Veijanen from Finland and the Guest Editor. I would like to thank all of them very much for their effort in improving quality of the papers. In this part, introduced by the Guest Editor, the following papers are presented:

1. ***Two-Phase Power Allocation in the Finnish Labour Cost Index Survey*** (by Outi Ahti-Miettinen from Finland).
2. ***Comparing Alternative Distributional Assumptions in Mixed Models Used for Small Area Estimation of Income Parameters*** (by Enrico Fabrizi, Maria Rosaria Ferrante and Silvia Pacei from Italy).
3. ***Danish Work Environment Cohort 2005: From Idea to Sampling Design*** (by Helene Feveile, Ole Olsen, Hermann Burr, Elsa Bach from Denmark).
4. ***Design-Based Inference from On-Site Samples*** (by Thomas Laitila and Annica Isaksson from Sweden)
5. ***On the Estimation of Variance of Calibrated Estimators of the Population Covariance*** (by A. Plikusas and D. Pumputis from Lithuania).
6. ***Inverse Probability of Censoring Weighting Method in Survival Analysis Based on Survey*** Data (by Marjo Pyy-Martikainen and Leif Nordberg from Finland).
7. ***Regression Composite Estimation with Application to the Finnish Labour Force Survey*** (by Riku Salonen from Finland)

8. ***Challenges in the Estimation and Quality Assessment of Services Producer Price Indices*** (by Markus Gintas Šova, John Wood and Ian Richardson from the United Kingdom)
9. ***Administrative and Statistical Registers in Business Statistics of Ukraine*** (by Olga A. Vasechko from Ukraine and Michel Grun-Rehomme from France)
10. ***Unbiased Nonlinear Estimators of a Finite Population Total: Do They Exist?*** (by Jan Wretman from Sweden)

The second part of the issue entitled **Other Articles** contains the following four papers:

11. ***Estimation of Unpaid Work in Polish Households*** (by Ilona Błaszczak-Przybycińska from Poland). The paper presents empirical estimates of unpaid household production in Poland. The following home services, not included in macroeconomic accounts are evaluated: household upkeep, food preparation, making and care of textiles, child care, adult care and volunteer work. The data came from the *Time Use Survey 2003-2004*. Two approaches were employed: the market cost method and the alternative cost method.
12. ***Some Aspects of Post-Enumeration Surveys in Poland*** (by Jan Kordos from Poland)**.** The paper is concerned with the evaluation of census quality methods in Poland. Out of several methods for evaluating the results of a census, one method is selected, i.e. a post-enumeration survey (PES). First purposes of PES are recounted and some methodological issues discussed. Next post-enumeration surveys in Poland are presented and selected methodological issues considered. In conclusion, some recommendations for the 2010 Census of Agriculture and the 2011 Census of Population and Housing in Poland are put forward.
13. ***On Basic Notions of Nonparametric Bayesian Inference*** (by Marek Męczarski from Poland). In this paper basic concepts of nonparametric approach to Bayesian statistical analysis are presented, in particular random probability measures and the Dirichlet process. Their fundamental properties and simple statistical applications are shown.
14. ***Asymptotic Tests for Receiver Operating Characteristic Curves*** (by Agnieszka Rossa from Poland). In the paper two significance tests for Receiver Operating Characteristic Curves are proposed. Both tests are based on an asymptotic χ^2 distribution of test statistics.

Part **Reports** contains three reports from Conferences:

a) ***Report from the Small Area Estimation 2007 Conference (SAE2007),*** Pisa, Italy, 3-5 September 2007 (prepared by: Tomasz Klimanek, University of Economics in Poznan and Tomasz Żądło, University of Economics in Katowice, Poland).

b) ***Report from the National Scientific Conference "Statistics Yesterday, Today and Tomorrow",*** Wroclaw, Poland, October 10 -12, 2007, devoted to celebration of 95th Anniversary of the Polish Statistical Association (prepared by Cyprian Kozyra and Joanna Dębicka, University of Economics, Wrocław, Poland).

c) ***XXVI Conference on Multivariate Statistical Analysis*** (MSA 2007) & ***VI Conference on Statistics in Social and Economic Practic***e (SwPSG 2007), Łódź, Poland, 5-7 November 2007 (prepared by Aleksandra Baszczyńska and Jacek Białek , University of Łódź, Poland).

This issue is concluded with the **Acknowledgements** of referees of Volume 8.

Please note that our journal may be found at the following website address: http://www.stat.gov.pl/gus/45_2638_ENG_HTML.htm

Papers of 24 issues of our journal as well as *Guidelines for Authors* may be found there.

Farewell to Associate Editors, Referees, Authors and Readers

As I mentioned at the beginning of this message, taking into account my age, 78, I decided to resign as an *Editor-in-Chie*f of this journal by the end of 2007. During fifteen yeas of editing it, I had the privilege and pleasure to cooperate with *Associate Editors, referees* and *authors* from different countries. I would like to express my gratitude and thanks to our *Associate Editors* and referees for their cooperation and fruitful assistance in editing the journal.

During these fifteen years we have published 46 issues in 8 volumes comprising about 8200 pages, 505 articles, 55 reports, 26 book reviews and 8 obituaries.

First version of the journal known as *Statistics in Transition* was published twice a year in 1993 - 2006. Some "*special issues*" were also published. The journal provided a forum for an exchange of ideas and experience in various fields of statistics, especially those relevant to economies undergoing transition from the centrally planed to the market-based system. Gradually we started extending our field of interest to a broader area of application of statistical methods, and started preparing a new series of the journal.

Statistics in Transition – new series (SIT-ns), began in 2007, is an electronic journal of the Polish Statistical Association published by the Central Statistical Office of Poland. The journal, SIT-ns, is now published three times a year (April, August, December). The journal is, to some extent, the continuation of the

previous journal, relating to volume numbering and logo. However, SIT-ns is now adopting a policy of extending its field of interest to a broad area of application of statistical methods.

Though the priority is given to analysis of the post-communist economies and other emerging markets, manuscripts concerned with application of statistical methods in different fields, teaching statistics, and understanding of statistics are now welcome. To achieve these aims, SIT-ns seeks to publish high-quality papers that describe development in the fields of interest.

My duties now will be confined to a function of *Founder Editor.* I hope that a new *Editorial Board* will continue the improvement of journal's quality for benefit of statisticians from different countries.

Jan Kordos
The Editor-in-Chief from 1993 to 2007

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 409—410

FROM THE GUEST EDITOR

The Second Baltic-Nordic Conference on Survey Sampling was held in June 2007 in Kuusamo, Finland. The conference belongs to a long series of scientific conferences and workshops, which was initiated in 1997 by Professor Gunnar Kulldorff of University of Umeå. The First Baltic-Nordic Conference on Survey Sampling was held in 2002 in Ammarnäs, Sweden. Workshops on research and education of survey sampling theory and methodology have been organized annually in different Baltic and Nordic countries; the main organizer has been the Baltic-Nordic Network in Survey Sampling. The network includes people from University departments, National Statistics Institutes and Statistical societies of the respective countries.

There were over 70 participants in the Second Baltic-Nordic Conference on Survey Sampling, coming from 18 different countries. The program included over 50 invited and contributed papers. The coverage of papers was wide: there were six sessions devoted to survey sampling and survey methodology, and two sessions on business surveys. Special thematic sessions were organized on calibration and model-assisted methods, small area estimation, and nonresponse. Additional special sessions addressed skewed samples and longitudinal surveys, and the future of Baltic-Nordic co-operation in survey statistics. As a consequence, the network co-operation now includes people from Ukraine in addition to the Nordic countries and the Baltic countries; the network thus can be called Baltic-Nordic-Ukrainian Network on Survey Statistics.

The educational flavor of the conference was strong: we had a privilege to follow keynote lectures given by two prominent statisticians, Professor Harvey Goldstein of University of Bristol and Professor Carl-Erik Särndal of University of Montreal. Before the conference, a Short Course on Multilevel Modelling was organized at the University of Helsinki with Prof. Harvey Goldstein as the lecturer; there were over 60 participants in the course.

We have included in this Special Issue of Statistics in Transition (new series) a total of ten papers that have their origin in the Second Baltic-Nordic Conference on Survey Sampling. The paper by Outi Ahti-Miettinen focuses on sample allocation in the Finnish Labour Cost Index Survey. Enrico Fabrizi, Maria Rosaria Ferrante and Silvia Pacei address small area estimation of income parameters using linear mixed models, for data from the ECHP Survey. The design of the Danish Work Environment Cohort Study 2005 is described in the paper by Helene Feveile, Ole Olsen, Hermann Burr and Elsa Bach. In their paper, Thomas Laitila and Annica Isaksson develop estimation procedures for the so-called on-site sampling situations. The estimation of variance of calibrated

estimators of the population covariance is discussed in the paper by Aleksandras Plikusas and Dalius Pumputis. Marjo Pyy-Martikainen and Leif Nordberg consider adjusting for panel attrition in the context of survival analysis based on survey data. In the next paper, Riku Salonen addresses regression composite estimation with an application to the Finnish Labour Force Survey. In their brief report, Markus Gintas Šova, John Wood and Ian Richardson discuss challenges in the estimation of Services Producer Price Indices, applied to the UK case. Olga Vasechko and Michel Grun-Rehomme describe in their paper the current state of administrative and statistical registers in business statistics in Ukraine. In the final paper, Jan Wretman considers nonlinear estimators of a finite population total and asks: Do they exist?

It is expected that additional papers having their origin in the Second Baltic-Nordic Conference on Survey Sampling will be published in the March 2008 issue of Statistics in Transition (new series).

Several persons (in addition to the Editor and Guest Editor) have served as reviewers of papers submitted for publication: Kari Djerf, Dan Hedlin, Danutė Krapavickaitė, Aleksandras Plikusas, Daniel Thorburn, Imbi Traat, and Ari Veijanen. I want to express my sincere thanks to all these people for their important contribution. Last but not the least, I address thanks to Jan Kordos, the Editor of Statistics in Transition (new series), for the possibility to publish conference papers in the journal.

Risto Lehtonen
The Guest Editor

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 411—422

TWO-PHASE POWER ALLOCATION IN THE FINNISH LABOUR COST INDEX SURVEY

Outi Ahti-Miettinen

ABSTRACT

In many business surveys such as the Finnish Labour Cost Index (FLCI), survey estimates are required for the whole private sector as well as for its main industry classes. Because these classes vary greatly from one to another in size and in other characteristics, a special consideration must be given to the allocation of the sample. Statistics for labour costs have been conducted in Finland every fourth year but the respective index is new and much more demanding. Our main target in designing the sample for the first year of the FLCI has been to pay special attention to reliable cross-sectional estimates. The paper presents the methodology for the sampling design of the 2007 FLCI. The procedure will not be changed dramatically in following years. The FLCI has several targets, the estimation of change in labour costs being one key issue. In the sampling design phase, thus when starting a new survey procedure, this is difficult to well take into account. Our analysis suggests exploiting power allocation. This method is used to draw an effective sample for the whole private sector as well as for its sub-populations. To achieve full advantage of stratification, allocation was done in two phases: first to main industry classes and then to size bands within these industries. Naturally, the allocation leads in some strata to take-all samples. We first provide the 'ideal' gross sample sizes for the strata, and then continue to the respective net sample sizes by anticipating the response rates, consequently.

Key words: Labour Cost Index; Industry stratification; Power allocation

1. Introduction

The Finnish labour cost index (FLCI) describes the development in the cost of labour for an hour worked in the private sector. The calculation of the labour cost index sets out from the forming of an index of wage and salary costs, to which an index of social costs is then added.

The FLCI system leads to conduct a new survey, because good quality data on short-term labour costs cannot be produced by combining existing data sources.

Our resources will allow for a sample size of around 2,000 enterprises. Estimates are needed both for the whole private sector and for its main industries. This necessitates the use of stratified sampling. Because industries vary greatly from one to another, equal-probability sampling can not be used. For unequal probability sampling, several allocation alternatives exist. We examined some of these, and our main solution is based on power allocation.

Our sampling is based on a cross-sectional approach for which purpose we were able to construct a good annual test file from 2004. Using the experience of simulation experiments based on this test file the sampling for the first year of the FLCI survey was designed and drawn. The frame here was naturally the newest possible, that is, constructed from the business register of October 2006.

The paper presents the methodology for the sampling design of the FLCI. Section 2 describes the structure of the labour cost index of the EU and gives the targets for the Finnish solution. Section 3 shows our methods and practical tools for allocating the desired gross sample to its strata. In section 4 are represented our empirical examination and the results in which the sample of 2007 Finnish Labour Cost Index is based on.

2. The labour cost index

In 2003, the European Union issued a regulation to establish a common framework for the production, transmission and evaluation of comparable labour cost indices for its Member States. The labour cost index (LCI) is defined as the Laspeyres index of labour costs per hour worked. The index will use chaining so that the base year will be changed every year, but the indices are provided quarterly. The formulae to be used in the calculation of the LCI and the structure of the weights are defined in the regulation. However, the regulation leaves open how the formulae should be interpreted in practice. The LCI shall be produced for both the private sector and for main economic activity classes.

The basic Laspeyres formula to be used to calculate the LCI for period t in year j with annual base period k is defined as:

$$LCI_{tj(k)} = \frac{\sum_i c_i^{tj} h_i^k}{\sum_i c_i^k h_i^k} = \frac{\sum_i (c_i^{tj} / c_i^k) c_i^k h_i^k}{\sum_i W_i^k} = \frac{\sum_i (c_i^{tj} / c_i^k) W_i^k}{\sum_i W_i^k}, \qquad (1)$$

where

c_i^{tj} = Labour costs per hour worked of employees in industry i of quarter t of year j

c_i^k = Labour costs per hour worked of employees in industry i of base year k

h_i^k = Hours worked by employees in industry i in base year k

$W_i^k = c_i^k h_i^k =$ Labour costs of employees in economic activity i in base period k.

The regulation defines formula for calculating the weights as $\frac{W_i^k}{\sum_i W_i^k}$. Thus these are changed every year in Finland. The last version of the formula is thus a standard weighted average for the ratio. This can be seen as a ratio estimation as described in Särndal, et al. (1992), among others. Even though LCI is in real terms a ratio of two ratios, after simplification, the ratio to be estimated is:

$$R = \frac{c_i^{tj}}{c_i^k}. \tag{2}$$

The estimator for the ratio R is $\hat{R} = \frac{\hat{c}_i^{tj}}{\hat{c}_i^k}$, a function of two random variables $\hat{c}_i^{tj}$ and $\hat{c}_i^k$. If changes in the samples are not significant over time, the theory for the usual ratio estimators can be used to calculate the variance of the labour cost index. If the sampling design is taken into account, the estimate of the LCI can be written as

$$LCI_{tj(k)} = \frac{\sum_i h_i^k c_i^{tj}}{\sum_i h_i^k c_i^k} = \frac{\sum_i h_i^k \frac{\sum_s \frac{N_{is}}{m_{is}} \sum_{l \in r_{is}} \frac{c_l^{tj}}{h_l^{tj}}}{N_i}}{\sum_i h_i^k \frac{\sum_s \frac{N_{is}}{m_{is}} \sum_{l \in r_{is}} \frac{c_l^k}{h_l^k}}{N_i}}. \tag{3}$$

Here $\frac{N_{is}}{m_{is}}$ = sampling weight following from the number of respondents m_{is} calculated by strata is in which $s =$ size band. Symbol c_l refers to the labour costs of sample enterprise l and h_l refers to hours worked in enterprise l. (Särndal, 2006)

The index formula needs to be used at industry level but the data are collected using an enterprise sample. There is a limited amount of funding available and, consequently, the sampling design requires careful planning.

This paper is focused on the sample allocation in which we try to take into account the requirements of the LCI. In our tests, we were not able to take into account the longitudinal nature of the index, but we believe that, if we are able to estimate the key cross-sectional parameters well, we will also succeed in the index

calculations. The parameters to be estimated are: (i) Totals of worked and paid hours for each quarter and each calendar year, (ii) Totals of labour costs (wages, etc.), for each quarter and calendar year, respectively, and (iii) Changes from one quarter to the next keeping the same enterprises in the sample as much as possible.

3. Stratification, take-all sizes and allocation

Because estimation of the labour cost index is also required for main industry classes, exploitation of stratified sampling is useful. In Finland, results will be calculated for 23 industries.

To avoid large difficulties in dealing with constantly changing inclusion probabilities, probability proportional to size sampling was not thought to be useful. Hence, another classification was maid for size groups, allowing varying sampling probabilities for different sizes of enterprises. As usual, all large enterprises were automatically included in the sample, but size classification allowed also for different take-all sizes for different industry classes.

Because industries vary substantially in sizes and in variables of interests, problems can arise in using standard allocation schemes, such as Neyman allocation that follows formula

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}, \tag{4}$$

where n is the overall sample size, n_h is the sample size in stratum h, N_h is the number of units in stratum h and S_h^2 is the population variance in stratum h. This traditional technique is often used when the goal is to minimise the coefficient of variation (CV) for the whole target population, but this may lead to greatly varying CVs at sub-population level and some classes can have very large CVs. On the other hand, if the CVs of all industry classes are minimised, the overall CV may be too large. Power allocation is a method for finding a compromise between these two allocations. Bankier (1988) has represented the formulae for power allocation so that the loss function is

$$F = \sum_{h=1}^{H} (x_h^q CV(\hat{y}_h))^2, \text{`} \tag{5}$$

where $CV^2(\hat{y}_h) = V(\hat{y}_h)/y_h^2$, x_h is the size of stratum h, q is a constant in the range $0 \le q \le 1$, y_h is the value of the characteristic of interest in stratum h and $\hat{y}_h$ its estimator. F is minimised subject to the constraint $\sum_{h=1}^{H} n_h = n$. The minimal value of F is achieved if

$$n_h = n\frac{S_h x_h^q / \bar{y}_h}{\sum_{h=1}^{H} S_h x_h^q / \bar{y}_h}, \qquad (6)$$

where $S_h^2 = \sum_{i=1}^{N_h}(y_{hi} - \bar{y}_h)^2/(N_h - 1)$ and $\bar{y}_h = y_h / N_h$. The value q is called the power of the allocation and then choosing its value between 0 and 1 can be viewed as the desired compromise between the Neyman allocation and the allocation of almost equal CVs. By setting $q = 1$ and letting $x_h = y_h$ leads to a Neyman allocation.

4. An empirical examination of the sampling design

4.1 Sampling frame and test data

For developing the sampling design, two test data sets were constructed, one from 2002 and another from 2004. These were derived from the business register of Statistics Finland and from the wage and salary survey. The main characteristics were the yearly averages of wages and hours paid by industry. We encountered major problems in constructing comparable files for both years and hence could not evaluate well changes in labour costs. Consequently, the testing of the allocation only concerns the cross-sectional year 2004.

The target population excludes agriculture, hunting, forestry and fishing. Activities of private households as employers and undifferentiated production activities of private households, as well as extraterritorial organisations and bodies were excluded from the frame. The sampling frame was limited to enterprises with more than 20 or 30 (dependent on industry) employees. Small enterprises' costs would be imputed to the index to avoid response burden.

In the test data we had the number of employees as the enterprise size category. Aggregates of these numbers at the industry level were used as the size of a stratum in the power allocation. Data on one year's earnings were used as the value of the characteristic of interest. This will not completely correspond to the real target variable of labour costs in the base year and even less so in quarterly statistics. However, we still believe that the tests were quite close to the real life situation we would run into with the survey. Standard deviation of earnings within an industry was used as S_h in the power allocation formula.

Table 1 shows main industry classes and number of enterprises in the frame of the Finnish Labour Cost Index survey.

Table 1. The industry classification and sizes of main industry classes in Finland.

Section	Description	*N*
C	Mining and quarrying	**16**
D	Manufacturing (D1-D7: subsets of D)	**1601**
E	Electricity, gas and water supply	**62**
F	Construction	**750**
G	Wholesale and retail trade: repair of motor vehicles, motorcycles and personal and household goods	**1349**
H	Hotels and restaurants	**251**
I	Transport, storage and communications	**380**
J	Financial intermediation	**152**
K	Real estate, renting and business activities (K1-K3: subsets of K)	**1299**
L	Public administration and defence; compulsory social security	**18**
M	Education	**214**
N	Health and social work	**402**
O	Other community, social and personal services activities	**322**

4.2 Experiments and monte calro design

To test different samples a total of 1,000 independent samples of around 2,000 enterprises were drawn from the frame population. The precision of the point estimates was evaluated by coefficient of variation (CV). The efficiency of each sample design was measured by

$$Mean(CV) = \frac{1}{1000}\sum_{k=1}^{1000} CV_k \ ,$$

where CV_k refers to coefficient of variation of the estimate in drawn sample k.

The tested allocation schemes were Neyman allocation and power allocation with power q between 0 and 1. The samples were drawn by using the SAS Surveyselect procedure and the calculation of the coefficient of variations was done with Claes Andersson's and Lennart Nordberg's SAS program CLAN, which calculates standard errors and point estimates for survey samples by taking the sampling design into account.

4.3 The effect of power q on sampling efficiency

Figures 1 and 2 are our main results from the two-phase allocation. Figure 1 includes four panels that illustrate differences by industry.

The topmost panel of Figure 1 consists of the industries that are partially uninteresting considering the allocation, since they are either very small in frame figures, or there is so large variance that all enterprises will be included in the sample. The second panel covers the largest industries. Neyman-Tschuprow allocation minimises the coefficient of variation for the whole population, so the allocation scheme gives the large industries a bigger sample when getting closer to Neyman-Tschuprow allocation and their coefficients of variation then decrease. The two bottom panels relate to medium size industries whose sample size depends mostly on the variation of the auxiliary variables. When the variation is small in the single auxiliary variable used in Neyman-Tschuprow allocation: Neyman-Tschuprow allocation will not consider them to be significant for the whole private sector estimation and gives a very small sample size to the industry and the sub-population's coefficient of variation increases notably. If the variation is greater, the coefficient of variation increases more constantly when moving towards the Neyman-Tschuprow allocation.

In Figure 2 we have taken some examples of industries from the three interesting panels of Figure 1 and included the coefficient of variation of the estimate for the whole private sector (PS) into one picture.

The effect of choosing the power in the allocation can be seen from Figure 2. By reducing the precision requirement for the whole population a little, the estimates of all industry classes can have CVs that are within acceptable limits. Because the industry classes vary so much from one to another, allocation with power $q = 0$ does not produce equal CVs for all industry classes and some estimates of industry classes end up having rather high CVs.

Figure 2 also shows quite well that there can be more convenient allocation to the sampling design when choosing the power allocation and its power q somewhere in between 0 and 1.

A good quality estimate for the whole private sector is one of the main interests in the study. But because estimates for all industry classes are needed, an other way to view the efficiency of overall sample is to calculate the weighted means of the CVs by industry classes. The weighting was done by the number of enterprises in the industry class. In Table 2 we see that this leads to same interpretation that we had from the Figure 2.

Table 2. Weighted means of the CVs by industry classes in all tested allocations.

q	**0.0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**	**Ney**
CV	2.40	2.31	2.24	2.20	2.16	2.16	2.16	2.19	2.22	2.29	2.31	2.42

4.4 Allocation in two phases

There are few reasons to believe that the single classification to industries is not enough in collecting the data for the LCI. It is hoped that a large part of the sum of wages paid to employees will be included in the sample and from earlier surveys it can be assumed that the unit cost of labour varies by size of enterprise. Moreover, because the distribution by size is skew among enterprises and large enterprises account for a major share of the total of wages, the representativeness of enterprises of all sizes categories must be ensured. Division of enterprises to size strata helps to achieve this goal.

It is not purposeful to publish any figures by size classes. Therefore, we questioned the need to allocate the sample into a two-dimensional sampling frame in one phase. One dimension in the table being industry classification and the other being size classification. If this was done, the allocation would also optimise the sample by size class. An effective sample was required primarily for the whole private sector and main industry classes. Therefore, we tested another possibility where the allocation was done in two phases: first to industry classes and then to size groups within industries.

To test a number of size classes and the two-phase allocation, we made the similar simulation experiment as in 4.3. It was done both as allocation in one phase and allocation in two phases to the two-dimensional frame. The efficiency of each design was measured by calculating the mean of the CVs over simulations of the whole private sector's estimate.

Figure 3 shows how the efficiency of the sample changes when the number of size classes varies. In addition, the figure shows the difference in efficiency with one-phase and two-phase allocation. It can be seen that the two-phase allocation is more efficient if there are more than two size classes.

Figure 1. The effect of power q in allocation in all industry classes

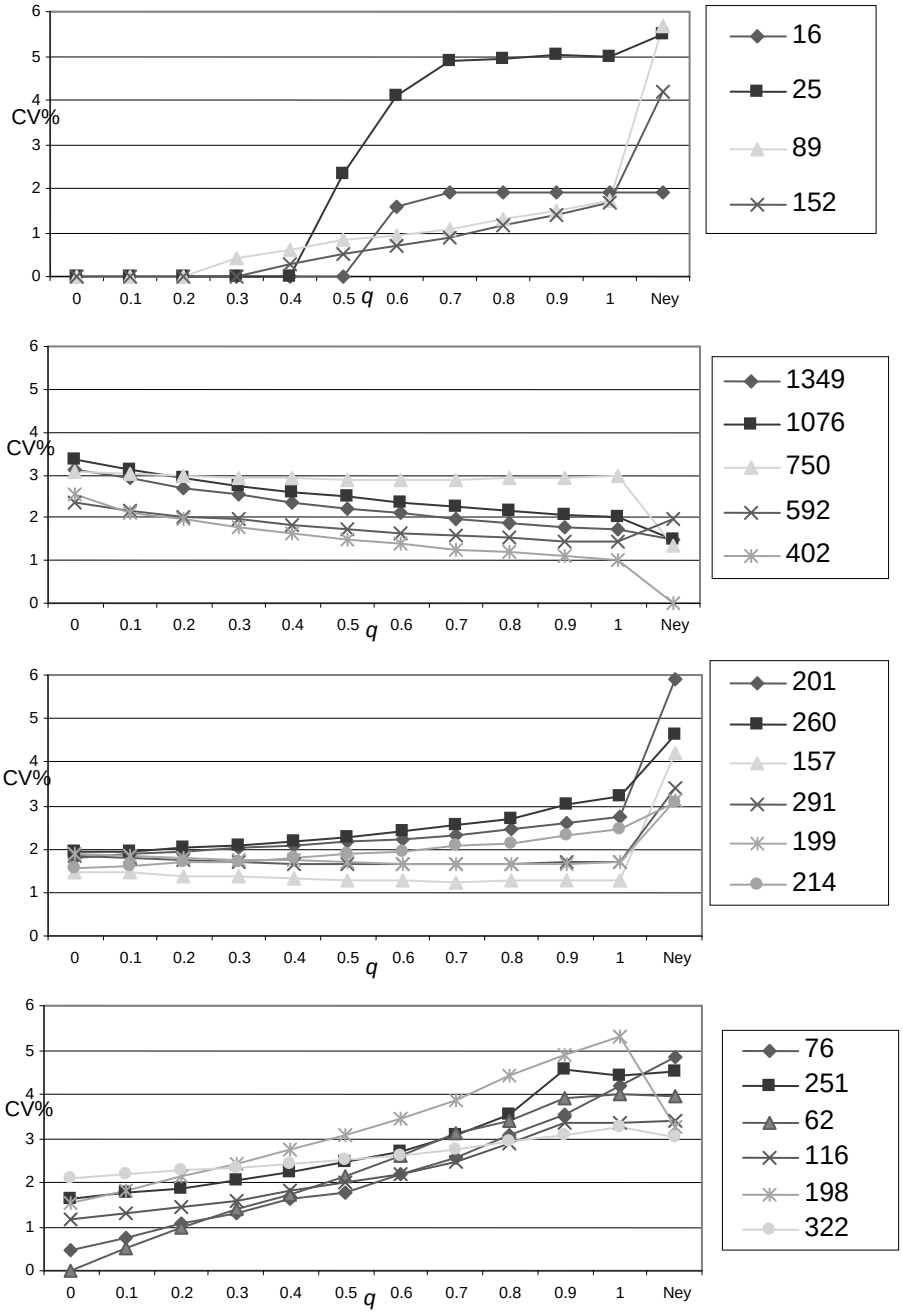


Figure 2. An example of the effect of power q in allocation in some industry classes and the whole private sector (PS).

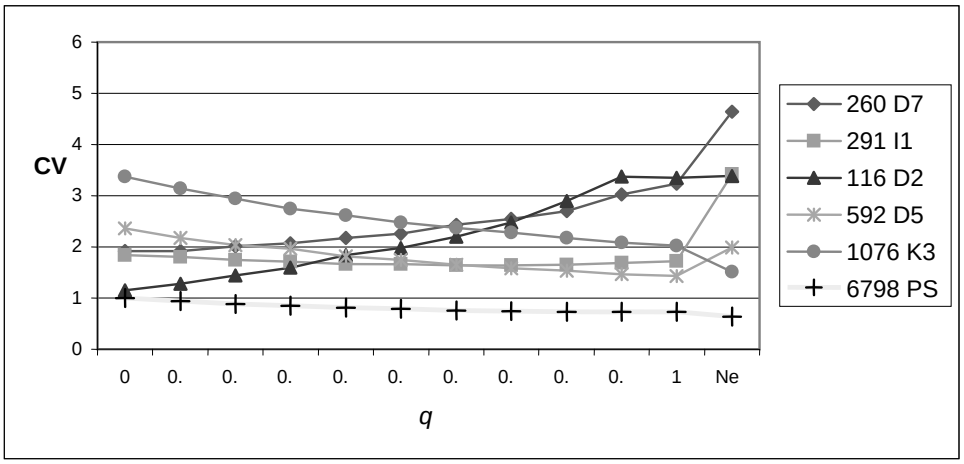
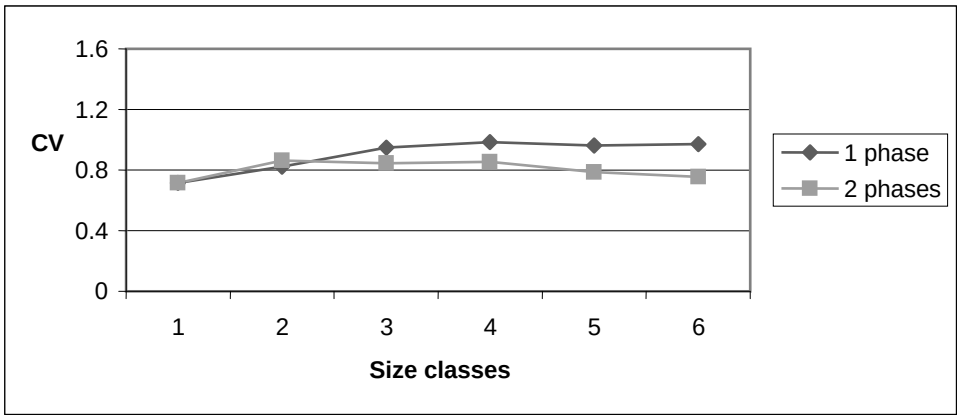


Figure 3. Difference in coefficient of variation between allocation in one or two phases.



4.5 Effect of non-response

Even if the LCI survey were to be made mandatory, we know that especially small enterprises would produce high non-response rates. This increases the CVs in the forthcoming estimates, compared to those presented in the empirical part of this paper. Basing on experiences from Finnish structural earnings statistics, we were able to anticipate the response rates by strata, approximately, and we hope that these will be rather correct after the fieldwork. Using this information some of the loss of efficiency can be prevented.

Because the sampling size of the survey was fixed to around 2,000 enterprises, we had to adjust the calculated gross sample allocation with response rates.

5. Discussion

The paper presents the sampling design for the first year of the FLCI. The design is based on stratified simple random sampling given that the maximum gross sample size is approximately fixed. The requirements for index estimates are demanding since results are needed for 23 industries, not only for the whole private sector. Hence a new sample allocation scheme has been developed, taking advantage of earlier results, especially from Bankier (1988). Naturally, other strategies, especially Neyman allocation, were examined as well as several alternatives for power allocation.

The allocation was made in two phases: first to industries and then to size bands within industries. This is because there was no need to optimise the sample by size band. The allocation results are not uniform over industries and hence it is not automatically clear which allocation should be used in practice.

The sampling frame of the FLCI has five size bands. As Figure 3 shows, even though six size bands would produce a slightly more effective sample, more than five size bands would create problem of a large number of enterprises moving from one stratum to another in our panel design. Five bands were also thought to give enough variety to the sampling probabilities.

Power allocation makes sure that all main industry classes would have a sample that would produce an acceptable, precise estimate of labour costs. According to our tests, in our sampling frame power allocation, with power q somewhere in between 0.4 and 0.6, would do this. In the sampling of the FLCI it was decided that allocation with q=0.5, also known as square-root allocation, would be used. Everything will be checked again after the first year fieldwork period at the end of 2007, but the sampling design cannot be changed much.

Due to longitudinal aspect of the survey and the forthcoming reforming of NACE classification of economic activities, size bands were kept the same between industries. Therefore optimal stratification of size bands within industry could not be done and classification from previous surveys had to be settled for. After the revision of NACE is done, the future research will focus on benefits of optimal size boundaries. Also the alternative approach of indirect allocation will be studied, where iterative procedure is used to find all industries the lowest CVs that the total sample size provides for.

Acknowledgements

I would like to thank Professor Seppo Laaksonen for introducing and guiding me to the topics of this research and supervising of my master's thesis, in which this paper is based on. He has had a significant influence in both the scientific research and writing of this paper.

REFERENCES

BANKIER, M. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. *Journal of the American Statistical Association, 42: 174—177.*

COCHRAN, W. (1977). *Sampling Techniques. Third Edition.* Wiley, New York.

THE EUROPEAN PARLIAMENT AND COUNCIL (2003). *Regulation (EC) No 450/2003 of the European Parliament and of the Council.*

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

SÄRNDAL, C.E. (2006). *Notes on LCI.* Unpublished memorandum. Statistics Finland.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 423—439

COMPARING ALTERNATIVE DISTRIBUTIONAL ASSUMPTIONS IN MIXED MODELS USED FOR SMALL AREA ESTIMATION OF INCOME PARAMETERS

Enrico Fabrizi, Maria Rosaria Ferrante, Silvia Pacei

ABSTRACT

Linear Mixed Models used in small area estimation usually rely on normality for the estimation of the variance components and the Mean Square Error of predictions. Nevertheless, normality is often inadequate when the target variable is income. For this reason, in this paper we consider Linear Mixed Models for the log-transformed income (which require back-transformation for prediction of means and totals on the variable's original scale) and a Generalized Linear Mixed Model based on the Gamma distribution. Various prediction methods are compared by means of a simulation study based on the ECHP data. Standard predictors obtained from Linear Mixed Model for the untrasformed income are shown to be preferable to the considered alternatives, confirming their robustness with respect to the failure of the normality assumption.

Key words: European Community Household Panel; Average Equivalized Income; Lognormal Linear Model; Prediction; Gamma Distribution.

1. Introduction

In the European Union, the demand for estimates about the distribution of income at the sub-national level, a fundamental tool for the implementation of social cohesion policies, has grown rapidly in recent years (Stewart, 2003). For the period from 1994 to 2001, income distribution parameters and poverty indicators may be estimated consistently across most of the member states using information collected by the European Community Households Panel (ECHP), a sample survey on households' income and social conditions, coordinated by Eurostat (Betti and Verma, 2002; Eurostat, 2002). This panel survey was designed to provide reliable estimates of main parameters of interest for large areas within

countries called NUTS1 (NUTS stands for the "Nomenclature of Territorial Units for Statistics"; Eurostat, 2003).

The ECHP survey was substituted in 2004 by a new rotating panel survey called EU-SILC (European Union — Statistics on Income and Living Conditions), based on new measurement methodologies and a larger sample (Eurostat, 2005). The two surveys are very similar under many aspects and ECHP data pertaining to Italy is used for the purposes of this paper.

We are interested in estimating the mean of equivalized household income for sub-national regions defining a partition of the country, for which direct estimators, that is, those applying standard weighted estimators to the region-specific part of the sample, lead to estimates with too large a variance. The solution to this problem involves the application of a 'Small Area' estimator, that is, an estimator using relevant auxiliary information to improve the precision of direct estimates (see Rao, 2003, for a general review). The auxiliary information may be exploited by specifying a (sometimes implicit) model that relates all the areas being studied.

In particular, in Fabrizi *et al.* (2007), we discuss several models within the class of 'unit level' Linear Mixed Models, where a linear relationship is assumed between the target variable and a set of auxiliary variables whose total is accurately known from the Census or some other sources, and random effects are introduced to model the correlation of residuals. In this approach, the models are linear for the equivalized household income considered on its original scale, and normality is assumed for the random effects and the residuals. In Fabrizi *et al.* (2007), we recognize that the normality assumption may not hold exactly for the considered data, but we find it to have a moderate impact on small area point predictors of equivalized mean income; moreover, provided that a robust strategy for the estimation of MSE is followed (for instance, the jackknife estimator of Jiang *et al.*, 2002), an estimate of MSE associated to these predictors with good properties may also be obtained.

Other authors (see for instance Elbers *et al.*, 2003) prefer to apply Linear Mixed Models to the log-transformation of income. This in principle may improve the fit of the models, but it has two related drawbacks: *i*) in order to predict area means or totals on the original scale of the study variable you need to back-transform individual predicted values, but the resulting prediction values will be biased (although several methods have been proposed to keep this bias low); *ii*) the prediction of individual values requires that the values of auxiliary variables are known for each member of the population outside the sample, whereas if the model is linear on the natural scale of the study variable, only the area means/totals of auxiliary variables are needed to predict area means/totals of the study variable.

In this paper, we do not consider this latter problem, we focus instead on the prediction of the mean of the equivalized income for a subset of the population by considering several alternative options. One is to consider a linear mixed model

on the natural scale of the equivalized income. To keep things simple, we will discuss the well-known nested error regression model introduced by Battese, Harter and Fuller (1988). Another is to also consider linear mixed models on the log-transformation of the equivalized income in association with different bias correction methods: naïve, smearing (Duan, 1983) and a ratio-adjusted-for-sample-total (RAST) method discussed in Chambers and Dorfman (2003). Finally, we also consider a Generalized Linear Mixed model, in which a more suitable distribution (positive and non symmetric) for the equivalized income is assumed conditionally on the auxiliary variables, the Gamma distribution.

The comparison of these options is based on a Monte Carlo exercise. To this purpose, the last wave (2001) of the ECHP survey is treated as a pseudo-population from which we bootstrap samples using the survey weights as the size variable. This solution may not be as good as that of using data from a real Census population, but it is hopefully more realistic than generating population values of household income from a parametric model.

The paper is organized as follows. Section 2 discusses the use of Mixed Linear Models for the small area estimation and describes the nested error regression model. Section 3 briefly reviews the ECHP survey and describes how we use this survey data to conduct the Monte Carlo simulation study. Section 4 presents distributional assumptions and predictors suggested as alternatives. Performances of the estimators derived from the proposed models are compared in section 5.

2. Linear mixed models in small area estimation

When the target parameter is an average or a total, Linear Mixed Models (LMM) are largely used. A brief description of LMM and the estimators they lead to is given below. For a more complete review of the application of this class of models in the context of small area estimation, see Rao (2003, ch. 5). A general linear mixed model can be described as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1 v_1 + ... + \mathbf{Z}_s v_s + \mathbf{e}\,, \tag{1}$$

where $\mathbf{y} = \left\{ y_{dj} \right\}$ is the *n*-vector of sample observations, *j* denotes the unit and *d* the small area ($j = 1,...,n_d$; $d = 1,...,m$), β a $p \times 1$ vector of regression coefficients, v_i is a $q_i \times 1$ vector of random effects ($i = 1,...,s$), $\mathbf{e} = \left\{ e_{dj} \right\}$ a vector of errors; $\mathbf{X}$ is assumed of rank *p*, $\mathbf{Z}_i = \left\{ \mathbf{z}_{idj}^T \right\}$ is a $n \times q_i$ matrix of the incidence of the *i*-th random effects. We assume that $E\left(v_i\right) = 0$, $V\left(v_i\right) = \mathbf{G}_i$, $E\left(\mathbf{e}\right) = 0$, $V\left(\mathbf{e}\right) = \mathbf{R}$ (all expectations are wrt. model (1)) and that $v_1,...,v_s,\mathbf{e}$ are mutually independent.

As a consequence, the variance-covariance matrix of $\mathbf{y}$ is given by:

$$\mathbf{V} = V(\mathbf{y}) = \sum_{i=1}^{s} \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}_i^T + \mathbf{R} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R},$$

where $\mathbf{Z} = [\mathbf{Z}_1 | ... | \mathbf{Z}_s]$. It is usually assumed that matrixes $\mathbf{G}, \mathbf{R}$ depend on a *k*-vector of variance components ψ, and so we can write $\mathbf{V}(\psi) = \mathbf{Z}\mathbf{G}(\psi)\mathbf{Z}^T + \mathbf{R}(\psi)$.

Note that at the level of individual observations, the model (1) can be rewritten as $y_{dj} = \mathbf{x}_{dj}^T \beta + \mathbf{z}_{1dj}^T v_1 + ... + \mathbf{z}_{sdj}^T v_s + e_{dj}$.

In small area estimation, the aim is to predict scalar linear combinations of fixed and random effects of the type $\eta = \mathbf{m}^T \beta + \mathbf{k}^T v$ where $\mathbf{m}$ and $\mathbf{k}$ are $p \times 1$ and $q \times 1$ vectors respectively, with $q = \sum_i q_i$. The best linear unbiased predictor (BLUP) of η can be obtained by estimating fixed effects and "realized values" of random specific area effects by GLS method:

$$\tilde{\eta}^{BLUP}(\psi) = \mathbf{m}^T \tilde{\beta}(\psi) + \mathbf{k}^T \tilde{v}(\psi). \qquad (2)$$

When the variance components in ψ are unknown, they may be estimated from the data and substituted into (2), thus obtaining "empirical BLUP" $\tilde{\eta}^{EBLUP}(\hat{\psi}) = \mathbf{m}^T \hat{\beta}(\hat{\psi}) + \mathbf{k}^T \hat{v}(\hat{\psi})$ (see Rao, 2003, ch. 6, and Jiang and Lahiri, 2006, for details). As far as the estimation of ψ is concerned, a number of methods have been proposed in the literature, such as Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) which assume the normality of random terms, and the MINQUE proposed by Rao (1971) which is non-parametric. In the present work we have opted for the REML method, thus assuming normality.

One simple example within the class of Linear Mixed Models is given by the standard one-fold nested error linear regression model of Battese, Harter and Fuller (1988), which has been widely applied in the small area literature:

$$y_{dj} = \mathbf{x}_{dj}^T \beta + \alpha_d + e_{dj} \qquad (3)$$

where y_{dj} is the *Y* value observed on unit *j* of area *d*, $\mathbf{x}_{dj}$ is the auxiliary vector for unit *j*, β is the fixed effects vector (common to all areas), α_d is the specific area *d* effect and e_{dj} is the residual term for unit *j*.

All random terms are assumed mutually independent and normally distributed with zero mean and constant variance:

$$\alpha_d \overset{ind}{\sim} N\left(0,\sigma_\alpha^2\right), \quad e_{dj} \overset{ind}{\sim} N\left(0,\sigma_e^2\right). \tag{4}$$

Therefore this random effects structure corresponds to the assumption of a constant covariance between units that belong to the same area. Note that it is a particular case of (1) obtained when $s=1$, $q_1=m$, $\mathbf{G}_1=\sigma_\alpha^2\mathbf{I}_m$ and $\mathbf{R}=\sigma_e^2\mathbf{I}_n$.

Model (3) — (4) will be considered as the benchmark in the comparison between alternative distributional assumptions on residuals. The EBLUP estimator of the small area mean ($\eta_d=\bar{\mathbf{x}}_d^T\beta+\alpha_d$) will be given by $\tilde{\eta}_d=\bar{\mathbf{x}}_d^T\tilde{\beta}+\tilde{\alpha}_d$, where fixed effects and "realized value" of random specific area effects are estimated as described above.

3. The simulation study based on the european community household panel data

We carried out a Monte Carlo simulation study using the last wave (2001) of the ECHP data available for Italy as our 'synthetic population'. The use of the sample as pseudo-population is necessitated by the fact that the information on household income is not collected by the Census. Nevertheless, this solution is hopefully more realistic than generating population values of household income from a parametric model. Similar simulation studies based on re-sampling can be found in Falorsi *et al*. (1999); Lehtonen *et al*. (2003); Singh *et al*. (1994).

We took the household to be the reference unit in our study. The households in the data set were selected from different strata (NUTS2 regions) and were given different weights that account for unequal selection probability, adjustments for non-response in the initial recruitment and subsequent attrition.

In our Monte Carlo experiment, samples were drawn with replacement from the ECHP data set using stratified probability proportional to size sampling, the size variable being given by survey weights. Strata were given by NUTS2 regions which also correspond our domain of interest, the 21 Italian Administrative Regions, and were therefore treated as 'fixed domains'. Moreover we note that replicated samples were drawn keeping the pseudo-population fixed, so that the simulation was aimed at evaluating the design-based properties of the estimators.

The size of replicated samples was fixed to $n=1{,}000$ (roughly 15% of the size of respondent households in the 2001 wave). The region-specific sample sizes we obtained ranged from 14 to 112, being on average equal to 48. The total number of simulated samples was set to 1,000. Monte Carlo errors associated with this number of replicates were small enough to ensure significance of all comparisons we discuss in section 5.

Our target variable was given by the total net household income equivalized with respect to household size and composition. Total net household income is obtained as the sum of net incomes of all members of the household. Equivalent

net income is calculated by dividing total net household income by equivalent household size according to the OECD scale used by Eurostat (which gives a weight of 1.0 to the first adult, 0.5 to the other persons aged 14 or over who are living in the household and 0.3 to children under the age of 14).

Regarding the characteristics of the obtained pseudo-population, its overall mean is 22,547 Euros and the coefficient of variation is 0.59. The distribution is positively skewed even though skewness is not extreme (skewness coefficient $\gamma_1 \cong 2.5$). The difference between mean and median is 9% of the mean. Looking at the different administrative regions, the small area averages of the target variable show very different values, thus reflecting the well-known regional disparities which characterised the country. For example, the highest regional average is about 70% higher than the lowest one. The variance varies among regions, increasing with the regional average, so that the coefficient of variation varies among small areas from a minimum of 0.28 to a maximum of 0.84. Also the skewness (γ_1 ranging from 0.1 to 4.6) shows that the distribution of our target variable is quite a bit different in different areas.

Of the many covariates available from the ECHP questionnaire, we considered only those for which area population means were available from the 2001 Italian Census results, because those means are necessary to calculate the EBLUP estimator. Thus the chosen covariates are as follows: the percentage of employed; the percentage of unemployed; the percentage of people with a high/medium/low level of education in the household; household typology (presence of children, presence of aged people, etc.); the number of rooms per-capita and the tenure status of the accommodation (rented, owned etc.).

The adjusted R^2 of the OLS regression is close to 0.35 in almost all repeated samples. This rather low figure is the result of the nature of the phenomenon under study (household income is not easy to predict) and the constraint represented by the need to include only those covariates for which the population total can be obtained from the Census.

4. Alternative predictors

We consider two classes of alternatives to the empirical best predictor (EBLUP) associated with the nested error regression model described in section 2: the first includes predictors based on the fitting of a nested error regression model onto the logarithm of the total net equivalized household income; whereas the second assumes that, conditionally on the covariates, the total net equivalized household income is Gamma distributed.

The logarithmic transformation is often used in models for income because the logarithm of values generated from a positively asymmetric distribution are generally more “normal” than the untransformed values. Also the Gamma model has often been considered for the study of income distribution. Reasons for its use

have been both theoretical (Mukerji, 1967) and practical, due to the better fit provided with respect to empirical distribution (Eltetö e Frigyes, 1968; Van Praag *et al*., 1983).

The two strategies are described in subsections 4.1 and 4.2.

4.1. Predictors based on modeling the logarithm of income

In this model the dependent variable is given by the (natural) logarithmic transformation of the household equivalised income:

$$z_{dj} = \log y_{dj} \qquad d = 1,...,m \quad j = 1,...,n_d$$

$$z_{dj} = \mathbf{x}_{dj}^T \beta + \alpha_d + e_{dj} \tag{5}$$

$$\alpha_d \sim N\left(0, \sigma_\alpha^2\right) \quad e_{dj} \sim N\left(0, \sigma_e^2\right) \quad \alpha_d \perp e_{dj}\,. \tag{6}$$

The usual hypotheses of independence, homoschedasticity and normality of residuals hold so variance components are estimated using the REML technique. The 'naïve predictors' discussed in subsection 4.1.1 rely on this normality assumption. However, we have proof that this assumption does not hold exactly in the case of our data. To overcome this problem, non parametric solution that do not rely on this assumption have been proposed. We discuss two different options within this class in subsections 4.1.2 and 4.1.3.

4.1.1. Naïve predictor

The quantity to be estimated is given, for area *d*, by $\overline{y}_{U,d} = N_d^{-1} \sum_{j=1}^{N_d} y_{dj} = N_d^{-1} \sum_{j=1}^{N_d} \exp\left(z_{dj}\right)$. A simple back transformation of the empirical best linear unbiased predictor (EBLUP) $\hat{\overline{z}}_{U,d} = \overline{\mathbf{x}}_d^T \hat{\beta} + \hat{\alpha}_d$, i.e. $\exp\left(\hat{\overline{z}}_{U,d}\right)$ cannot be used since it would be severely biased. A slightly better predictor may be obtained as:

$$\hat{\overline{y}}'_{U,d} = \frac{\sum_{j \in s} y_{dj} + \sum_{j \notin s} \exp\left(\hat{z}_{dj}\right)}{N_d} \tag{7}$$

with $\hat{z}_{dj} = \mathbf{x}_{dj}^T \hat{\beta} + \hat{\alpha}_d$. However, also this predictor is biased low because, in general:

$$E\left\{\exp\left(\mathbf{x}_{dj}^T \beta + \alpha_d + e_{dj}\right)\right\} \neq E\left\{\exp\left(\mathbf{x}_{dj}^T \beta + \alpha_d\right)\right\} \tag{8}$$

even when $E\left(e_{dj}\right)=0$.

In the literature, different strategies have been suggested to overcome this problem. Some of them keep the normality distribution assumption for the transformed variable, others escape this restriction. In the first group of methods there is, for instance, the naïve lognormal predictor (Chambers and Dorfman, 2003), which uses a first order bias correction. In the case of model (5) – (6) it becomes:

$$\hat{\bar{y}}'_{U,d}=\frac{\sum_{j\in s}y_{dj}+\sum_{j\notin s}\exp\left(\hat{z}_{dj}+\frac{\hat{V}\left(z_{dj}\right)}{2}\right)}{N_d} \tag{9}$$

where the estimated variance of z_{dj} is $\hat{V}\left(z_{dj}\right)=\hat{\sigma}_e^2+\hat{\sigma}_\alpha^2$. However this predictor is still biased ($O\left(n^{-2}\right)$ bias). It is possible to demonstrate that the correction for the negative bias leads to the overestimation of *Y* values (Chambers and Dorfman; 2003).

An alternative predictor, which is strongly based on the assumption of log-normality and characterized by the same order of bias as the previous one ($O\left(n^{-2}\right)$), has been discussed by Kalberg (2000).

4.1.2. The rast predictor

The aim of the method is to yield a predictor calibrated on the sample average (or total) of *Y*, that is, such that $\sum_{j\in s}y_{dj}=\sum_{j\in s}\hat{y}_{dj}$. The lognormal predictors discussed in the previous paragraph do not possess this property. The 'naïve' predictor (7) is modified so that it will possess this property (Chambers and Dorfam, 2003). It is necessary to find new formulas for GLS estimators of β and α_d so that:

$$\sum_{j\in s}y_{dj}=\sum_{j\in s}\exp\left(z^*_{dj}\right)=\sum_{j\in s}\exp\left(\mathbf{x}_{dj}^T\beta^*+\alpha_d^*\right).$$

It is easy to show that the equality holds by simply adding to the intercept a correction given by $\gamma\left(\hat{\beta},\hat{\alpha}_d\right)=\log\sum y_{dj}-\log\sum\exp\left(\mathbf{x}_{dj}^T\hat{\beta}+\hat{\alpha}_d\right)$. Therefore, assuming *K* covariates, $\beta^*=\left(\hat{\beta}_0+\gamma\left(\hat{\beta}\right),\hat{\beta}_1,...,\hat{\beta}_K\right)^T$ and $\alpha_d^*=\hat{\alpha}_d$. The resulting predictor of the population mean is:

$$\hat{\bar{y}}_{U,d}^{RAST} = \frac{\sum_{j\in s} y_{dj} + \sum_{j\notin s} \exp\left(z_{dj}^{*}\right)}{N_d} = \frac{1}{N_d}\left(\sum_{j\in s} y_{dj} + \frac{\sum_{j\notin s} \exp\left(\sum_{k=1}^{K} \hat{\beta}_k x_{kdj}\right) \sum_{j\in s} y_{dj}}{\sum_{j\in s} \exp\left(\sum_{k=1}^{K} \hat{\beta}_k x_{kdj}\right)}\right) \qquad (10)$$

It is possible to note that in (10) both intercept and specific effects disappear, but their effect is taken into consideration in the estimation of the other coefficients. In effect, the ratio between the sample sum of *Y* values and the sample sum of their predictions is used to correct individual non sampled predictions.

4.1.3. The smearing predictor

With the aim of estimating the untransformed scale expectation $E\left(Y\middle|\mathbf{X}\right) = \int \exp\left(\mathbf{X}'\beta + \varepsilon\right) dF\left(\varepsilon\right)$, without knowing the error distribution function *F* or a reliable parametric form for it, Duan (1983) suggests substituting *F* by its empirical estimate $\hat{F}_n$ so as to obtain what she calls *smearing estimate*:

$$\hat{E}\left(Y\middle|\mathbf{X}\right) = \int \exp\left(\mathbf{X}^T\beta + e\right) d\hat{F}_n\left(e\right) = \frac{1}{n}\cdot\sum_{j\in s} \exp\left(\mathbf{X}^T\hat{\beta} + \hat{e}_j\right)$$

where the $\hat{e}_j$ are the sample residuals from the ordinary least squares fit of $\log y_j$ onto $\mathbf{x}_j$. Following this idea, for an arbitrary estimator $\hat{\beta}$ of β, the smearing predictor of the population mean may be written as:

$$\hat{\bar{y}}^{SMEARING} = \frac{\sum_{j\in s} y_j + \sum_{j\notin s} y_j^{*}}{N} = \frac{\sum_{j\in s} y_j + \sum_{j\notin s} n^{-1} \sum_{h\in s} \exp\left(\mathbf{x}_j^T\hat{\beta} + \hat{e}_h\right)}{N} =$$

$$= \frac{\sum_{j\in s} y_j + \sum_{j\notin s} \exp\left(\mathbf{x}_j^T\hat{\beta}\right)\cdot n^{-1} \sum_{h\in s} \exp\left(\hat{e}_h\right)}{N}$$

where the observations for the non sampled units are predicted by correcting the "naïve" back transformation by a factor given by the average of the sample residuals $sc\left(\mathbf{e}\right) = n^{-1}\sum_{j\in s} \exp\left(\hat{e}_j\right)$.

In the case of model (5) – (6), the smearing predictor for area *d* will be given by:

$$\hat{\bar{y}}_{U,d}^{SMEARING} = \frac{\sum_{j\in s} y_{dj} + n_d^{-1}\sum_{h\in s}\exp(\hat{e}_{dh})\cdot\sum_{j\notin s}\exp\left(\mathbf{x}_{dj}^T\hat{\beta}+\hat{\alpha}_d\right)}{N_d} = \frac{\sum_{j\in s} y_{dj} + sc(\mathbf{e}_d)\cdot\sum_{j\notin s}\exp\left(\mathbf{x}_{dj}^T\hat{\beta}+\hat{\alpha}_d\right)}{N_d} \quad (11)$$

where fixed and random effects are estimated by GLS as usual and the correction factor $sc(\mathbf{e}_d)$ is calculated as the average of the residuals within area *d*.

It is possible to show that the smearing predictor in (11) may be also obtained as:

$$\hat{\bar{y}}_{U,d}^{SMEARING} = \frac{1}{N_d}\left\{\sum_{j\in s} y_{dj} + \sum_{j\notin s}\exp\left(\sum_{k=1}^{K}\hat{\beta}_k x_{kdj}\right)\cdot\left(n_d^{-1}\sum_{h\in s}\frac{y_{dh}}{\exp\left(\sum_{k=1}^{K}\hat{\beta}_k x_{kdh}\right)}\right)\right\}. \quad (12)$$

Also in this expression, as happened for RAST predictor in (10), the intercept and the specific effects of the model disappear, but their effect is taken into consideration in the estimation of the other coefficients.

We note that the smearing method is based on the estimation of the distribution function of the study variable separately for each of the areas whose mean is being predicted. In general, these estimators will have good asymptotic properties but they may perform rather poorly when applied to small samples.

4.2. The predictor based on the gamma linear mixed model for income

Consider the following model:

$$y_{dj} \mid \alpha_d, \beta \sim Gamma\left(\nu, \nu/\mu_{dj}\right) \quad (13)$$

with $\mu_{dj} = \mathbf{x}_{dj}^T\beta + \alpha_d$. Since $E\left(y_{dj} \mid \alpha_d, \beta\right) = \mu_{dj}$ and $V\left(y_{dj} \mid \alpha_d, \beta\right) = \mu_{dj}^2/\nu$ we then have $CV\left(y_{dj} \mid \alpha_d, \beta\right) = \nu^{-1/2}$.

As a consequence, in our Gamma model we assume a constant coefficient of variation. This assumption may be useful in situations where the variance of the observations increases with the mean (McCullagh and Nelder, 1989, p. 285). This hypothesis appears very sensible in our case. It does not allow for a direct and immediate comparison either with the benchmark model (equations 3 – 4) or with the model for the logarithm of income whose residuals are assumed homoschedastic, but the aim of this work is to compare predictors rather than their related models.

The predictor of $\overline{y}_{U,d}$ associated with this model may be easily obtained as:

$$\hat{\overline{y}}_d = \frac{\sum_{j\in s} y_j + \sum_{j\notin s}\left(\overline{\mathbf{x}}_d^T\hat{\beta} + \hat{\alpha}_d\right)}{N_d}. \tag{14}$$

The estimates $\hat{\beta}$ and $\hat{\alpha}_d$ are obtained using the Maximum Likelihood method as implemented in the GLIMMIX procedure of SAS (SAS Institute, 2006). We note that this estimator differs from the EBLUP derived under the normality distribution assumption only in the variance and covariance matrix used to estimate β and α_d.

5. Results from the simulation study

In summary, the predictors for the regional averages of the equivalized per-capita income that we are going to compare in the simulation exercise are the following: *i*) the EBLUP obtained from the normal Linear Mixed Model of (3) and (4) (LMM); *ii*) the naïve and naïve lognormal predictors (respectively NAÏVE and NALOG); *iii*) the RAST and the SMEARING predictors obtained from the normal Linear Mixed Model for the logarithm of *Y* (say RAST and SMEAR); *iv*) the predictor obtained from the Linear Mixed Model for income with Gamma distributed observations (G).

The performances of estimators will be evaluated by averaging not only over the Monte Carlo replicates but also over the small areas, following an approach common in the literature (see Rao, 2003, section 7.2.6). In particular we chose to show three measures: the average absolute relative bias (*AARB*), the average relative bias (*AARB'*) and the average relative mean squared error (*ARMSE*):

$$\begin{aligned}
AARB &= m^{-1}\sum_{d=1}^{m}\left|R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dr}}{\overline{y}_{U,d}}-1\right)\right| \\
AARB' &= m^{-1}\sum_{d=1}^{m}\left\{R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dr}}{\overline{y}_{U,d}}-1\right)\right\} \\
ARMSE &= m^{-1}\sum_{d=1}^{m}\left\{R^{-1}\sum_{r=1}^{R}\left(\frac{\tilde{y}_{dr}}{\overline{y}_{U,d}}-1\right)^2\right\}
\end{aligned} \tag{15}$$

where $\tilde{y}_{dr}$ is the prediction of the pseudo-population small area mean $\overline{y}_{U,d}$ obtained from the r^{th} simulated sample. Both *AARB* and *AARB'* summarize the bias of the predictors: *AARB* averages the size of the relative bias over the areas, while *AARB'*, which considers the sign of this bias, is helpful to understand

whether there is a systematic tendency of the predictors to under or overestimate the actual $\bar{y}_{U,d}$. The third indicator, *ARMSE*, is a measure of accuracy of the predictors.

Results related to these indicators are reported in Table 1. Besides the indicators of (15), the average relative efficiency

$$AEFF_{dir} = \frac{ARMSE_{*}}{ARMSE_{dir}} \tag{16}$$

is also reported. Note that $ARMSE_{dir}$ pertains to the direct estimator (i.e. the Horwitz-Thompson estimator of $\bar{y}_{U,d}$ based on the inverse inclusion probabilities) in order to make the evaluation of the advantages associated with the various model-based predictors more readily comparable. In (16), * stands for LMM, NAÏVE, NALOG, RAST, SMEAR or G.

From Table 1, we note how all the considered small area strategies lead to more efficient estimates, on average, than the direct estimator, except for the NAÏVE estimator as was expected. The gain in efficiency, calculating from $AEFF_{dir}$, is relevant for all the small area estimators and varies from 35% for the RAST predictor to 53% for the LMM estimator. Therefore, the EBLUP estimator derived from a normal Linear Mixed Model shows the best performance in terms of accuracy, followed by G. On the other hand, the approximation necessary to produce the parameters' estimates in their actual scale, when the logarithmic transformation is applied, causes a greater instability in the results which makes the NALOG, RAST and SMEAR predictors less reliable than the other two. We also note that the simpler NALOG is more efficient than the two non parametric solutions, RAST and SMEARING, even though it is more biased.

Table 1. Performance measures

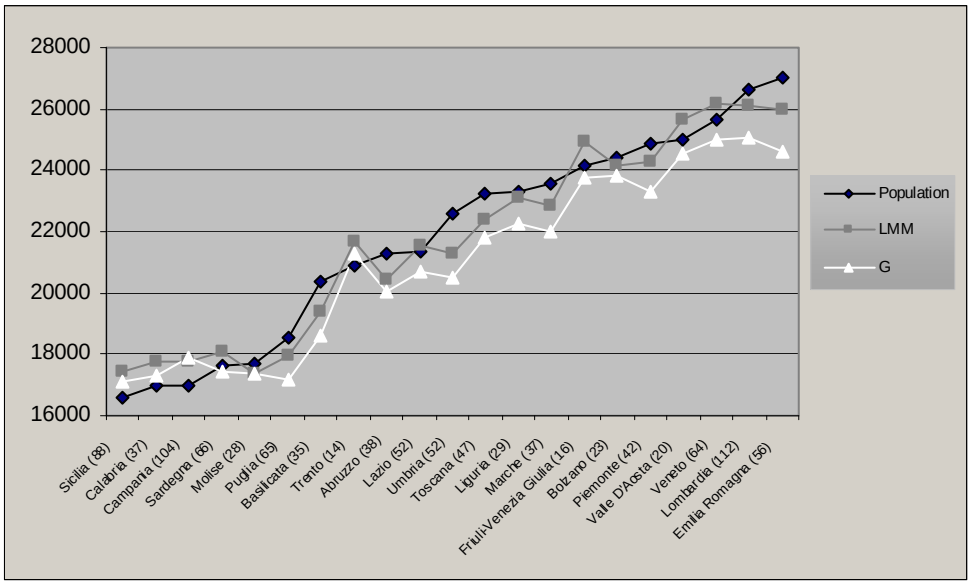
Estimator	*ARMSE%*	*AARB%*	*AARB%*	$AEFF_{Dir}$ %
Direct	0.787	0.0	0.0	100.0
NAÏVE	0.899	7.8	-7.8	114.2
NALOG	0.428	3.2	2.7	54.4
LMM	0.368	2.4	0.1	46.8
RAST	0.513	0.2	0.0	65.2
SMEAR	0.477	1.0	0.2	60.6
G	0.386	2.8	-1.2	49.0

Therefore, departure from normality seems to have a slight impact on punctual values of predictors. BLUP formulas can be derived without normality. Moreover, there are sound reasons to expect REML (and ML) estimators of variance components to perform well even if normality does not hold (see Jiang, 1996, for details).

Regarding the bias, very different results have been obtained for the different estimators. Looking at the *AARB* indicator, the least biased estimator is RAST even if, as we have already discussed, it is one of the less efficient ones. But this result is not surprising because, as explained in section 4, it is constructed as to control the transformation bias. Also the SMEAR estimator is not very biased, followed by LMM and then by G, which tends to underestimate the regional means (see *AARB´* column). Evidently the Gamma distribution is not completely suitable for the empirical income distribution that we have, even though the EBLUP derived by the Gamma model has the lowest variance. But the naïve estimators are the most biased. As expected, NAÏVE tends to seriously underestimate the parameters while NALOG corrects this bias but generates a positive bias, even though less important then the NAÏVE's one.

In order to further investigate the performance of the two preferable estimators (LMM and G), we looked at what happens in each region. To this purpose, in Figure 1, the means of the simulation replications obtained for them in the 21 regions are shown. To have a better view of the eventual effect on estimates of the value of the parameter, regions are ordered increasingly from the left to the right with respect to their population mean. This arrangement corresponds approximately to the regions arrangement form the south to the north because of the well-known greater incidence of poverty in the south of the country. We observe that, while the regional averages of the normal EBLUP are sometimes higher and sometimes lower than the population means, those related to the Gamma model are almost always lower except for the lowest levels of the population means. This means that G estimator tends to underestimate the high levels of income and to overestimate the low ones.

Figure 1. Regional means (sample size in brackets)



6. Conclusions

The main finding of the paper is that the empirical best predictor associated with the Battese Harter and Fuller model on the original scale of the study variable *Y* (given by equivalized household income) compares favourably to the back transformed predictors based on modelling the logarithmic transformation. On the one hand, this result is not completely surprising since, when obtaining the EBLUP, the normality assumption is used only in the REML estimation of the variance components, and this estimation method has been proven to be consistent even without normality (Jiang, 1996).

On the other hand, the result is relevant because the prediction of area means and totals using a linear mixed model on the original scale of *Y* requires that only the area means of the auxiliary variables are known, whereas methods considering the logarithmic transformation (as any other nonlinear transformation) need individual values for all units outside the sample. Moreover, as already noted in the introduction, in Fabrizi *et al.* (2007), we showed how the jackknife MSE estimator proposed by Jiang *et al.* (2002) is a good estimator of the design based MSE of this predictor. Besides, the extension of the work of Prasad and Rao (1990) and Datta and Lahiri (2000) to the estimation of MSE of predictors based on the non-linear transformations of the study variables have not been fully developed yet (see Slud, 2006 for more details).

The predictor based on the Gamma Generalized Linear Model, although in the case of our analysis it was outperformed by the Normal Linear Mixed Model, offers, at least in principle, an interesting alternative, since it does not require back-transformations and the MSE estimator of Jiang *et al.* (2002) may be applied.

Acknowledgements

Research was partially funded by Miur-PRIN 2003 "Statistical analysis of changes of the Italian productive sectors and their territorial structure", coordinator Prof. C. Filippucci. The work of Enrico Fabrizi was partially supported by the grants 60FABR06 and 60BIFF04, University of Bergamo.

We thank ISTAT for kindly providing the data used in this work.

REFERENCES

BATTESE G.E., HARTER R.M., FULLER W.A. (1988) An Error Component Model for Prediction of County Crop Areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28—36.

BETTI G., VERMA V. (2002) Non-monetary or Lifestyle Deprivation, in EUROSTAT (2002) *Income, Poverty Risk and Social Exclusion in the European Union*, Second European Social Report, 87—106.

CHAMBERS R.L., DORFMAN A.H. (2003) Transformed Variables in Survey Sampling, *Working paper* M03/21, Southampton Statistical Sciences Research Institute.

DATTA G.S., LAHIRI P. (2000) A Unified Measure of Uncertainty of Estimated of Best Linear Unbiased Predictors in Small Area Estimation Problems, *Statistica Sinica*, 10, 613—627.

DUAN N. (1983) Smearing Estimate: a Nonparametric Retransformation Method, *Journal of the American Statistical Association*, 78, 605—610.

ELBERS C., LANJOUW J.O., LANJOUW P. (2003) Micro-Level Estimation of Poverty and Inequality, *Econometrica*, 71, 355—364.

ÈLTETÖ O., FRYGIES E. (1968) New Income Inequality Measures as Efficient Tools for Causal Analysis and Planning, *Econometrica*, 36, 383—396.

Eurostat (2002) *European social statistics — Income, poverty and social exclusion*, 2nd report.

Eurostat (2003) *Regions. Nomenclature of territorial units for statistics, NUTS — 2003*, Methods and Nomenclatures series.

Eurostat (2005) *The continuity of indicators during the transition between ECHP and EU-SILC*, Working papers and studies, 2005 Edition.

FABRIZI E., FERRANTE M.R., PACEI S. (2007) Small Area Estimation of Average Household Income based on Unit Level Models for Panel Data, *Survey Methodology*, forthcoming.

FALORSI P.D., FALORSI S., RUSSO A. (1999) Small Area Estimation at Provincial Level in the Italian Labour Force Survey, *Journal of the Italian Statistical Society*, 1, 93—109.

JIANG J. (1996) REML Estimation: Asymptotic Behavior and Related Topics, *The Annals of Statistics*, 24, 255—286.

JIANG J., LAHIRI P., WAN S.M. (2002) A Unified Jackknife Theory for Empirical Best Predictor with M-estimation, *The Annals of Statistics*, 30, 1782—1810.

JIANG, J., LAHIRI, P. (2006) Mixed model prediction and small area estimation, Editor's invited discussion paper, *Test*, Vol. 15, 1, 1—96.

LEHTONEN R., SÄRNDAL C.-E., VEIJANEN A. (2003) The Effect of Model Choice in Estimation for Domains, Including Small Domains, *Survey Methodology*, 29, 1, 33—44.

KARLBERG F. (2000) Population Total Prediction Under a Lognormal Superpopulation model, *Metron*, 58, 53—80.

MCCULLAGH P., NELDER J.A. (1989) *Generalized Linear Models*, Chapman and Hall, London, England.

MUKERJI V. (1967) Type III Distribution and its Stochastic Evolution in the Context of Distributions of Income, Landholdings and Other Economic Variables, Sankhia, 29, A, 405—416.

PRASAD N., RAO J.N.K. (1990) Estimation of Mean-Squared Errors in Small Area Estimation, *Journal of the American Statistical Association*, 85, 163—171.

RAO C.R. (1971) Estimation of Variance Components – MINQUE Theory, Journal of Multivariate Analysis, 1, 257—275.

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

SAS Institute inc. (2006) The GLIMMIX Procedure. User Manual. June 2006. Downloadable at the address http://support.sas.com/rnd/app/papers/glimmix.pdf

SINGH A.C., MANTEL H.J., THOMAS B.W. (1994) Time Series EBLUPs for Small Areas Using Survey Data, *Survey Methodology*, 20, 1, 33—43.

SLUD E.V., MAITI T. (2006) Mean-squared Error Estimation in Transformed Fay-Herriot Models, *Journal of the Royal Statistical Society, ser.* B, 68, 239—257.

STEWART K. (2003) Monitoring social exclusion in Europe's regions, *Journal of European Social Policy*, 13, 4, 335—356.

VAN PRAAG B., HAGENAARS A., VAN ECK W. (1983) The influence of classification and observation errors on the measurement of income inequality, *Econometrica*, 51, 1093—1108.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 441—458

DANISH WORK ENVIRONMENT COHORT STUDY 2005: FROM IDEA TO SAMPLING DESIGN

Helene Feveile, Ole Olsen, Hermann Burr, Elsa Bach

ABSTRACT

The Danish Work Environment Cohort Study has been conducted every fifth year since 1990. This series of national surveys were designed primarily for surveillance of the work environment, but included cohorts permitting prospective, epidemiological analyses. Up until 2005 it was a split panel survey with essentially identical target population and sampling frame. In 2005 the survey was supplemented with a sample from specific industries and jobs. Additionally, the primary mode of data collection was changed from telephone interview to mail questionnaire and frame undercoverage emerged due to change in administrative praxis. Implementation of the initial ideas led to a non-standard design. This paper describes the 2005 survey; the initial ideas, the realized sampling design, the data collection. A weighting scheme and strategy for analysis are proposed.

Key words: National survey; stratified sample; weighting.

1. Introduction

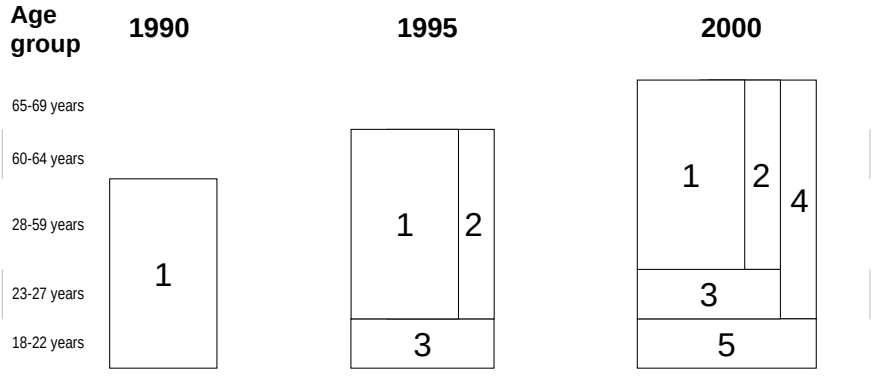
Implementing general ideas of design in real life surveys can raise all kinds of challenges. The aim of this paper is to encourage exchange of experiences when carrying out larger surveys over longer time periods. We will do this by presenting the design of the Danish Work Environment Cohort Study (DWECS).

DWECS is a series of national surveys conducted in 1990, 1995, 2000 and 2005 and has two overall objectives. The primary objective of the survey is surveillance of the work environment. For this purpose a battery of questions about noise and vibrations, physical demands and body postures, thermal, chemical and psychosocial working environment and working hours was included along with some questions about personal and occupational background factors. The secondary objective is to enable analytical, epidemiological analyses, i.e. association between occupational exposures and subsequent development of health symptoms. For this purpose a battery of health and well-being questions

was included along with questions about symptoms and diseases. To enable description of changes over time many questions were kept from one survey to the next. On the other hand new occupational issues appear on the agenda (due to e.g. technological development or change of political focus) and thus new questions were incorporated in the survey. A comprehensive description of DWECS 1990-2000 is given in Burr, Bjorner, Kristensen, Tüchsen and Bach (2003).

The design of DWECS reflects the multipurpose nature. Until the 2005-wave DWECS was a split panel survey combining independent, repeated cross-sections and built-in cohorts, thus enabling both the surveillance and the analytical, epidemiological aspect. The 1990-wave was based on a simple random sample of people from 18 to 59 years of age. In 1995 the sample from 1990 was contacted again, irrespectively of participation in 1990, along with a sample of 18 to 22-year-olds and a sample of people immigrated during the past 5 years. These additional samples were drawn in order to adjust for immigration and the ageing of the 1990 sample and the sizes were determined by proportional allocation (the group constitute the same proportion in the sample as in the total population). This design procedure was repeated in 2000. The design of the first three waves of DWECS is illustrated in figure 1.

Figure 1. The design of DWECS 1990—2000.



1. Cross-sectional sample 1990
2. Immigrant sample 1995
3. Young sample 1995
4. Immigrant sample 2000
5. Young sample 2000

In the first three waves of DWECS, data was collected using primarily telephone interviews with personal interviewing as second alternative.

In 2005 the financial situation for DWECS was enhanced and it was decided to enlarge the survey with a cross-sectional sample of people in the ages most active in employment. Furthermore additional samples within a chosen few industries and job groups of special interest were added since rare groups were not efficiently described in an overall national sample. Finally a different framework was imposed, as data should primarily be collected using mail questionnaires with telephone interviews as second alternative.

Since 1995 inhabitants in Denmark have had the possibility to request survey exemption; neither administrative or research orientated surveyors can obtain their address from centralised registers. In 2000 the procedure for requesting survey exemption was changed. Previously citizens had to contact the administration specifically to request survey exemption; afterwards the request was as a tick box on the change of address form. As a consequence the fraction of people with survey exemption increased from a total number of 24 to above 10% of the population in 2006 (Thorsted 2007). The coverage of the 2005-wave was affected by this.

Compared to the previous waves of DWECS, the 2005-wave met three new challenges; the change in sampling design, the change in mode of data collection and the increasing propensity to request survey exemption. As DWECS is an important part of the system for surveillance of the work environment in Denmark there is a need for a uniform, transparent and comprehensible procedure for estimation and reporting.

This paper focuses on the development of the sampling design of DWECS 2005 with the modes of data collection and survey exemption as additional challenges. The aim is to describe the conduct of DWECS 2005 including the initial ideas, the actual sampling design, the data collection, and to discuss the gap between the idealized ideas and the more fuzzy reality. The chosen weighting scheme is presented.

2. Sampling and data collection of DWECS 2005

The Danish, centralised civil register (CRS) contains information for every individual who is or has been an inhabitant of Denmark since 1968 and is updated daily (CRS link). This register contains information on among other things gender, age, address, citizenship and whether or not the individual has requested survey exemption.

Since 1981 the entire population in Denmark has been classified annually according to employment status, industry and occupation in the register based labour force statistics (RAS). Information on occupation is updated more often in two different registers: Indices of average earnings for the private sector (LPS) and indices of earnings for the public sector (LOS). The industry and job specific samples were drawn from these registers.

The target population for DWECS was the 18-74-year-old inhabitants in Denmark. The sampling frame for the study consisted of those without survey exemption among 18-74-year-olds, registered in CRS. In reality the sampling was done in CRS as well as in certain parts of RAS, LPS and LOS (corresponding to specific industry and job codes). The units in the latter three registers were also registered in the CRS and comprised by the survey exemption issue.

2.1. Sample selection

The DWECS sample in 2005 consisted of five subsamples:

The *follow-up sample* consisted of 10,131 people who were invited to participate in DWECS 2000 and still alive and living in Denmark in 2005 (they were between 23 and 74 years of age). The sample is made up by survivors originally sampled in CRS.

The *young adult(s) sample* was a simple random sample of 943 18—22-year-old people from CRS.

The *immigrant sample* was a simple random sample from CRS of 236 23—59-year-old immigrants not residing in Denmark in 2000.

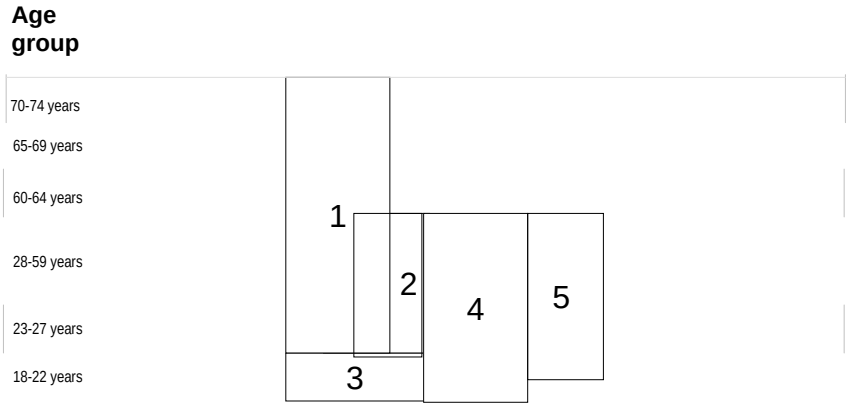
The *cross-sectional sample* was a simple random sample of 8,545 18—59-year-old people from CRS.

The *industry and job specific sample* was a stratified sample of 6,957 people between 20 and 59 years of age and working within nine specific industries or six specific jobs.

As in the previous waves of DWECS, the sizes of the young adult(s) sample and the immigrant sample were determined by proportional allocation (relative to the 18—59-year-olds in the follow-up sample). The cross-sectional sample was an enlargement of this age group in the national sample.

People from the specific industries were sampled from RAS (RAS link) and people from the specific jobs were sampled from either LOS (LOS link) or LPS (LPS link). A stratified two-phase sampling design was applied. For each industry and job a sample from the register was first obtained (a total of 6,957 PSUs) then followed by a screening of respondents in the beginning of the telephone interview. If the respondent was considered not to belong to the specific industry or job under investigation, the interview was terminated, otherwise it was completed. The aim was to attain 200 completed interviews in each industry or job group. People registered in RAS, LOS and LPS are also registered in CRS, but unfortunately data on classification according to RAS, LOS and LPS were not requested for the part of the sample derived directly from CRS. Therefore DWECS 2005 was not based on a stratified sample, even though it appeared so at first glance. The five subsamples constituting DWECS 2005 is illustrated in figure 2.

Figure 2. The design of DWECS 2005. Numbered boxes illustrate subsamples; cross-hatched areas indicate telephone interviewing as initial mode of data collection.



1. Follow-up sample
2. Immigrant sample
3. Young sample
4. Cross-sectional sample
5. Industry and job specific sample

2.2. Data collection

As mentioned above, data in 2005 should primarily be collected using mail questionnaires with subsequent telephone interviews of initial nonresponders. However there were some exceptions to this strategy (cross-hatched areas in figure 2). In two of these scenarios telephone interview was the primary mode of data collection followed by mail questionnaires to initial nonresponders. Since description of changes over time was one of the main aims of DWECS and mode of data collection can affect the response patterns (Feveile, Olsen and Hogh 2007), a randomised trial was incorporated in DWECS: In the part of the follow-up sample below 60 years of age, 2000 people were randomised to telephone interview while the remaining 5725 received a mail questionnaire. The immigrant sample was, due to practical requirements, contacted by SFI-SURVEY using telephone interview as initial mode of data collection.

People who received a mail questionnaire were offered the opportunity to complete the questionnaire on the web instead of the paper version. The telephone interviews were conducted by SFI-SURVEY, a department at the Danish National Institute of Social Research.

The industry and job specific sample was contacted by Statistics Denmark using telephone interview as only mode of data collection, since only Statistics Denmark are authorized to sample and identify persons in RAS, LOS and LPS.

The survey consisted of 17 demographic items, 102 items on health and 130 items on occupation (plus 8 specifically for employees or 4 specifically for self-employed). Telephone, internet and mail questionnaire respondents were asked the same questions. The mail questionnaire was 44 pages long and the telephone interview typically lasted 45 minutes. In all data collection formats the demographic items were first; in the mail questionnaire the items concerning health were printed before the items concerning occupation while the ordering was reversed in the telephone interview (as well as in the web-version of the questionnaire). All respondents were asked all items concerning health, irrespectively of their labour market status.

Data collection by questionnaires and telephone interviews by SFI-SURVEY took place in the period October 2005 — May 2006. The data collection performed by Statistics Denmark took place November 2005 — April 2006 with consecutive start of the 15 jobs and industries over the period. Due to the time lag in the registers, the register classification of industry was 2—2½ years old at the time of data collection, for the private occupations it was 5—8 months old, and for the public occupations it was 3—6 months old**.**

2.3. Weighting

The design was chosen so the distributions of industry and jobs were different in the total sample than in the frame. This should be reflected in the design weights.

As already noted, we lacked information about RAS, LOS and LPS classification of the follow-up, young adult(s), immigrant and cross-sectional sample. This indeterminate strata affiliation made us unable to calculate the correct inclusion probabilities. To overcome this obstacle and maintain a transparent approach we regarded and analysed the 2005-wave of DWECS as a stratified simple random sample with 17 pseudo strata: One pseudo stratum of 18—59-year-old inhabitants of Denmark, one pseudo stratum of 60—74-year-old inhabitant of Denmark, nine specific industry pseudo strata and six specific job pseudo strata.

Let $h \in \{1,\ldots,17\}$ index these 17 pseudo strata and let N_h and n_h denote the population size and the sample size in pseudo stratum h. We defined the design weight for an individual in pseudo stratum h as the inverse sampling fraction, that is

$$w_h^d = \frac{N_h}{n_h}.$$

Experience shows that response rates differ between age and gender groups and it was also expected to differ between the specific industries and jobs. Thus a post-stratification was defined by the 15 industry and job specific pseudo strata, the 60-74-year-olds divided by gender and the 18-59-year-old pseudo stratum divided by gender and into 4 age groups (18-29, 30-39, 40—49, 50—59); a total of 25 pseudo strata.

Let $k \in \{1,\ldots,25\}$ index the 25 post-stratification pseudo strata. We note that the post-stratification was a further partition of the initial 17 pseudo strata. Thus there exists a surjective function $k \mapsto h(k)$. Let n_k and m_k denote the sample size and the number of respondents in post-stratification pseudo stratum *k*. We defined the nonresponse weight for an individual in post-stratification pseudo stratum *k* as

$$w_k^r = \frac{n_k}{m_k}.$$

Thus respondents in post-stratification pseudo stratum *k* were assigned the weight

$$w_k = w_k^r \cdot w_{h(k)}^d = \frac{n_k}{m_k} \cdot \frac{N_{h(k)}}{n_{h(k)}}.$$

The final weights were defined from the w_k –weights by dividing with the sum of weights and multiplying with the population size in the specific (of the 17) pseudo stratum. By applying these final weights it was ensured that the weights summed to the population size in each of the 17 pseudo strata. However the total number of inhabitants in Denmark was overestimated since the 15 industry and job specific pseudo strata were subsets of the large pseudo stratum of 18-59-year-old inhabitants in Denmark (see discussion).

No weighting was applied in the 1990, 1995 and 2000 surveys.

2.4. Statistical method

Prevalence and confidence limits were estimated using the PROC SURVEYFREQ procedure of SAS 9.1. The design was specified as stratified simple random sampling (no finite population correction) with 17 strata. The 25 different individual weights previously defined were employed.

3. Results

Table 1 shows the population size, the sample size, the sampling fraction, number of respondents and response rates. In the ages most active in employment and exclusive of the industry and job specific sample, the sampling fraction was 0.57%. The sampling fraction was 0.31% among the oldest (corresponding to the sampling fraction in previous waves of DWECS). In the

industry and job specific samples the sampling fractions ranged from 1.08% to 22.78%. The goal of at least 200 interviews in the industries and jobs in special focus was not achieved but most of the discrepancy was caused by a technicality (by mistake no age restrictions were imposed in the sample).

Table 1. Population size, sample size, sampling fraction, number of respondents and response rate in DWECS 2005, in each pseudo stratum

Pseudo* stratum	Population size	Sample size	Samp. frac. (%)	Number of resp.	Response rate (%)
Inhabitants of Denmark (age ≤ 59)	3,053,614	17,449	0.57	10,771	61.7
Inhabitants of Denmark (60 ≤ age ≤ 74)	782.646	2.406	0.31	1,642	68.3
Industry: Manufacture of textiles and leather	8,580	404	4.71	191†	77.2‡
Industry: Poultry dressing station, fish and animal feed plant	9,112	485	5.32	203†	71.6‡
Industry: Manufacture of food, beverages and tobacco	35,377	382	1.08	197†	81.4‡
Industry: Manufacture of means of transport and shipyard	14,908	415	2.78	197†	63.1‡
Industry: Fire-fighting service and salvage corps	7,834	400	5.11	197†	59.5‡
Industry: Personal care	22,519	730	3.24	197†	66.3‡
Industry: Agriculture	37,048	981	2.65	204†	61.7‡
Industry: Horticulture and forestry	13,978	460	3.29	201†	72.8‡
Industry: Hotels and restaurants	46,118	689	1.49	199†	66.3‡
Job: Carpenter, joiner	4,619	371	8.03	183†	57.4‡
Job: Concrete worker or general worker	1,633	372	22.78	194†	74.7‡
Job: Taxi driver	9,401	311	3.31	134†	67.2‡
Job: Bus driver	1,533	330	21.53	166†	54.9‡
Job: Police	2,510	290	11.55	188†	66.2‡
Job: Social worker	6,733	337	5.01	176†	80.1‡

* Not a genuine stratification
† Number of respondents that belonged in the specific industry or job
‡ Respondents (correctly or incorrectly classified) / sample

The overall response rate in the combined follow-up, immigrant, young adult(s) and cross-sectional sample was 62.5%. The response rates in the industry and job specific samples ranged from 54.9% to 81.4%.

Table 2 shows the fraction of the respondents that passed through the screening questions and were judged to belong to the specific industry or job under investigation. A considerable difference between the fractions of correctly classified persons in the register depending on the specific industry/job was observed; the fractions ranged from 33.72%, among those belonging to the agricultural trade according to the register, to 97.92% among those registered as policemen.

Table 2. Percentage of respondents considered correctly classified in RAS, LOS and LPS

Industry or job specific strata	Sample size	Percentage in industry/job
Industry: Manufacture of textiles and leather	404	61.22
Industry: Poultry dressing station, fish and animal feed plant	485	58.50
Industry: Manufacture of food, beverages and tobacco	382	63.34
Industry: Manufacture of means of transport and shipyard	415	75.19
Industry: Fire-fighting service and salvage corps	400	82.77
Industry: Personal care	730	40.70
Industry: Agriculture	981	33.72
Industry: Horticulture and forestry	460	60.00
Industry: Hotels and restaurants	689	43.54
Job: Carpenter, joiner	371	85.92
Job: Concrete worker or general worker	372	69.78
Job: Taxi driver	311	64.11
Job: Bus driver	330	91.71
Job: Police	290	97.92
Job: Social worker	337	65.19

Table 3 shows the sample size and the number of usable completed questionnaires /interviews in the follow-up, the young adult(s), the immigrant and the cross-sectional sample broken down by mode of data collection (cf. figure 2). Combining both mail questionnaires and telephone interviews ensured a response

rate over 65% for the follow-up part of the sample. The young adult(s) sample was much more reluctant to participate and this tendency was found again in the cross-sectional sample where the response rate was 48.4% among the 775 18—22-year-olds. The immigrant sample had a response rate of 33%.

Table 3. Sample size, number of respondents to initial and secondary modes of data collection and total response rate, by sample and data collection strategy

Sample	Sample size	Respondents Initial mode	Respondents Secondary mode	Total response rate %
Follow-up phone*	2000	1251	67	65.9
Follow-up mail†	5725	3297	450	65.4
Follow-up old†	2406	1520	122	68.2
Young adult(s)†	943	334	118	47.9
Immigrant*	236	69	9	33.1
Cross-sectional†	8545	4475	701	60.6

* Initial mode of data collection was telephone interview. Secondary mode was mail or internet questionnaire.
† Initial mode of data collection was mail or internet questionnaire. Secondary mode was telephone interview.

Due to the multiplicity of combinations of subsamples and data collection strategies a complete flow chart of DWECS 2005 is beyond this paper. Figures 3 and 4 are examples of flow charts that arise from the two main data collection strategies in DWECS 2005.

Figure 3. Flow chart of the data collection in the part of the follow-up sample randomised to mail questionnaire.

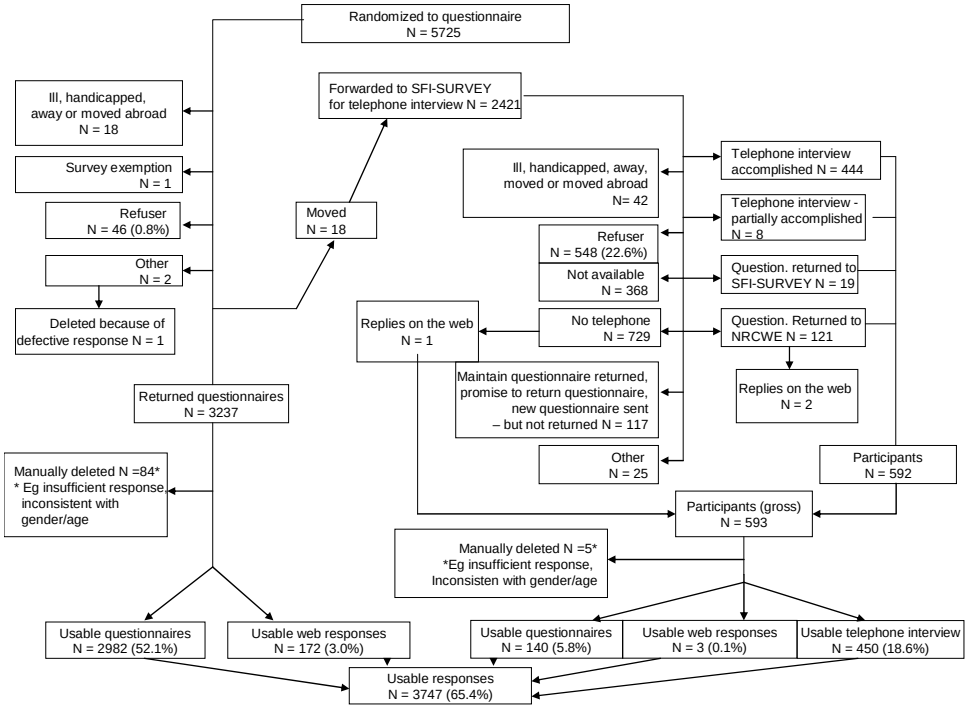


Figure 4. Flow chart of the data collection in the part of the follow-up sample randomised to telephone interview.

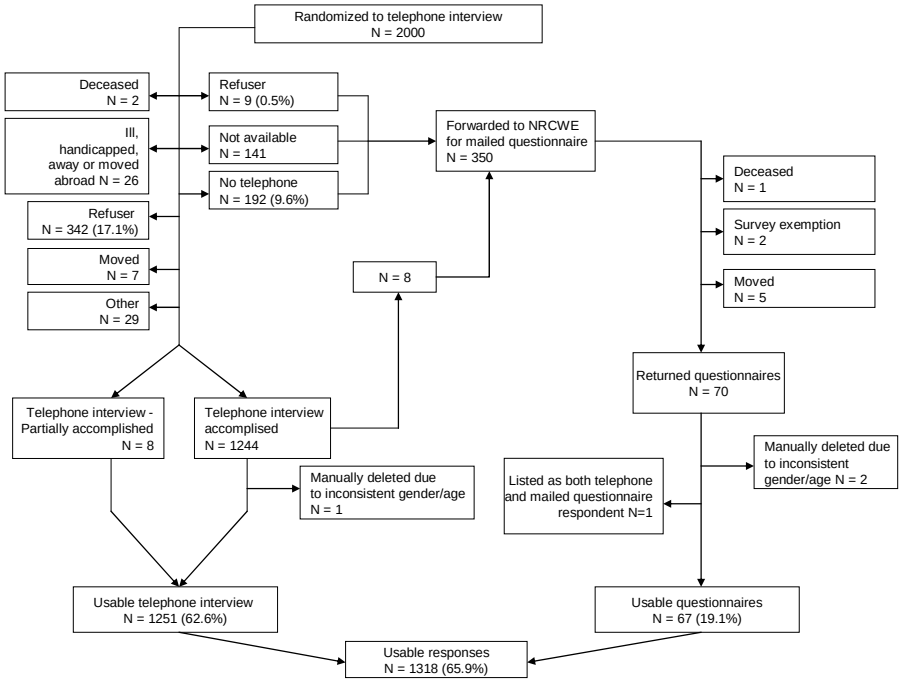


Table 4 shows the overall response rate (irrespectively of the how data was collected) for men and women stratified by age. The table is based on the total sample exclusive of the job and industry specific sample. The association between age and inclination to response that was indicated in table 3 is elaborated. Marginally the response rate increased with age, and men had a lower response rate than women. The total response rate of 62.5% covered a range from 46.1% among the young men to 72.2% among the 50—59-year-old women.

Table 4. Total response rate in follow-up, immigrant, young adult(s) and cross-sectional sample, stratified by age

Age	Men		Women		Total	
	Sample Size	Response rate (%)	Sample Size	Response Rate	Sample Size	Response Rate
18—29	2087	46.1	1963	57.1	4050	51.4
30—39	2144	57.0	2216	64.5	4360	60.8
40—49	2259	60.1	2250	70.6	4509	65.3
50—59	2229	64.3	2301	72.2	4530	68.3
60—74	1178	70.0	1228	66.6	2406	68.3
Total	9897	58.6	9958	66.4	19855	62.5

The age and gender distributions in the total population of 18—74-year-old inhabitants in Denmark (based on registers) were compared to

i. the estimated age and gender distributions we got from the total sample using the design weights $\left(w_h^d\right)$ and
ii. the estimated age and gender distributions based on the respondents using the nonresponse adjusted weights (the normalized w_k).

Table 5 shows the raw age and gender distributions in the total population in Denmark (including inhabitants with survey exemption) and among the respondents. Table 5 also shows the estimated age and gender distributions based on the total sample using the design weights and the estimated age and gender distributions based on the respondents in DWECS 2005 using the final weights (adjusted for nonresponse). The same four age distributions were calculated for the two gender domains but the results were omitted since the tendencies were the same as those in table 5.

Table 5. Distribution of age and gender in the total population (18-74 years of age), among the respondents in DWECS 2005, and obtained by weighted estimation in the sample and among respondents in DWECS (%)

	Population (Register based)	Resp.	Estimated from sample (design weights)	Estimated from respondents (final weights)
Age				
18—29	19.4	15.9	18.5	18.4
30—39	20.5	22.1	20.3	20.3
40—49	20.4	25.3	21.0	21.0
50—59	19.3	25.9	20.9	20.9
60—74	20.4	10.8	19.3	19.3
Gender				
Men	50.1	50.9	50.3	50.4
Women	49.9	49.1	49.7	49.6

As expected the age and gender distribution among the respondents differed from the population proportions. Applying the weights remedied this to some extent.

3.1. Presentation of results

The results of the survey were presented in Danish in an overall report and six leaflets. These are available on the web site (NRCWE link) along with additional material in the form of tables and bar charts (also in Danish). The presentation

reflected the primary objective of describing the work environment uniformly and transparently.

4. Discussion

In the initial planning of DWECS 2005, the idea was to repeat the design procedure from the previous waves, thus obtaining a representative, national sample providing an overview of the work environment in Denmark and at the same time following up on previously defined cohorts. Furthermore it was the ambition to enable analyses of a few industries and job groups of special interest by adding subsamples of these. Had we had access to registers that stratified the population perfectly according to industry and job at the date of sampling, we would have ended up with a stratified, simple random sample.

Implementing these stratification ideas in practice led to a more complex design. The classifications according RAS, LOS and LPS were not available for the follow-up, immigrant, young adult(s) and cross-sectional sample precluding a genuine stratified sample and calculation of stringent design weights. Furthermore RAS, LOS and LPS were not perfect sampling frames for the industry and job specific samples. Due to both misclassification and update frequency of the registers, a fraction of the people sampled according to RAS, LOS and LPS did not work in the specific industry or job under investigation at the time of the interview. For economic reasons, these interviews were terminated. The major challenge in determining the correct weights and estimation procedures was the fuzziness of the stratification. We therefore introduced the term “pseudo stratification” to emphasise this deviation from the ideal textbook design.

4.1. Coverage

The fraction of persons sampled from the industry/job registers who had their industry/job confirmed at the time of interview varied according to job (range 64—98%) and industry (range 34—83%) (cf. table 2). This reflected a combination of the validity of the registers and the loyalty towards the industry/job. By validity we mean accuracy of the register at the time of classification (Statistics Denmark considers the information on industry in RAS to be of high quality (RAS link) and the information in LPS and LOS to be of reasonably high quality (LPS link and LOS link)) and by loyalty we understand to what extent people remain in the industry/job from the time of classification to the time of interview.

Of those registered as (publicly employed) policemen, 98% were still considered policemen when interviewed 3—6 months after the register had been processed, whereas only 65% of those registered as social workers 3-6 months prior were still considered social workers when interviewed. For jobs in the private sector the processing time for the register was slightly longer, 5—8

months, and the range of percentages were slightly lower: 64% (taxi drivers) to 92% (bus drivers). For the industry samples the time span between register classification and interview was much longer, 2—2½ years, and thus the range of percentages was much lower: 34% (agriculture) to 83% (fire-fighting service, etc.). For three industries less than half the contacted persons could at the time of interview confirm that they belonged to the industry (agriculture, personal care, hotels and restaurants). The respondents in these samples may be biased in the sense that they mainly represent employees with stable job conditions in an industry with an otherwise high turnover.

Overall using RAS, LPS and LOS as sampling frames did not capture people in the specific industries and jobs with only a few months experience/seniority and respondents may in general have had a different work environment than the average worker in the industry/job.

The new procedure for requesting survey exemption introduced in 2000, is a challenge for conducting surveys based on data from CRS, and the survey exemption is maintained until the individual requests for its annulment. In 2006 about 600,000 had requested survey exemption corresponding to 20—25% of those having changed address. The rate of survey exemption varied with demographic and socioeconomic status (e.g. age and education). Before 2000 CRS was an ideal sampling frame for national population surveys in Denmark and the undercoverage was essentially zero. In 2005 more than 10% of the population was not covered when using CRS as sampling frame. Since the rate of undercoverage varied with age and education and these factors were highly correlated with many occupational exposures this might have resulted in substantial bias in the estimates.

4.2. Response rate

With three exceptions the industry and job specific samples had response rates greater than or equal to the overall national sample in the same age interval (Table 1) despite the fact that the total response rate in the overall national sample (exclusive of the industry and job specific sample) of 62.5% was the result of a mixture of data collection strategies (mail questionnaire with subsequent telephone interviews of initial nonresponders and vice versa). The randomised trial of mode of data collection resulted in similar total response rates (65.9% versus 65.4%), and in both cases the response rate benefited from changing mode of data collection among initial nonresponders (Table 3). A mixture of modes of data collection however increased the complexity of data collection process (cf. figures 3 and 4).

The response rate in the overall national sample was moderate, and furthermore response rates varied among men and women and between the various age groups (cf. Table 4). The variation between groups made some nonresponse adjustment seem attractive. Table 5 shows that applying the final

weights (adjusted for nonresponse) resulted in estimates vary similar to those obtained from the sample using the design weights. Since the nonresponse weights were constructed from age and gender, this result was not surprising and does not justify the nonresponse weighting altogether.

The response rates among the youngest (47.9%) and especially among the immigrants (33.1%) made it obvious that some kind of extra effort should be done with respect to increasing the response rate when conducting surveys in these subpopulations.

It appears from table 5 that the estimated age and gender distribution was acceptable when based on the sample using the design weights. Since citizens with survey exemption were removed from the sampling frame, the estimates could not be expected to hit the mark. The discrepancies could however be due to poor specification of design or calculation of weights as well as the described disagreement between sampling frame and target population.

4.3. Weighting and statistical method

The design weights were constructed based on the sampling fractions in the pseudo strata. The industry and job specific sample was added to the national sample, since these groups were too rare to be efficiently described based on the overall sample. Evidently the sampling fractions for these pseudo strata were larger than those for the overall national sample. Since the pseudo strata were not disjoint, the total population size was overestimated by the population count of the specific industries and jobs, i.e. 221,903 individuals (5.5%), but the sampling fraction within each pseudo stratum was correctly specified.

The design weights were multiplied with nonresponse weights in order to reduce differential nonresponse bias. The follow-up, immigrant, young adult(s) and cross-sectional sample were post-stratified according to age and gender and nonreponse weights were calculated in these 10 post-stratification pseudo strata. The nonresponse weights in the 15 industry and job specific subsamples were calculated as sample size divided by respondents willing to participate in the telephone interview; this was done whether or not the interview continued or was terminated because the interviewer judged that the respondent did not work in the specific job or industry. The nonresponse weighting cannot be expected to eliminate nonresponse bias, but hopefully it decreases the bias in some instances.

We analysed as if the sample size had been fixed in advance. In reality the size of the industry and job specific sample was random, since it was determined by the criteria of 200 completed interviews in each industry/job. Thus, the number of persons sampled in each industry and job depended on the response rate, the percentage of respondents considered correctly classified in RAS, LOS and LPS and the seniority within the industry/job.

The two-phase sampling design in the industry and job specific sample was not specifically taken into account in the analyses, since the fact that strata

affiliation was not determined for all members of the population remains the main problem. For the same reason alternative weighting schemes have not been considered. E.g. calibration with respect to the estimated counts of the correctly classified industry and job specific subpopulations seems contradictory, just as calibration with respect to the observed counts of the same subpopulations appear unnatural, since many of the respondents are excluded based on screening for relevance of the classification.

For many questions response patterns differ with mode of data collection (Feveile, Olsen and Hogh 2007). For such questions it can be necessary to restrict analysis (and thus recalculate the weights) to embrace only those assigned to telephone interview (cf. figure 2); in this case the industry and job specific respondents constitute the larger part of the available observations (although not of the weighted observations).

4.4. Conclusion

When designing larger surveys using various overlapping registers, the possibilities for obtaining a genuine stratification should be considered. When simple random sampling is employed for each subsample (as in the presented example DWECS 2005), register data to be used for stratification should be obtained for all units in the sample. If a more complex sampling design is used for some of the subsamples, it should be considered whether some extended merging of register data for the entire frame is necessary in order to obtain a genuine stratification. Another option would be trying to adapt the design to some kind of two-phase sampling for stratification.

With time-varying register affiliation it should be kept in mind that not only the accuracy of the register affects the coverage. Also the time-gap between register classification and the survey along with turnover in the register might lead to bias in the selection.

Acknowledgement

The authors were inspired by a well-written paper by Michael Davidsen and Mette Kjøller (Davidsen and Kjøller 2002).

REFERENCES

BURR, H., BJORNER, J. B., KRISTENSEN, T. S., TÜCHSEN, F. AND BACH, E. (2003). Trends in the Danish work environment in 1990—2000 and their associations with labour-force changes. *Scandinavian Journal of Work, Environment & Health*, 29, 270—279.

CRS link. http://www.cpr.dk/cpr/

DAVIDSEN, M. AND KJØLLER, M. (2002). The Danish Health and Morbidity Survey 2000 - Design and Analysis. *Statistics in Transition*, 5, 927—942.

FEVEILE, H., OLSEN, O. AND HOGH, A. (2007). A randomized trial of mailed questionnaires versus telephone interviews: response patterns in a survey. *BMC. Medical Research Methodology*, 7, 27.

LOS link.
http://www.dst.dk/HomeUK/Guide/documentation/Varedeklarationer/emnegruppe/emne.aspx?sysrid=84002

LPS link.
http://www.dst.dk/HomeUK/Guide/documentation/Varedeklarationer/emnegruppe/emne.aspx?sysrid=865

NRCWE link. http://www.nrcwe.dk/

RAS link.
http://www.dst.dk/HomeUK/Guide/documentation/Varedeklarationer/emnegruppe/emne.aspx?sysrid=848

THORSTED, B. L. (2007). Forskerbeskyttelse i CPR. In Symposium i anvendt statistik, Linde, P. (eds) pp 74—84. København: Institut for Økonomi Aarhus Universitet, Danmarks Statistik (abstract - in Danish).

HELENE FEVEILE, National Research Centre for the Working Environment, Lersø Parkallé 105, DK-2100 Copenhagen, Denmark

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 459—474

DESIGN-BASED INFERENCE FROM ON-SITE SAMPLES

Thomas Laitila, Annica Isaksson

ABSTRACT

On-site sampling is used in surveys where a frame of the population of interest is not near at hand. The population may for instance consist of individuals visiting some fishing-waters or a shopping mall. We demonstrate how general conclusions can be drawn from on-site sample data by use of design-based inference. For an on-site sample of individuals, first- and second-order inclusion probabilities are derived, thus making Horvitz-Thompson and variance estimation possible. The performances of some alternative estimators of the population mean are compared in a simulation, based on real survey data on anglers visiting the Kaitum river. The derived inclusion probabilities are also used to evaluate an often-made on-site sampling assumption: that the individuals' inclusion probabilities are proportional to their number of visits to a site. It is shown that for the sampling design under study, the assumption holds for at least one special case, but not in general.

Key words: Inclusion probability; Angler survey; Shopping Center Sampling;

1. Introduction

Pollock, Jones and Brown (1994) propose on-site sampling methods for angler surveys. The idea is to sample anglers at fishing sites for collection of data on angler effort and catch. Their sampling unit of interest is "fishing trip" and their aim is to estimate characteristics of the population of fishing trips made during a season. The authors advocate multistage sampling designs, e.g., with selection of time periods in the first stage, selection of access points in the second stage, and selection of anglers in the third stage are often advocated. If probability sampling is used in each stage, standard methods for estimation and inference can be applied.

Shaw (1988) also considers on-site sampling for surveys of visitors to recreational sites such as angling sites. He is interested in the population of visitors rather than the population of visits. An important study variable is the

total number of visits made by the visitor; information utilized for valuation of site characteristics. Standard techniques for inference to the population of visitors require simple random sampling of visitors, and thus are not appropriate for on-site sample data.

There are several other areas aside from angler surveys where on-site sampling is used to collect data intended for inference to a population of visitors. One area is marketing, where visitors to shopping malls are sampled for studies of consumer behavior and consumer attitudes. Two recent examples are Keillor, D'Amico and Horton (2001) and Keen et al. (2004). Block et al. (2002) use shopping mall sampling for evaluation of effects of antidrug advertisement. Brenner (1998) uses on-site samples in a study of the potentials of introducing everyday mathematics into classrooms.

There are only a few contributions in the literature on the topic of estimation and designing on-site sample surveys for inference to the population of visitors. Sudman (1980) proposes locations at site for sampling, application of quota sampling techniques, and weighting of observations with respect to frequency of visits. Nowell and Stanley (1991) utilize results by Cox (1969) for correction of length biased samples. Shaw (1988) derives sample inclusion probabilities using a superpopulation model and applies results for estimation of a Poisson model of visit frequencies. Laitila (1998) derives concentrated log-likelihood functions for the study of site choices under on-site sampling. A generic assumption behind the proposed correction methods for estimation of population characteristics using on-site samples is the proportionality of the sample inclusion probability to the number of visits to the site (or the length of stay). This is in contrast to e.g. Cox (1969) who provides a description of the sampling process of fibers making the proportionality assumption feasible. In general, the sampling of individuals on-site is not well described and the validity of the proportionality assumption is not known.

One potential method for performing on-site sampling is indirect sampling and the application of the generalized weight share method (Lavallée, 2007). This method can be used when the population of interest is unknown but can be identified via links to a known population. In our context it might be possible to apply indirect sampling by treating time periods as the known population. Each fishing trip made by an angler is linked to a specific time period. Thus, the unknown population can be identified via a population of time periods.

This paper contributes with derivations of first- and second-order inclusion probabilities for population units under a multistage sampling design for on-site sampling. The paper rests upon traditional theory for sampling from finite populations (e.g. Särndal, Swensson and Wretman, 1992) and provides a new framework for on-site sampling studies. Our derivations are useful in at least three ways. First, they make it possible to apply traditional estimation methods on on-site data. Second, they can be used to validate the commonly made assumption of proportional inclusion probabilities. Third, they facilitate comparisons between

traditional methods of inference and alternative methods such as indirect sampling combined with generalized weight sharing.

The theoretical framework is introduced in Section 2: the target population and the sampling design under consideration are defined. Section 3 deals with on-site sampling in the multi-period/multi-entrance case, while Section 4 deals with two simple special cases. A numerical illustration is given in Section 5. A discussion is saved for the final section.

2. Population, sampling design and estimators

2.1. Population

In all surveys, it is crucial to define the population of interest. Unless the population is clearly defined, the properties of a given sampling design can not be established, and the foundation for statistical inference is weak. There are two potentially useful population definitions when collecting on-site sample data. The first one is considered by Kalton (1991) who describes methods for sampling "flows of mobile populations" (such as visitors to a summer sculpture exhibition in a city park). In this context, Kalton defines a population of visits. This definition implies that two visits made by the same person are viewed as two different elements of the population. When the population is defined in this manner, sampling and estimation is, theoretically, not very problematic. The situation can be handled by a multi-stage sampling approach (e.g, Särndal et al., 1992, Ch. 2) where time periods and entrances are selected at different stages (Kalton, 1991).

In this paper, we are interested in cases when the population is defined as the set of visitors – see Definition 2.1. Then, even though a standard multi-stage sampling procedure is applied, standard multi-phase estimators are not appropriate. The reason is that a single element in the population is not uniquely allocated to one of the "groups" of elements defined by the multi-stage sampling procedure. A person visiting the site more than once has several possible occasions to be selected into the sample. This feature must be accounted for when drawing inference to the population stated in Definition 2.1.

Definition 2.1: The population of interest consists of the set of individuals visiting the site at least once during the connected time period τ. This set of individuals is denoted *U* and is of size *N*.

2.2 Sampling design

In order to survey the population *U* in Definition 2.1, a three-stage sampling design is proposed. The design is based on the following assumption.

Assumption 2.1: Two different visits to the site made by an individual are at least δ time units apart.

The time interval τ is partitioned into N_I periods, labeled $i = 1,...,N_I$, of equal time length ρ, where $\rho < \delta$. By this restriction, the visits made by an individual during τ are allotted to separate time periods. At each visit to the site, the individual enters the site only once. The number of available entrances is denoted by N_{II}.

A three-stage design with time periods as primary sampling units (PSU), entrances as secondary sampling units (SSU) and visitors as tertiary sampling units (TSU) is now considered. This order of selection seems natural and coincides with sampling designs considered by e.g. Pollock et al. (1994). Someone might prefer to use entrances as PSUs and time periods as SSUs. We expect our derivations in Section 3 to hold also for this situation (after an appropriate change of notation). The set of PSUs (of size N_I) is symbolically represented by $U_I = \{1,...,i,...,N_I\}$. The set of SSUs (of size N_{II}) is represented by $U_{II} = \{1,...,q,...,N_{II}\}$. For PSU *i* and SSU *q*, the number of visitors is denoted N_{iq}, and the visitors are labeled $k = 1,...,N_{iq}$. The visitors are represented by their labels, and the population of TSUs for PSU i and SSU q is given by $U_{iq} = \{1,...,k,...,N_{iq}\}$. The three-stage sampling design now reads:

Stage 1. A sample s_I of PSUs of size n_I is drawn from U_I by simple random sampling without replacement (SRS).

Stage 2. For every PSU $i \in s_I$, a sample s_{IIi} of SSUs of size n_{II} is drawn from U_{II} by SRS.

Stage 3. For every SSU $q \in s_{IIi}$, a sample s_{iq} of TSUs of size n_{iq} is drawn from U_{iq} by Bernoulli sampling (BE) with inclusion probability α.

The final sample of visitors is given by

$$s = \bigcup_{i \in s_I} \bigcup_{q \in s_{IIi}} s_{iq}$$

of random size *n*. Note that a sample unit enters the final sample *s* only once although it may have been sampled several times.

A BE design is chosen for the stage 3 sampling of respondents. An alternative design frequently employed in field studies is systematic sampling (SY). However, for the design to be measurable, SY sampling with at least two random starts needs to be applied in stage 3. The BE design provides a measurable design, and the independence of selections makes it theoretically more tractable.

A number of interesting special cases can be defined from this design. If $n_I = 1$ and $N_{II} = 1$ the design is called a *one-period one-entrance* design. With $n_I \geq 2$ and $N_{II} = 1$ the design is referred to as a *multi-period one-entrance*

design. The general case $n_I \geq 2$ and $n_{II} \geq 2$ is called a *multi-period multi-entrance* design. The relevant inclusion probabilities for these sampling designs are treated in Section 3.

2.3 Estimators

We focus on the problem of estimating the mean of a variable *y* for the population *U* in Definition 2.1,

$$\bar{y}_U = \frac{1}{N}\sum\nolimits_U y_k \tag{1}$$

where y_k is the fixed value of *y* for visitor $k \in U$ and $\sum_U y_k$ is abbreviation for $\sum_{k\in U} y_k$. (If D is a set of elements such that $D \subseteq U$ then Σ_D is written for $\Sigma_{k\in D}$.)

In order to estimate $\bar{y}_U$, the three-stage sampling procedure described in Section 2.2 is used to select a sample *s* of visitors to the site. For each visitor $k \in s$, the study variable value y_k is observed, as well as the individual's number of visits to the site during time period τ , z_k .

One possible estimator of $\bar{y}_U$ is the Horvitz-Thompson (HT) estimator,

$$\hat{\bar{y}}_{U\pi} = \frac{1}{N}\sum\nolimits_s \frac{y_k}{\pi_k} \tag{2}$$

In (2), π_k denotes the probability of visitor $k \in U$ to be included in the sample – the first-order inclusion probability. The main advantages of the HT estimator is that it is unbiased for $\bar{y}_U$ with respect to the sampling design and that it is fairly simple to calculate. The population size *N* is usually unknown when on-site sampling is used. This rules out the HT estimator for estimation of the population mean. On the other hand, the HT-estimator is applicable for the estimation of the population total.

The variance of the HT estimator is given by

$$V\left(\hat{\bar{y}}_{U\pi}\right) = \frac{1}{N^2}\sum\sum\nolimits_U \left(\pi_{kl} - \pi_k\pi_l\right)\frac{y_k}{\pi_k}\frac{y_l}{\pi_l} \tag{3}$$

where π_{kl} is the probability that visitors *k* and *l* are both included in the sample $\left(k,l \in U\right)$ – the second-order inclusion probability. (If *k*=*l*, then $\pi_{kl} = \pi_k$.) The double sum $\sum\sum_U$ in Equation (3) is an abbreviation for $\sum_{k\in U}\sum_{l\in U}$.

An alternative to the HT estimator is the Hájek (1971) estimator of $\bar{y}_U$,

$$\tilde{y}_s = \frac{\sum_s \frac{y_k}{\pi_k}}{\sum_s \frac{1}{\pi_k}} \tag{4}$$

This estimator usually has smaller variance than the HT estimator, especially when the sample size is variable. An additional advantage of the Hájek estimator is that it can be used even though the population size N is unknown. The Hájek estimator is not design-unbiased, but for large samples the bias is small. The approximate variance of the Hájek estimator, approximated by use of Taylor linearization, is given by

$$AV(\tilde{y}_s) = \frac{1}{N^2}\sum\sum_U (\pi_{kl} - \pi_k\pi_l)\left(\frac{y_k - \bar{y}_U}{\pi_k}\right)\left(\frac{y_l - \bar{y}_U}{\pi_l}\right) \tag{5}$$

(Särndal et al., 1992, Result 5.7.1).

3. Estimation under the multi-period multi-entrance design

For the application of the estimators presented in Section 2.3 to the three-stage sampling design, the first- and second-order sample inclusion probabilities must be known. The first-order inclusion probabilities are presented in Subsection 3.1; the second-order inclusion probabilities in Subsection 3.2.

3.1 First-order inclusion probabilities

Theorem 3.1: Under the multi-period multi-entrance design and Assumption 2.1, the first order inclusion probabilities are given by

$$\pi_k = \sum_{j=1}^{\min(z_k, n_I)} \left\{ p(j : n_I, z_k, N_I) \cdot \sum_{g=1}^{j} \binom{j}{g} \left(\frac{n_{II}}{N_{II}}\right)^g \left(\frac{N_{II} - n_{II}}{N_{II}}\right)^{j-g} \left[1 - (1-\alpha)^g\right] \right\} \tag{6}$$

where

$$p(j : n_I, z_k, N_I) = \frac{\binom{z_k}{j}\binom{N_I - z_k}{n_I - j}}{\binom{N_I}{n_I}}$$

The structure of the probability stated in the theorem is intuitive. Given that j out of z_k time periods, and g out of j entrances (the j entrances used in the j time periods) are selected in the first two stages, the probability of including the kth

individual is $1-(1-\alpha)^g$. The probability of selecting g out of j entrances in the second stage is $\binom{j}{g}(n_{II}/N_{II})^g((N_{II}-n_{II})/N_{II})^{j-g}$, a probability from the binomial $bin(j,\pi_{II})$ distribution with $\pi_{II}=n_{II}/N_{II}$. Finally, the first-stage probability of selecting j out of z_k time periods is given by the hypergeometric distribution. Thus, the probability of the intersection of the events: A="selecting j out of the z_k time periods when the kth individual visits the site" (the first stage), B="selecting g of the j entrances used by the kth individual at these j visits" (the second stage), and C="selecting the kth individual to the sample when sampling is made at g of the entrances used by the kth individual" (the third stage), is obtained as the product of these probabilities, i.e.

$$P(A\cap B\cap C)=p(j:n_I,z_k,N_I)\binom{j}{g}(n_{II}/N_{II})^g((N_{II}-n_{II})/N_{II})^{j-g}\left(1-(1-\alpha)^g\right)$$

The inclusion probability is now obtained by summing over all possible combinations (j,g) such that $g\le j\le \min(n_I,z_k)$. The case j=g=0 is excluded since then the individual can not be a part of the sample. A formal proof of Theorem 3.1 can be obtained from the authors.

3.2 Second-order inclusion probabilities

For the second-order inclusion probabilities, some new notation is required. First, let z_k^- denote the number of PSUs such that individual k visits the site but individual l does not (z_l^- is defined correspondingly). Let z_{kl}^s denote the number of PSUs such that both individuals visits the site and use the same entrance. Further, let z_{kl}^d be the number of PSUs such that both individuals visit the site but use different entrances. With this notation, $z_k=z_k^-+z_{kl}^s+z_{kl}^d$ and $z_l=z_l^-+z_{kl}^s+z_{kl}^d$.

In the first-stage sampling of time periods, let j_k^- and j_l^- denote the number of PSUs selected among the z_k^- and z_l^- time periods respectively. Similarly, let j_{kl}^s and j_{kl}^d time periods be selected from the z_{kl}^s and z_{kl}^d time periods, respectively. In the second-stage sampling, let g_k^- of the j_k^- entrances used by individual k be selected. Let g_l^- and g_{kl}^s be similar notations corresponding to j_l^- and j_{kl}^s, respectively. For the j_{kl}^d time periods let g_k^d be the number of entrances used by individual k which is selected to the second-stage sample. g_l^d is similarly

defined. Summing up, $g_k = g_k^- + g_k^d + g_{kl}^s$ is the number of times individual *k* enters the site through an entrance selected in the second stage, and $j_k = j_k^- + j_{kl}^s + j_{kl}^d$ is the number of time periods selected given that individual *k* visits the site. Thus, $g_k \le j_k \le z_k$.

Define the set $J = \left\{ j_k^-, j_l^-, j_{kl}^d, j_{kl}^s \middle| 1 \le j_k \le \min(z_k, n_I), 1 \le j_l \le \min(z_l, n_I) \right\}$ and let $\eta \in J$ index an element in *J*. Define the conditional set

$$G_\eta = \left\{ g_k^-, g_l^-, g_{kl}^s, g_k^d, g_l^d \middle| g_k \ge 1, g_l \ge 1, g_k^- \le j_k^-, g_l^- \le j_l^-, g_{kl}^s \le j_{kl}^s, g_k^d \le j_{kl}^d, g_l^d \le j_{kl}^d, \eta \right\}$$

The set *J* defines possible combinations of $(j_k^-, j_l^-, j_{kl}^d, j_{kl}^s)$ such that the selected first-stage sample includes at least one period when individual *k* visits the site and at least one period when individual *l* visits the site. The set G_η gives possible combinations of $(g_k^-, g_l^-, g_{kl}^s, g_k^d, g_l^d)$, given $\eta \in J$, such that both individuals pass through at least one entrance each selected in the second-stage sample.

Theorem 3.2: Under the multi-period multi-entrance design and Assumptions 2.1-2.3, the second-order inclusion probabilities for $k \ne l$ are given by

$$\pi_{kl} = \sum_J P(j_k^-, j_l^-, j_{kl}^d, j_{kl}^s) \sum_{G_\eta} P(g_k^-, g_l^-, g_{kl}^s, g_k^d, g_l^d \middle| j_k^-, j_l^-, j_{kl}^d, j_{kl}^s) \left(1 - \beta^{g_k} - \beta^{g_l} + \beta^{g_k + g_l}\right)$$

where

$$P(j_k^-, j_l^-, j_{kl}^d, j_{kl}^s) = \frac{\binom{z_k^-}{j_k^-}\binom{z_l^-}{j_l^-}\binom{z_{kl}^d}{j_{kl}^d}\binom{z_{kl}^s}{j_{kl}^s}\binom{N_I - z_k^- - z_l^- - z_{kl}^d - z_{kl}^s}{n_I - j_k^- - j_l^- - j_{kl}^d - j_{kl}^s}}{\binom{N_I}{n_I}},$$

$$\begin{aligned} P(g_k^-, g_l^-, g_{kl}^s, g_k^d, g_l^d \middle| j_k^-, j_l^-, j_{kl}^d, j_{kl}^s) = {} & \binom{j_k^-}{g_k^-}\left(\frac{n_{II}}{N_{II}}\right)^{g_k^-}\left(1 - \frac{n_{II}}{N_{II}}\right)^{j_k^- - g_k^-} \\ & \times \binom{j_l^-}{g_l^-}\left(\frac{n_{II}}{N_{II}}\right)^{g_l^-}\left(1 - \frac{n_{II}}{N_{II}}\right)^{j_l^- - g_l^-} \\ & \times \binom{j_{kl}^s}{g_{kl}^s}\left(\frac{n_{II}}{N_{II}}\right)^{g_{kl}^s}\left(1 - \frac{n_{II}}{N_{II}}\right)^{j_{kl}^s - g_{kl}^s} \\ & \times \binom{j_{kl}^d}{g_k^d}\left(\frac{n_{II}}{N_{II}}\right)^{g_k^d}\left(1 - \frac{n_{II}}{N_{II}}\right)^{j_{kl}^d - g_k^d} \times P(g_l^d \middle| g_k^d, j_{kl}^d) \end{aligned}$$

with

$$P(g_l^d \,|\, g_k^d, j_{kl}^d) = \sum_{\substack{g_{1l}^d, g_{2l}^d \\ g_{1l}^d + g_{2l}^d = g_l^d}} \binom{j_{kl}^d - g_k^d}{g_{1l}^d} \left(\frac{n_{II}}{N_{II}-1}\right)^{g_{1l}^d} \left(1 - \frac{n_{II}}{N_{II}-1}\right)^{j_{kl}^d - g_l^d - g_{1l}^d} \times \binom{g_k^d}{g_{2l}^d} \left(\frac{n_{II}-1}{N_{II}-1}\right)^{g_{2l}^d} \left(1 - \frac{n_{II}-1}{N_{II}-1}\right)^{g_k^d - g_{2l}^d}$$

and $\beta = (1-\alpha)$.

The structure of the second-order inclusion probability is similar to the one of the first-order inclusion probability. For instance, given that sampling is made at g_k of the entrances used by the *k*th individual and at g_l of the entrances used by the *l*th individual, the probability of including both individuals in the sample is $(1-\beta^{g_k} - \beta^{g_l} + \beta^{g_k+g_l})$. A formal proof of the theorem can be obtained from the authors.

4. Special cases

4.1 Multi-period one-entrance design

A special case of the three-stage design in Section 2.2 arises when there is only one entrance to the site. Then, the second stage of the three-stage design is omitted. Or, put differently, $N_{II} = 1$ and $n_{II} = 1$ whereby the entrance is selected with probability 1. It follows that *g=j* in Equation (6) and the following result is obtained:

Lemma 4.1: Under the multi-period one-entrance sampling design, the first-order inclusion probabilities are given by

$$\pi_k = \sum_{j=1}^{\min(z_k, n_I)} \left\{ p(j : n_I, z_k, N_I) \cdot \left[1 - (1-\alpha)^j\right] \right\}$$

Regarding the second-order inclusion probabilities, we note that $z_{kl}^d = j_{kl}^d = 0$, since individuals *k* and *l* can not use different entrances. Also $g_k^- = j_k^-$, $g_l^- = j_l^-$, and $g_{kl}^s = j_{kl}^s$. Introducing these restrictions into the formula in Theorem 3.2 yields the result stated in Lemma 4.2.

Lemma 4.2: Under the multi-period one-entrance sampling design, the second-order inclusion probabilities for $k \neq l$ are given by

$$\pi_{kl} = \sum_J P(j_k^-, j_l^-, 0, j_{kl}^s)\left(1 - \beta^{j_k} - \beta^{j_l} + \beta^{j_k + j_l}\right)$$

4.2 One-period one-entrance design

The simplest special case considered here is the design where only one of the N_I time periods is selected and there is only one entrance. This design may seem unrealistic (cf e.g. Sudman, 1980), but the results presented here have some implications for the general literature on on-site sampling. Lemma 4.3 states the first and second order inclusion probabilities obtained for this sampling design.

Lemma 4.3: Under the one-period one-entrance sampling design, the first- and second- order inclusion probabilities are given by

$$\pi_k = cz_k \text{ and } \pi_{kl} = N_I c^2 z_{kl}^s$$

where $c = \alpha N_I^{-1}$ and $k \neq l$.

Lemma 4.3 is obtained from lemmas 4.1 and 4.2 by setting $j = n_I = 1$, $j_k^- = j_l^- = 0$, and $j_{kl}^s = j_k = j_l = 1$.

An important conclusion from Lemma 4.3 is that the one-period design is not measurable: the second order inclusion probability π_{kl} equals zero if $z_{kl}^s = 0$. Thus, unbiased variance estimators based on sample information are generally not available (see Särndal et al., 1992, Remark 2.4.3).

In the special case covered by Lemma 4.3, the individual's first-order inclusion probability is proportional to his or her number of visits. Thus, this is an example of a situation where the often advocated assumption of proportionality of sample inclusion probabilities to frequency formally holds.

5. Simulation

5.1 Simulation design

The simulation is based on data from three annual surveys (1999, 2000 and 2001) of anglers visiting the Kaitum river in northern Sweden (Paulrud and Laitila, 2004). We use the merged data as a pseudo-population *U* of size *N*=2,391. For each individual in *U*, we know the total number of days that he or she was engaged in fishing during the year of observation. This variable, the number of fishing days, is our study variable *y* in the simulation. Our pseudo-population data set contains information on the number of times that each individual visited the river during the specified fishing season (ten consecutive summer weeks). It lacks however information on exactly when the visits were made. For each individual in

U, we therefore allocate his or her declared number of visits over the ten weeks. The allocation is made randomly without replacement (that is, no individual is allowed to visit the river more than once during a given week).

In each simulation round, a sample of visitors is drawn from *U*. In practice, interviewers walk along the river side and sample those anglers that are intercepted at site. This is treated as a single-entrance situation. The sampling design studied thus corresponds to a two-stage sampling design with weeks as PSUs and anglers as SSUs. Two different sample sizes for Stage 1 are considered; , (the one-period design) and $n_I = 2$ (the two-period design). Three different inclusion probabilities for the BE sampling in Stage 2 are tried in the simulations; $\alpha = .2$, .5 and .8. Let the sample selected in iteration *g*, *g*=1,…,*G*, be denoted $s_{(g)}$ (of random size $n_{(g)}$). For each choice of sample size in Stage 1 and inclusion probability in Stage 2, we make *G*=1,000 iterations.

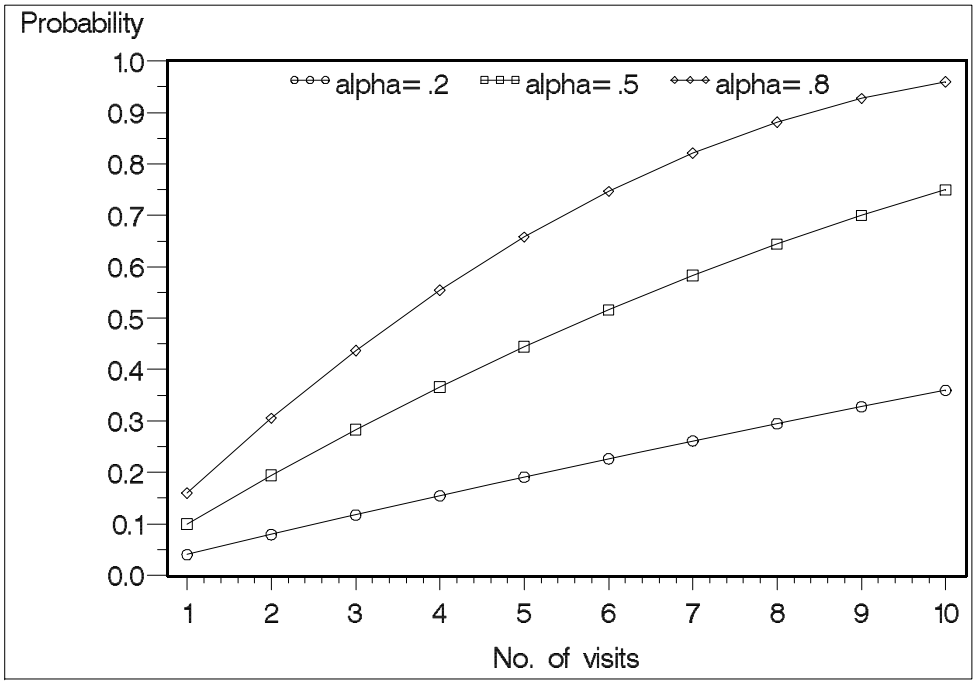
In the estimation, the first-order inclusion probabilities of selected anglers need to be calculated. For the one-period design, inclusion probabilities are given by Lemma 4.3.

For $n_I = 2$ (and $N_I \geq 2$) the formula for π_k in Lemma 4.1 simplifies to

$$\pi_k = \begin{cases} \dfrac{2\alpha}{N_I} & \text{if } z_k = 1 \\ \dfrac{2\alpha z_k}{N_I(N_I - 1)}\left[(N_I - z_k) + (z_k - 1)\left(1 - \dfrac{\alpha}{2}\right)\right] & \text{if } z_k \geq 2 \end{cases} \tag{7}$$

Figure 5.1 shows inclusion probabilities calculated according to Equation (7) with $\alpha = .2$, .5 and .8. The inclusion probabilities are plotted against the angler's number of visits to the site during the ten week season. We see that a visitor's first-order inclusion probability is approximately proportional to his or her number of visits if $\alpha = .2$ but less so for larger values on α.

Figure 5.1. First-order inclusion probabilities vs. number of visits for the two-period single-entrance design.



As a complement to the simulation, the variance is also computed for the HT estimator under a standard reference design: direct sampling by SRS from the population frame. The variance is calculated for the expected sample sizes relevant under the one-period and two-period designs and different values on α. This is accomplished by using $n = \sum_U \pi_k$ for each case.

5.2 Results

The simulation output for the one- and two-period design is presented in Table 5.1 where the bias estimates of various estimators of the mean are presented. Note that the HT estimator is unbiased and the observed bias is only due to the variability of the simulation results. Estimates of the variances of the bias estimates can be derived from the MSE values presented. Using separate t-tests the biases for the HT and Hájek estimators in the table are all insignificant. The bias estimates for the sample mean is significant, however. In relation to the true mean, the bias is not very large: the largest observed bias (for $n_1 = 1,\ \alpha = .8$ and $\bar{y}_s$**)** only represents 1.8 % of the true mean.

Among the estimators the HT estimator has the largest variance and the sample mean the smallest, and the same is true for their MSEs. Consequently, in terms of MSE, the sample mean is superior to the others. However, this advantage steadily decreases as the (expected) sample size increases. This makes sense since the variance is reduced by increasing sample size, whereas the bias is not, and the sample mean has the largest bias. For larger sample sizes than those in use here, the Hájek estimator is probably preferable.

Table 5.1. Bias and MSE (in parentheses) estimates of estimators of the population mean ($\bar{y}_U = 25.79$).

		Estimator			
n_I	α	$\hat{\bar{y}}_{U\pi}$	$\tilde{\bar{y}}_s$	$\bar{y}_s$	$\hat{\bar{y}}_{U\pi,\mathrm{dir}}$
1	.2	.064	.062	-.438	0
		(18.1)	(8.46)	(3.38)	(4.50)
	.5	-.060	-.011	-.464	0
		(7.21)	(3.37)	(1.32)	(1.69)
	.8	.017	-.033	-.468	0
		(4.52)	(2.10)	(0.86)	(0.98)
2	.2	-.072	.006	-.414	0
		(9.64)	(4.04)	(1.80)	(2.27)
	.5	-.032	.002	-.435	0
		(3.70)	(1.52)	(0.80)	(0.87)
	.8	.001	.013	-.446	0
		(1.87)	(0.82)	(0.55)	(0.52)

The estimator $\hat{\bar{y}}_{U\pi,\mathrm{dir}}$ in Table 5.1 is the HT estimator under direct SRS from the population frame. The estimator corresponds to the sample mean and is unbiased whereby the bias 0 is inserted in the table. Sample sizes equal expected sample sizes for the corresponding on-site samples. The expected sample sizes for the one-period designs are 96, 240 and 384 for α equal to 0.2, 0.5 and 0.8, respectively. For the two-period designs the corresponding expected sample sizes are 183, 427, and 632, respectively.

According to the results in Table 5.1, the variance of the HT estimator under direct sampling is smaller than the MSEs of the HT estimator and the Hájek estimator under the on-site. The larger the sample size, the more advantageous is the HT estimator under direct sampling, since it is both unbiased and has quite small variance. For our maximum sample size, $n = 632$, it already outperforms all the others.

6. Discussion

This paper contributes with inclusion probabilities for a three-stage sampling design suggested for on-site sampling. The expressions obtained are complex but feasible to use, especially if the numbers of time periods and entrances are small. The inclusion probabilities were derived using the conditioning principle. As suggested by an anonymous reviewer, the inclusion – exclusion principle might provide a simpler approach.

The expressions show that the often assumed proportionality of inclusion probabilities to frequency of visits is not generally true. Within the three-stage design, proportionality is obtained only for a very simple case, the one-period one-entrance design. On the other hand, our numerical illustration shows that proportionality might serve as a good approximation. This problem deserves further studies as proportionality yields more simple expressions.

The simulation illustrates that if dependence of the inclusion probabilities on the frequencies of visits is not accounted for, as in the case of the sample mean, bias results. This result was expected and can be shown formally. In the present case the bias was small, probably due to a low correlation between the study variable and the number of visits, and the MSE of the uncorrected estimator was generally smaller than those of the corrected estimators. Variables with stronger relations to the frequency of visits are expected to be associated with estimates with larger biases and MSEs. However, the relatively large variances of the corrected estimators illustrate a loss in precision when applying on-site sampling.

For the application of the three-stage design and calculation of the inclusion probabilities, information on all visits made by an individual is needed. At the time of the interview at the site, future visits by the respondent are typically not known. Thus, information on the total number of visits needs to be collected after the time interval defining the study population. This can be made in several ways. Paulrud and Laitila (2004) utilize recorded addresses for a mail questionnaire study. Dillman, Dolsen and Machlis (1995) provide guidelines for studies based on hand-out of questionnaires at interception of respondents at site. These and other examples show that collection of information on visits after the study period is feasible. Consequently, the approach for on-site sampling and estimation treated in this paper can be used in practice.

Care has to be taken in the application and design of an on-site sampling design. The number of visits made to the site during the time interval must be modest. Examples where the design can be used are surveys of visitors to recreational areas like hiking parks and angling sites. Other examples include surveys of visitors to zoos and museums. It is also possible to apply the design to more frequently visited sites, like shopping centers and airports, if the time interval defining the population is short. A problem for future research is to combine a series of on-site studies in order to cover a longer time interval.

The problems involved with on-site sampling resemble those with indirect sampling and it is of interest to study the relationships between these methods more closely. One interesting topic is how the inclusion probabilities relate to the weights defined by the generalized weight share method. Another topic is the efficiency of the generalized weight share estimators relative to estimators based on direct calculation of inclusion probabilities. Note, however, that the indirect sampling method does not circumvent the problem of collecting information from respondents after the studied time interval.

Acknowledgement

The authors acknowledge the helpful comments of Carl-Erik Särndal and anonymous reviewers.

REFERENCES

BLOCK, L.G, MORWITZ, V.G, PUTSIS JR, W.P. and S.K. SEN (2002). Assessing the Impact of Antidrug Advertising on Adolescent Drug Consumption, American Journal of Public Health, 92, 1346—1351.

BRENNER, M.E. (1998). Meaning and Money, Educational Studies in Mathematics, 36, 123—155.

COX, D. (1969). Some Sampling Problems in Technology, in Johnson U. L. and H. Smith, (eds.), New Developments in Survey Sampling, Wiley Interscience, New York.

DILLMAN, D.A, DOLSEN, D.E. and G.E. MACHLIS (1995). Increasing Response to Personally-Delivered Mail-Back Questionnaires, Journal of Official Statistics, 11, 129—139.

HÁJEK, J, (1971). Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One", in Godambe, V.P. and D.A. Sprott (eds.) Foundations of Statistical Inference, Holt, Rinehart and Winston, Toronto.

KALTON, G. (1991). Sampling Flows of Mobile Human Populations, Survey Methodology, 17:2, 183—194.

KEEN, C, WETZELS, M, DE RUYTER, K. and R. FEINBERG (2004). E-tailors versus Retailers – Which Factor Determines Consumer Preferences, Journal of Business Research, 57, 685—695.

KEILLOR, B.D. D'AMICO, M. and V. HORTON (2001). Global Consumer Tendencies, Psychology and Marketing, 18, 1—19.

LAITILA, T. (1998). Estimation of Combined Site-Choice and Trip-Frequency Models of Recreational Demand using Choice-based and On-Site Samples, Economics Letters, 64, 17—23.

LAVALLÉE, P. (2007). *Indirect Sampling*, Springer Science+Business Media, New York.

NOWELL, C. and L.R. STANLEY (1991). Length-Biased Sampling in Mall Intercept Surveys, Journal of Marketing Research 28, 475—479.

PAULRUD, A. and T. LAITILA (2004). Valuation of Management Policies for Sport Fishing on Sweden's Kaitum River, Journal of Environmental Planning and Management, 42, 863—879.

POLLOCK, K.H, JONES, C.M. and T.L. BROWN (1994). Angler Survey Methods and Their Applications in Fisheries Management, American Fisheries Society Special Publication 25, American Fisheries Society, Bethseda.

SHAW, D. (1988). On-Site Samples' Regression, Problems of Non-negative Integers, Truncation, and Endogenous Stratification, Journal of Econometrics, 37, 211—223.

SUDMAN, S. (1980). Improving the Quality of Shopping Center Sampling. Journal of Marketing Research, 17, 423—431.

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

THOMAS LAITILA, Statistics Sweden, SE-701 89 Örebro, Sweden.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 475—486

ON THE ESTIMATION OF VARIANCE OF CALIBRATED ESTIMATORS OF THE POPULATION COVARIANCE

Aleksandras Plikusas, Dalius Pumputis

ABSTRACT

The problem of variance estimation of the calibrated estimators of the finite population covariance is considered. The calibrated estimators of the population covariance are more efficient compared to the straight estimators provided the auxiliary variables are well correlated with the study variables. In the case of low correlated auxiliaries, all estimators are of a similar quality. There are many ways to define a calibrated estimator of the covariance. We consider three types of estimators having one weighting system. The estimators with two or three weighting systems can be also defined. Due to the complexity of calibrated weights, the calculation of variance of respective estimators is complicated. In the case of a nonlinear calibration, calibrated weights are expressed by recurrent equations, and an explicit analytical expression for the variance is not available. The variance estimators for such a case are derived using a rough Taylor expansion or some pseudo linearization. The estimators derived are compared by simulation with empirical variance and the Jackknife variance estimator.

Key words: finite-population covariance; auxiliary variables; calibration; nonlinear calibration.

1. Introduction

Calibrated estimators of the finite population total are widely used in statistics to improve the quality of estimators, using auxiliary information. Estimators of this type are mostly used in official statistics, especially in social surveys. The idea of calibration technique for estimating the population totals was presented in (Deville and Särndal, 1992).

The estimation of more complicated parameters, using auxiliary variables, is not widely studied in the literature. Calibrated estimators of the covariance that are defined in a slightly different way are studied in (Sitter and Wu, 2002), and the importance of extending the calibration approach for the estimation of more

complicated parameters is also declared. One type of a calibrated estimator of the ratio of two totals was considered by (Krapavickaitė and Plikusas, 2005). Calibrated estimators of quantiles are studied in (Harms and Duchesne, 2006).

The more efficient estimators of the finite population covariance may be used, for example, for the estimation of the regression coefficient. The calibrated estimators of the covariance, considered in this paper, were introduced in (Plikusas and Pumputis, 2007). They are constructed using different calibration equations and different loss functions. In some cases an explicit solution of calibration equations does not exist. Iterative formulas are used to calculate calibrated weights, and estimation of the variance of calibrated estimators is complicated.

In this paper, some approximate estimators of the variance of the calibrated estimators of the population covariance are examined. The simulation study for the skewed population is presented.

2. Calibrated estimators of the covariance

Consider a finite population $U = \{u_1, u_2, \ldots, u_N\}$ of N elements, where the unit u_k has unknown study variables' values $\{ y_k, z_k \}$ for each k.

Denote the finite population covariance of the study variables y and z by

$$C(y,z) = \frac{1}{N-1}\sum_{k=1}^{N}(y_k - \mu_y)(z_k - \mu_z),$$

where $\mu_y = \frac{1}{N}\sum_{k=1}^{N} y_k$, $\mu_z = \frac{1}{N}\sum_{k=1}^{N} z_k$.

In this section, we present three types of calibrated estimators of the covariance proposed in (Plikusas and Pumputis, 2007).

Denote by s, $s \subset U$, the probability sample set drawn from the population U by the design with the first order inclusion probabilities π_k, and second order inclusion probabilities π_{kl}. The sample design weight of the element k is denoted by $d_k = 1/\pi_k$.

Some straight estimators of covariance can be found, for example, in the book (Särndal, Swensson and Wretman, 1992, p.186). Two of them are

$$\hat{C}_1(y,z)=\frac{1}{N-1}\sum_{k\in s}d_k\left(y_k-\frac{1}{N}\sum_{l\in s}d_l y_l\right)\left(z_k-\frac{1}{N}\sum_{l\in s}d_l z_l\right)$$

and

$$\hat{C}_2(y,z)=\frac{1}{N}\sum_{k\in s}\frac{y_k z_k}{\pi_k}-\frac{1}{N(N-1)}\sum_{k,l\in s,\ k\neq l}\frac{y_k z_l}{\pi_{kl}}. \tag{1}$$

We will use these estimators for comparison with the calibrated estimators defined in the paper below.

Suppose, that for the unit u_i, the known auxiliary variables' values $\{a_i, b_i\}$, for each *i*, are available. Denote their known covariance by $C(a,b)$.

In general, the calibrated estimator is defined by the calibrated weights w_k, $k\in s$, that are used instead of the design weights d_k, and satisfy the following conditions:

a) the weights w_k of the calibrated estimator satisfy some calibration equation;
b) the distance between the weights d_k and calibrated weights w_k is minimal according to some loss function *L*.

Let us define three calibrated estimators $\hat{C}_w(y,z)$ of the covariance of the following shape:

$$\hat{C}_w(y,z)=\frac{1}{N-1}\sum_{k\in s}w_k\left(y_k-\hat{\mu}_{yw}\right)\left(z_k-\hat{\mu}_{zw}\right), \tag{2}$$

where $\hat{\mu}_{yw}=\frac{1}{N}\sum_{k\in s}w_k y_k$, $\hat{\mu}_{zw}=\frac{1}{N}\sum_{k\in s}w_k z_k$, and weights w_k we define below in a different ways using three different calibration equations as condition a), and some loss functions for condition b).

I) Nonlinear calibration. Let us consider the calibration equation

$$\frac{1}{N-1}\sum_{k\in s}w_k\left(a_k-\hat{\mu}_{aw}\right)\left(b_k-\hat{\mu}_{bw}\right)=C(a,b), \tag{3}$$

$$\hat{\mu}_{aw}=\frac{1}{N}\sum_{k\in s}w_k a_k,\ \hat{\mu}_{bw}=\frac{1}{N}\sum_{k\in s}w_k b_k.$$

We call this case a nonlinear calibration because the calibration equation (3) is nonlinear with respect to the calibrated weights w_k.

II) Linear calibration. Define the second calibration equation as follows

$$\begin{aligned} &\frac{1}{N-1}\sum_{k\in s} w_k (a_k - \mu_a)(b_k - \mu_b) = C(a,b), \\ &\mu_a = \frac{1}{N}\sum_{k=1}^{N} a_k,\ \mu_b = \frac{1}{N}\sum_{k=1}^{N} b_k. \end{aligned} \tag{4}$$

We refer to this case as a linear calibration, because we are actually calibrating the total of the variable $(a-\mu_a)(b-\mu_b)$.

III) Calibration of totals. The system of calibrated weights w_k of the third type is defined by calibration of totals of the auxiliary variables, so the calibration equation is

$$\left(\sum_{k\in s} w_k a_k, \sum_{k\in s} w_k b_k\right) = \left(\sum_{k=1}^{N} a_k, \sum_{k=1}^{N} b_k\right). \tag{5}$$

We present below a list of loss functions that we used to define the calibrated weights:

$$L_1 = \sum_{k\in s} \frac{(w_k - d_k)^2}{d_k q_k}, \qquad L_3 = \sum_{k\in s} 2\frac{(\sqrt{w_k} - \sqrt{d_k})^2}{q_k},$$

$$L_6 = \sum_{k\in s} \frac{1}{q_k}\left(\frac{w_k}{d_k} - 1\right)^2, \qquad L_7 = \sum_{k\in s} \frac{1}{q_k}\left(\frac{\sqrt{w_k}}{\sqrt{d_k}} - 1\right)^2.$$

The functions L_1, L_3 and three other loss functions are mentioned in (Deville and Särndal, 1992). The loss functions L_6 and L_7 are introduced in (Plikusas, 2003). The free additional weights q_k can be used to extend the class of estimators. In our simulation we put $q_k = 1$ for all *k*. Simulation results show that the difference in efficiency of the calibrated estimators, depending on the loss function used, is minor. The impact of the calibration equation used is more significant.

Let us formulate two propositions that define four different systems of calibrated weights using loss functions L_1, L_3, L_6, L_7.

Proposition 1. The weights $w_k = w_k^{(i)}$, $k \in s$, $i = 1,3,6,7$, which satisfy (3) and minimize the loss function L_i, satisfy the equation $w_k^{(i)} = d_k g_k^{(i)}$. Here

$$g_k^{(1)} = 1 + \lambda^{(1)} q_k e_k, \quad g_k^{(3)} = \left(\frac{1}{2} \lambda^{(3)} q_k e_k - 1 \right)^{-2},$$

$$g_k^{(6)} = 1 + \lambda^{(6)} d_k q_k e_k, \quad g_k^{(7)} = \left(\lambda^{(7)} d_k q_k e_k - 1 \right)^{-2},$$

$$\lambda^{(1)} = \hat{A} \left(\sum_{k \in s} d_k q_k e_k a_k b_k \right)^{-1},$$

$$\hat{A} = (N-1) C(a,b) + N \left(2 - \frac{\hat{N}_w}{N} \right) \hat{\mu}_{aw} \hat{\mu}_{bw} - \sum_{k \in s} d_k a_k b_k, \ \hat{N}_w = \sum_{k \in s} w_k,$$

$$e_k = (a_k - \hat{\mu}_{aw})(b_k - \hat{\mu}_{bw}) - \left(1 - \frac{\hat{N}_w}{N} \right) \left(\frac{\hat{\mu}_{aw}}{a_k} + \frac{\hat{\mu}_{aw}}{a_k} \right) a_k b_k;$$

$\lambda^{(3)}$ is a properly chosen root of the equation $\alpha_2 \lambda^2 + \alpha_1 \lambda + \alpha_0 = 0$ with

$$\alpha_0 = \hat{A}, \ \alpha_1 = -\sum_{k \in s} q_k w_k a_k b_k e_k, \ \alpha_2 = \frac{1}{4} \sum_{k \in s} q_k^2 w_k a_k b_k e_k^2, \ \lambda^{(6)} = \hat{A} \left(\sum_{k \in s} d_k^2 q_k a_k b_k e_k \right)^{-1};$$

$\lambda^{(7)}$ is a properly chosen root of the equation $\beta_2 \lambda^2 + \beta_1 \lambda + \beta_0 = 0$ with

$$\beta_0 = \hat{A}, \ \beta_1 = -2 \sum_{k \in s} d_k q_k w_k a_k b_k e_k, \ \beta_2 = \frac{1}{4} \sum_{k \in s} d_k^2 q_k^2 w_k a_k b_k e_k^2.$$

Proposition 2. The weights $w_k = w_k^{(i)}$, $k \in s$, $i = 1,3,6,7$, which satisfy (4) and minimize the loss function L_i, satisfy the equation $w_k^{(i)} = d_k f_k^{(i)}$. Here

$$f_k^{(1)} = 1 + (t_c - \hat{t}_c) \left(\sum_{l \in s} d_l q_l c_l^2 \right)^{-1} q_k c_k,$$

$$f_k^{(3)} = 4 \left(2 - (N-1) C(a,b) \left(\sum_{l \in s} \frac{q_l w_l^{3/2} c_l^2}{2\sqrt{w_l} - \sqrt{d_l}} \right)^{-1} q_k c_k \right)^{-2},$$

$$f_k^{(6)} = 1 + (t_c - \hat{t}_c)\left(\sum_{l\in s} d_l^2 q_l c_l^2\right)^{-1} d_k q_k c_k,$$

$$f_k^{(7)} = \left(1 - (N-1)C(a,b)\left(\sum_{l\in s} d_l q_l w_l^{3/2} c_l^2\right)^{-1} d_k q_k c_k\right)^{-2},$$

$$c_k = (a_k - \mu_a)(b_k - \mu_b),\ t_c = \sum_{k=1}^{N} c_k,\ \hat{t}_c = \sum_{k\in s} d_k c_k.$$

The proofs of Propositions 1 and 2 are given in (Plikusas and Pumputis, 2007).

3. Variance estimation

There are several approaches for estimating variance of complex estimators. The two essential techniques are the analytical approach (e.g., linearization technique) and replication (sub-sampling) methods. Employing the linearization technique, a linear approximation of the estimator is used to derive an expression for the approximate variance. The estimator of the variance is derived from this expression of the approximate variance.

Let us construct some estimators of the variance of the considered calibrated estimators. We have examined two rough variance estimators of the calibrated estimators of the population covariance. The same simplified estimators are constructed for all types of calibrated estimators.

The jackknife and several bootstrap methods can be mentioned as sub-sampling methods that can be used for the variance estimation as well.

In our case, the analytical approach faces some difficulties: we have no explicit expression for the random calibrated weights w_k in the case of nonlinear calibration and estimators are more complicated as compared to that of the totals.

Below we present two variance estimators constructed by a rough linearization.

The approximate linearized estimator. The first proposed estimator of the variance of the calibrated estimators of covariance is of the form

$$\hat{V}\left(\hat{C}_w(y,z)\right) = \left(\frac{1}{N-1}\right)^2 \sum_{k,l\in s}\left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right)\frac{\hat{e}_k}{\pi_k}\frac{\hat{e}_l}{\pi_l}, \qquad (6)$$

here

$$\hat{e}_k = \frac{w_k}{d_k}\left((y_k - \hat{\mu}_{yw})(z_k - \hat{\mu}_{zw}) + \left(\frac{\hat{N}_w}{N} - 1 \right)(y_k \hat{\mu}_{zw} + z_k \hat{\mu}_{yw}) \right),$$

$$\hat{N}_w = \sum_{k \in s} w_k \, .$$

The motivation of this estimator is as follows. The estimator $\hat{C}_w(y,z)$, defined in (2), can be expressed in the following form

$$\hat{C}_w(y,z) = \frac{1}{N-1}\left(\hat{t}_{yzw} - \frac{2}{N}\hat{t}_{yw}\hat{t}_{zw} + \frac{1}{N^2}\hat{t}_{yw}\hat{t}_{zw}\hat{N}_w \right),$$

here

$$\hat{t}_{yzw} = \sum_{k \in s} w_k y_k z_k, \quad \hat{t}_{yw} = \sum_{k \in s} w_k y_k, \quad \hat{t}_{zw} = \sum_{k \in s} w_k z_k .$$

Denote

$$t_{yzw} = \sum_{k=1}^{N} \frac{w_k}{d_k} y_k z_k, \quad t_{yw} = \sum_{k=1}^{N} \frac{w_k}{d_k} y_k,$$

$$t_{zw} = \sum_{k=1}^{N} \frac{w_k}{d_k} z_k, \qquad N_w = \sum_{k=1}^{N} \frac{w_k}{d_k}.$$

Let us take the linear part of the Taylor series expansion of $\hat{C}_w(y,z)$ at the point $\left(\hat{t}_{yzw}, \hat{t}_{yw}, \hat{t}_{zw}, \hat{N}_w\right) = \left(t_{yzw}, t_{yw}, t_{zw}, N_w\right)$:

$$\hat{C}_{wl}(y,z) = \frac{1}{N-1}\left(\hat{t}_{yzw} + \frac{t_{zw}}{N}\left(\frac{N_w}{N} - 2 \right)\hat{t}_{yw} + \frac{t_{yw}}{N}\left(\frac{N_w}{N} - 2 \right)\hat{t}_{zw} + \right.$$

$$\left. + \frac{t_{yw}t_{zw}}{N^2}\hat{N}_w + \frac{2t_{yw}t_{zw}}{N}\left(1 - \frac{N_w}{N} \right) \right).$$

The variance of the linearized estimator is equal to

$$V\left(\hat{C}_{wl}(y,z)\right) = \left(\frac{1}{N-1} \right)^2 V\left(\sum_{k \in s} d_k e_k \right),$$

where

$$e_k = \frac{w_k}{d_k}\left(y_k z_k + \frac{t_{zw}}{N}\left(\frac{N_w}{N} - 2\right) y_k + \frac{t_{yw}}{N}\left(\frac{N_w}{N} - 2\right) z_k + \frac{1}{N^2} t_{yw} t_{zw} \right).$$

Assuming the calibrated weights w_k are non-random and using Result 2.8.1 from (Särndal, Swensson and Wretman, 1992), we get the expression of the variance

$$V\left(\hat{C}_{wl}(y,z)\right) = \left(\frac{1}{N-1}\right)^2 \sum_{k,l=1}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{e_k}{\pi_k} \frac{e_l}{\pi_l}.$$

As the estimator of variance, we take the expression

$$\hat{V}\left(\hat{C}_{wl}(y,z)\right) = \left(\frac{1}{N-1}\right)^2 \sum_{k,l \in s} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{\hat{e}_k}{\pi_k} \frac{\hat{e}_l}{\pi_l}.$$

The values $\hat{e}_k$ are defined by replacing unknown parameters t_{yw}, t_{zw} and N_w with their estimates $\hat{t}_{yw}$, $\hat{t}_{zw}$, $\hat{N}_w$.

Using different weight systems from Proposition 1 and Proposition 2, we obtain the estimators of variance for each case (nonlinear, linear, calibration of totals) and each loss function (L_1, L_3, L_6, L_7).

Let us present another variance estimator, which we call a *pseudo-linearization* estimator. The covariance estimator (2) may be written in the following form:

$$\hat{C}_w(y,z) = \sum_{k \in s} d_k e_k^{(p)},$$

where d_k are sample design weights,

$$e_k^{(p)} = \frac{w_k}{(N-1)d_k}\left(y_k - \hat{\mu}_{yw}\right)\left(z_k - \hat{\mu}_{zw}\right).$$

Using a similar motivation as for estimator (6), we can get the variance estimator

$$\hat{V}_P\left(\hat{C}_w(y,z)\right) = \sum_{k,l \in s} \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) \frac{e_k^{(p)}}{\pi_k} \frac{e_l^{(p)}}{\pi_l}. \qquad (7)$$

The expression of the variance estimator (6) is much more complicated than expression (7). It is difficult to choose a preference between the two estimators proposed.

Replication methods such as the jackknife, bootstrap and balanced half-samples may be also used for the estimation of variances of complex estimators. In the following section we will compare by simulation both variance estimators with the jackknife variance estimator and empirical variance.

4. Simulation study

The real population of size 300 from the Lithuanian Enterprise Survey having a skewed distribution, which is close to exponential, is taken for the simulation. The population is stratified into two strata by the size of the survey variable *y*. We denote the stratum size by $N_h, h=1,2, N=N_1+N_2$. The stratified simple random sample is used as a sample design. The sample size $n=n_1+n_2=100$ is distributed to strata, using Neyman's optimal allocation. So, the inclusion probability of the element *k* into the sample is $\pi_k=n_h/N_h$, if the element *k* belongs to the stratum *h*. It is assumed in the simulation that the additional weight $q_k=1$ for all *k*.

We draw $m=500$ independent samples s_j. The empirical variance, and the mean values of the linearized variance estimator (6), pseudo-linearized variance estimator (7) and the jackknife variance estimator are calculated. To show the advantages using the auxiliary variables, we present empirical and jackknife variance for the standard estimators of the population covariance (1). Three different collections of the auxiliary variables *a* and *b* are used. In the first case, both auxiliary variables *a* and *b* are highly correlated with the study variables *y* and *z*; in the second one, there is one correlated auxiliary; in the third case, no highly correlated auxiliaries. The simulation results depending on the different auxiliaries, having different correlation ρ with the study variables are presented in Table 1.

The calibrated estimators of the covariance are denoted by $\hat{C}_{wi}^{(non)}(y,z)$, $\hat{C}_{wi}^{(lin)}(y,z)$, $\hat{C}_{wi}^{(tot)}(y,z)$, depending on the calibration equation used: (3), (4), (5). Here the index *i* refers to the number *i* of the loss function L_i that is used to define the estimator.

The calibrated estimators that are derived under the same calibration equation are very similar despite the loss function used. This is the reason why we provide the numerical results only for the loss functions L_1 and L_6.

Our simulation shows that estimators constructed using the linear calibration in the case of a high correlation between the auxiliary and study variables, have the smallest variance.

The estimators constructed using the calibration of the totals, are of the lowest efficiency in the case of highly correlated auxiliary variables. The motivation to consider this type of estimators follows from the survey practice when the same calibrated weights are used for the estimation of all the parameters needed.

The behaviour of the approximate linearized variance estimators (6) and (7) is very similar in all the cases considered.

The Jackknife estimator seems to be more adaptive. It is closer to the empirical variance and better reflects the real situation. We can suggest using it during the estimation procedure. The linearized estimator (6) and pseudo-linearized estimator (7) of the variance can be used provided only approximate variance estimation is needed. They are simple and require less computing time as compared to jackknife. The second variance estimator is preferable, because it is simpler and closer to the empirical and jackknife variances.

Table 1. Empirical variance and estimates of the variance of calibrated estimators. (population with exponential distribution, N = 300, sample size: n = 100)

Estimator	Empirical variance $\times 10^{-13}$	Linearized variance estimator $\times 10^{-13}$	Pseudo- linearized variance estimator $\times 10^{-13}$	Jackknife variance estimator $\times 10^{-13}$
$\rho(y,a)=0.8,\ \rho(z,b)=0.9$				
$\hat{C}_{w1}^{(non)}(y,z)$	2.960	5.729	5.541	2.603
$\hat{C}_{w6}^{(non)}(y,z)$	2.967	5.730	5.524	2.601
$\hat{C}_{w1}^{(tot)}(y,z)$	5.458	5.672	5.634	5.012
$\hat{C}_{w6}^{(tot)}(y,z)$	5.539	5.675	5.634	4.980
$\hat{C}_{w1}^{(lin)}(y,z)$	2.310	6.157	5.663	2.480
$\hat{C}_{w6}^{(lin)}(y,z)$	2.274	6.128	5.616	2.419
$\hat{C}_{1}(y,z)$	9.840	—	—	7.878
$\hat{C}_{2}(y,z)$	9.876	—	—	7.907
$\rho(y,a)=0.2,\ \rho(z,b)=0.9$				
$\hat{C}_{w1}^{(non)}(y,z)$	7.052	7.485	7.478	5.682
$\hat{C}_{w6}^{(non)}(y,z)$	7.107	7.481	7.471	5.742
$\hat{C}_{w1}^{(tot)}(y,z)$	4.871	5.654	5.616	5.042
$\hat{C}_{w6}^{(tot)}(y,z)$	5.002	5.675	5.637	5.032
$\hat{C}_{w1}^{(lin)}(y,z)$	10.029	7.267	7.315	7.869
$\hat{C}_{w6}^{(lin)}(y,z)$	10.018	7.254	7.309	7.866
$\hat{C}_{1}(y,z)$	10.376	—	—	7.878
$\hat{C}_{2}(y,z)$	10.411	—	—	7.907
$\rho(y,a)=0.2,\ \rho(z,b)=0.3$				
$\hat{C}_{w1}^{(non)}(y,z)$	11.555	7.403	7.228	7.745
$\hat{C}_{w6}^{(non)}(y,z)$	18.823	9.480	7.544	7.899
$\hat{C}_{w1}^{(tot)}(y,z)$	10.023	6.975	6.943	7.610
$\hat{C}_{w6}^{(tot)}(y,z)$	10.031	6.997	6.969	7.536
$\hat{C}_{w1}^{(lin)}(y,z)$	10.409	7.206	7.249	7.950
$\hat{C}_{w6}^{(lin)}(y,z)$	10.416	7.199	7.242	7.938
$\hat{C}_{1}(y,z)$	10.306	—	—	7.878
$\hat{C}_{2}(y,z)$	10.398	—	—	7.907

5. Conclusions

In conclusion we state that the calibrated estimators of the finite population covariance have a lower variance compared to the standard estimators, particularly, if the auxiliaries are strongly correlated with the study variables. They are also more efficient if only one highly correlated auxiliary is available. If the auxiliary variables are weakly correlated, the calibrated estimators may have a higher variance compared to the standard estimators. The loss function used has a minor impact on the efficiency of the estimators.

The linearized and pseudo-linearized variance estimators proposed are approximate and can be used only when an approximate variance is needed. The jackknife estimator seems to be more suitable, but it requires more calculation time.

Acknowledgments

The research is supported by the Grant of Lithuanian science foundation, T-24/07.

REFERENCES

DEVILLE, J-C., SÄRNDAL, C-E. (1992), Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, pp. 376—382.

HARMS, T., DUCHESNE, P. (2006), On calibration estimation for quantiles. *Survey Methodology*, 52, 1, pp. 37—52.

KRAPAVICKAITĖ, D., PLIKUSAS, A. (2005), Estimation of a Ratio in the Finite Population. *Informatica*, 16(3), pp. 347—364.

PLIKUSAS, A. (2001), Calibrated estimators of the ratio. *Lithuanian Math. J.*, 41 (special issue), pp. 457—462.

PLIKUSAS, A. (2003), Calibrated weights for the estimators of the ratio. *Lithuanian Math. J.*, 43, pp. 543—547.

PLIKUSAS, A., PUMPUTIS, D. (2007), Calibrated estimators of the population covariance. *Acta Applicandae Mathematicae*, 97(1-3), pp. 177—187.

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

SITTER, R., WU, C. (2002). Efficient Estimation of Quadratic Finite Population Functions in the Presence of Auxiliary Information. *Journal of the American Statistical Association*, 97, pp. 535—543.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 487—501

INVERSE PROBABILITY OF CENSORING WEIGHTING METHOD IN SURVIVAL ANALYSIS BASED ON SURVEY DATA

Marjo Pyy-Martikainen[1], Leif Nordberg[2]

ABSTRACT

In survival analysis based on survey data, attrition implies that a part of the event times are right-censored: it is only known that the true time exceeds that observed. To simplify the analysis, it is usually assumed that the process generating right-censoring is independent of the remaining event time. In practice, the assumption of independent censoring may not always hold. Dependent censoring may cause a bias in survival analysis. An inverse probability of censoring weighting (IPCW) method has been proposed to adjust for bias in survival analysis due to dependent censoring. To our knowledge, however, there are no empirical applications of the method in a complex survey data context. We use simulation methods to study the statistical properties of IPCW method in an artificial 2-wave panel survey. Our simulation study shows that the IPCW method is able to reduce bias in survival estimation also when there is only little information about the determinants of the right-censoring mechanism.

Key words: Inverse probability of censoring weights; survival analysis; complex surveys; simulation study.

1. Introduction

In panel surveys, data on durations spent in various states is often collected. Examples include duration of unemployment or employment, duration of poverty, duration of social assistance benefit receipt etc. In the analysis of durations or spells based on survey data, attrition implies that the ending date of some of the spells is unknown. For these *right-censored* spells, it is only known that the

[1] Department of Economics and Statistics, Åbo Akademi University and Statistics Finland, e-mail: marjo.pyy-martikainen@stat.fi
[2] Department of Economics and Statistics, Åbo Akademi University, e-mail: lnordber@abo.fi

length of the spell was at least that observed. Right-censoring may also occur because of end of follow-up time.

It is usually assumed, in order to make the analysis easier, that the right-censoring mechanism is *independent* of the remaining event time. This means that the right-censoring mechanism does not remove individuals from the survey because of particularly long or short durations. Under an independent right-censoring mechanism censoring does not cause bias and can thus be ignored in the analysis. In social surveys, the probability of attrition may be related for example to social exclusion which may be manifested by a long duration of unemployment or poverty. In such a situation, analyses of unemployment or poverty duration that ignore the censoring mechanism will lead to biased estimates.

Robins (1993) introduced an *inverse probability of censoring weighting* (IPCW) method that aims to correct for bias due to dependent right-censoring utilizing auxiliary variables related to both censoring and the duration of interest. He showed that if right-censoring is conditionally independent given the auxiliary variables, then using IPCW versions of Kaplan-Meier and Cox partial likelihood estimators result in consistent estimation. In simulation studies by van der Laan and Hubbard (1997) and van der Laan and Robins (1998) it has been shown that IPCW-based estimators perform remarkably well in a non- or semiparametric setting and in situations where the information about survival times is very limited.

In social surveys it is likely that all auxiliary variables needed to achieve conditional independence of censoring and event times are not observed. The censoring mechanism may thus contain some information on the event time of interest even after the IPCW correction. We conduct a simulation study in order to investigate the performance of the IPCW method in such a less-than-perfect situation. Our aim is to find out how strong the auxiliary information has to be for the IPCW method to be a useful tool. We take a design-based approach in the analysis. Consequently, our target parameters are the finite population regression coefficient B and survival function $S(t)$ that would be obtained from the estimation procedure if all data values in the finite population were available instead of having a sample only. The use of the IPCW method in a complex survey data context has previously been discussed by Lawless (2003a). However, we are unaware of any empirical applications of the method in the analysis of complex survey data. Our simulation study indicates that the method may be very useful even when the censoring mechanism is only partially known.

The paper is organized as follows. Section 2 gives a short introduction to some basic concepts of survival analysis. Section 3 discusses model-based and design-based approaches to survival analysis and introduces the design-based versions of Kaplan-Meier estimator and the partial likelihood function used to estimate the parameters of the Cox proportional hazards model. Section 4 introduces the concept of independent censoring and the IPCW method aimed to

adjust for bias caused by violation of this assumption. The performance of the IPCW method is studied by simulation methods in section 5. Section 6 concludes by discussing the findings from the simulation study.

2. Basic concepts of survival analysis

We are interested in making inferences about a duration or spell variable T. Because of censoring, only $t = \min(T, C)$ and $\delta = I(T \leq C)$ are observed, where C is a censoring time and δ is an event indicator. If $\delta = 1$, T is observed, and if $\delta = 0$, then we know only that the event time is longer than the censoring time. In longitudinal surveys, C depends on the length of the follow-up time, on the time at which the duration began and on the time of attrition (Lawless 2003a). The survival function and the hazard function are the two most important ways to express the distribution of a duration variable T. The value of the survival function

$$S(t) = P(T \geq t)$$

at time t is the probability that the spell is at least as long as t. The value of the hazard function

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid T \geq t)}{dt}$$

at time t describes the conditional probability of spell completion at time t, given that the spell has lasted until t. Models for the duration variables are usually constructed by defining the way covariates affect the hazard function.

3. Model-based and design-based approaches to survival analysis

A classical model-based analysis assumes that observations for different units are independent and that the sampling design is noninformative. Under a noninformative sampling design, the sample inclusion probabilities are not related to the values of outcome variables (for two alternative more formal definitions of noninformative sampling, see Chambers and Skinner, 2003, p.4 and Pfeffermann and Sverchkov, 2003, p.176). Longitudinal social surveys often have a complex sampling design with unequal probabilities of selection, stratification and clustering of observations. As a consequence, both the assumptions of independence and noninformativeness may be violated. In such a case it is necessary to take the impact of the sampling design into account in the analysis. A common approach to the design-based inference about a model parameter θ is to specify a finite population parameter θ_U that would be obtained from the model

estimation procedure if all data values in the finite population U were available instead of having a sample only. An estimate of θ_U is then obtained using sample data values and sample weights. This approach is also followed in our study. We are interested in estimating the finite population regression coefficient B from a Cox proportional hazards model. The design-based theory of Cox proportional hazards model was developed by Binder (1992). Similarly, we are interested in estimating the empirical survival function $S(t) = \frac{1}{N}\sum_{i=1}^{N} I(T_i \geq t)$ based on all units in the finite population (Lawless, 2003a). The following two subsections introduce the design-based versions of Kaplan-Meier estimator and the partial likelihood function used to estimate the parameters of the Cox proportional hazards model. In both cases we assume the population is constant over time and that we have at most one spell per unit.

Without going deeper in the discussion concerning the relative merits of design-based versus model-based analysis (see e.g. Pfeffermann, 1993), we see it for many reasons as both interesting and worthwhile to try to derive optimal estimators for the results one would get if one made the same analysis in an "ideal" situation, i.e. in the case where all data values in the finite population U were available.

3.1. Kaplan-Meier estimator

The Kaplan-Meier estimator (Kaplan and Meier, 1958) is a nonparametric estimator of the survival function $S(t)$. Folsom, Lavange and Williams (1989) developed an estimator that is appropriate when survival data is obtained from a complex survey. A lucid discussion of this estimator can be found in Lawless (2003b). Let $t_i, i = 1,\ldots,n$ be the observed event and censoring times in the sample of size n. Let $t_{(1)},\ldots,t_{(h)},\ldots,t_{(r)}$ be the ordered event times. The weighted number of observations undergoing an event at $t_{(h)}$ is $D_{(h)} = \sum_{i=1}^{n} I(t_i = t_{(h)})\delta_i w_i$, where w_i is the weight attached to observation i and δ_i is the event indicator defined earlier. The weighted number of observations with event or censoring times exceeding $t_{(h)}$ is $N_{(h)} = \sum_{i=1}^{n} I(t_{(h)} \leq t_i) w_i$. The weighted Kaplan-Meier estimator of the survival function is defined as

$$\hat{S}(t) = \prod_{h=1}^{r} \left(1 - \frac{D_{(h)}}{N_{(h)}}\right)^{I(t_{(h)} \leq t)} . \qquad (1)$$

Note that $D_{(h)}$ estimates the number of population units that undergo an event at time $t_{(h)}$ and $N_{(h)}$ estimates the population size of the risk set at time $t_{(h)}$. $\hat{S}(t)$ is thus an estimator of a population survival function that would be obtained if all the units of the finite population of interest were available for analysis.

3.2. Cox proportional hazard model

It is often of interest to find out how certain covariates $x = (x_1, \ldots, x_p)$ are related to the event time T. One of the most popular tools to study the association between T and x is the Cox proportional hazards model (Cox, 1972). The model specifies the hazard function as a product of two terms:

$$\lambda(t \mid x) = \lambda_0(t) \exp(x\beta),$$

where $\lambda_0(t)$ is a baseline hazard function that depends only on the event time and $\exp(x\beta)$ defines the way covariates x affect the hazard function. One reason for the popularity of the Cox proportional hazards model is the fact that the model parameters β can be estimated without assuming any parametric distribution for the event time variable T.

For survival data obtained from a complex survey, Binder (1992) used a pseudo-likelihood method to estimate the parameters and their variances for a Cox proportional hazards model. The unequal selection probabilities are taken into account by using sample weights. The dependence between observations is not modelled explicitly but is taken into account in variance estimation. The model is estimated by maximising a partial likelihood function. For a population of N units, the partial likelihood function is defined as

$$PL = \prod_{i=1}^{N} \left[\frac{\lambda(t_i \mid x_i)}{\sum_{j=1}^{N} I(t_i \le t_j) \lambda(t_i \mid x_j)} \right]^{\delta_i},$$

where x_i is the covariate vector, t_i is the spell length, and δ_i is the event indicator related to unit i. $I(t \le t_j)$ indicates whether the spell of unit j is still going on at time t. The sum $\sum_{j=1}^{N} I(t \le t_j)$ defines the size of the risk set, i.e. the number of spells still going on at time t. Note that the part of the hazard function that depends on event time only is common to each unit and cancels from the expression. The partial likelihood function can thus be expressed as

$$PL = \prod_{i=1}^{N} \left[\frac{\exp(x_i B)}{\sum_{j=1}^{N} I(t_i \le t_j) \exp(x_j B)} \right]^{\delta_i},$$

where B is the vector of population regression coefficients. B is determined as the solution to the score equations:

$$\frac{\partial \log PL}{\partial B} = \sum_{i=1}^{N} \delta_i \left[x_i - \frac{\sum_{j=1}^{N} I(t_i \le t_j) x_j \exp(x_j B)}{\sum_{j=1}^{N} I(t_i \le t_j) \exp(x_j B)} \right] = 0.$$

As noted by Roberts and Kovacevic (2007), if all units of the finite population do not experience spells, then N is the size of the subpopulation that experiences spells. To estimate the population regression coefficient B from a sample of n observations, Binder (1992) proposed the following pseudo-score estimating equations:

$$\sum_{i=1}^{n} w_i \delta_i \left[x_i - \frac{\sum_{j=1}^{n} w_j I(t_i \le t_j) x_j \exp(x_j \hat{B})}{\sum_{j=1}^{n} w_j I(t_i \le t_j) \exp(x_j \hat{B})} \right] = 0, \qquad (2)$$

where $w_j, j = 1, \ldots, n$ are the sample weights attached to the sample observations. The estimator $\hat{B}$ that solves equation (2) is the pseudo-maximum likelihood estimator of B (Binder, 1992). Binder (1992) and Roberts and Kovacevic (2007) discuss the design-based estimation of variance of $\hat{B}$.

4. Dependent censoring and the IPCW method

Censoring of spells occurs because of the shortness of the follow-up period or because of attrition. Censoring is thus related to the data collection and not to the phenomenon under study. Therefore, censoring should not affect the analysis of spells. To make analysis easier, it is usually assumed that the censoring mechanism is *independent* of the remaining event time (see e.g. Kalbfleisch and Prentice, 1980, pp. 119-121). For independent censoring mechanisms, the cause-specific hazard of T equals the marginal hazard:

$$\begin{aligned} \lambda_T(t \mid x, C \geq t) &= \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid T, C \geq t, x)}{dt} \\ &= \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid T \geq t, x)}{dt} = \lambda(t \mid x). \end{aligned} \tag{3}$$

Assumption (3) means that censoring and failure mechanisms are conditionally independent, given x. A (conditionally) independent right-censoring mechanism does not remove units from the survey because of particularly long or short spells. Under a (conditionally) independent censoring mechanism censoring does not cause bias and can thus be ignored in the analysis. In reality the assumption of independent censoring may not always hold. In social surveys, attrition may be related to particularly long or short spells of unemployment, poverty, supplementary benefit receipt etc. If this is not properly taken into account, the corresponding estimates from spell analyses are biased. It should be noted that bias is an outcome-specific issue. Attrition that is selective with respect to duration of benefit receipt may not cause bias in the analysis of unemployment duration.

4.1. Inverse probability of censoring weights

The inverse probability of censoring weighting (IPCW) method was introduced by Robins in 1993. The method aims to correct for bias due to dependent censoring. The assumption of independent censoring (equation (3)), can be shown to be equivalent to

$$\lambda_C(t \mid x, T, T > t) = \lambda_C(t \mid x, T > t), \tag{4}$$

which means that the cause-specific hazard of censoring does not depend on the (possibly unobserved) event time T. Dependent censoring mechanisms thus violate equation (4). The fundamental assumption underlying the IPCW method is that, given a vector of auxiliary variables z,

$$\lambda_C(t \mid x, z, T, T > t) = \lambda_C(t \mid x, z, T > t). \tag{5}$$

Given assumption (5), equation (4) is true if z does not predict censoring, ie. if

$$\lambda_C(t \mid x, z, T > t) = \lambda_C(t \mid x, T > t). \tag{6}$$

The assumption of independent censoring can thus be tested by modelling the cause-specific hazard of censoring using e.g. the Cox proportional hazard model. If the auxiliary variables z explain the cause-specific hazard of censoring, then censoring is dependent, which has to be taken into account in the analysis. The auxiliary variables z are variables which are not of interest as such, but which are

used to correct for bias due to dependent censoring. In order to be effective in this respect, the auxiliary variables should be associated with both censoring and event times. Robins (1993) showed that if the assumption (5) holds, then using in equations (1) and (2) weights defined by

$$w_i(t) = \frac{1}{S_C(t \mid x_i, z_i)},$$

where $S_C(t \mid x_i, z_i)$ is the cause-specific survival function for unit i, results in consistent estimation under dependent censoring. The estimate of $S_C(t \mid x_i, z_i)$ can be based on a fit of a Cox proportional hazard model with censoring as the event of interest. The weights $w_i(t)$ are time-dependent and inversely proportional to the conditional probability of having remained uncensored until time t, given x_i and z_i.

The approach of Robins is purely model-based. The sampling design has no role in the analysis and the weights are used only to correct for dependent censoring. Lawless (2003a) discussed the use of IPC weights in the estimation of survival function based on complex survey data. He showed that, in general, weights related to both the sampling design and censoring mechanism are needed for consistent estimation. He proposed the use of weights defined by

$$w_i^*(t) = \frac{1}{\pi_i \times S_C(t \mid x_i, z_i)}, \qquad (7)$$

where $\pi_i = P(i \in s)$ is the sample inclusion probability for unit $i, i = 1, \ldots, N$.

5. Simulation study

If the censoring model is correctly specified, then the IPC weighted Kaplan-Meier and Cox partial likelihood estimators can fully correct the bias due to dependent censoring. In longitudinal surveys, attrition depends on many variables, some of which may not be observed. Thus, the estimated censoring model should be considered as an approximation of the true model and, consequently, there is likely to be some residual dependency between T and C even after conditioning on z. Our aim is to study the bias-reducing power of the IPCW method in the presence of dependent censoring under the following scenarios:

1. The variable determining the censoring mechanism, z, is known.
2. We observe a variable that is either a) strongly or b) weakly associated with the variable z.
3. The variable z is unknown.

We assume that we are interested in the survival function and how a single covariate affects the hazard function. The parameters of interest are thus the values of the finite population survival function $S(t)$ at certain time points and the finite population regression coefficient B of the covariate. The statistical properties of IPC weighted estimators $\hat{B}$ and $\hat{S}(t)$ using weights defined in equation (7) were studied by simulation methods. Four different artificial populations corresponding to the above scenarios were generated and from each population, $K = 500$ independent samples were drawn. For each sample s_j, $j = 1\ldots, K$, drawn, estimates $\hat{B}_j$ and $\hat{S}_j(t)$ were calculated. The distribution of the K estimates was used as an approximation of the sampling distribution and the following estimators

$$\overline{\hat{B}} = \frac{1}{K}\sum_{j=1}^{K}\hat{B}_j$$

$$S^2_{\hat{B}} = \frac{1}{K-1}\sum_{j=1}^{K}(\hat{B}_j - \overline{\hat{B}})^2$$

were used to estimate the mean and variance of the sampling distribution of IPC weighted $\hat{B}$. For the IPC weighted Kaplan-Meier estimator, the mean and variance of the sampling distribution were estimated at time points $t = 1,25,50,75,100,125,150$.

5.1. Generation of the populations

The longitudinal population of interest is defined as individuals belonging to the target population at the beginning of the survey. The population is assumed to remain constant over time. Thus, there are no exits from or entrances to the target population. Each of the 4 populations are of size $N = 10000$ individuals and consist of $D = 8$ subgroups defined by binary variables sex, level of education and social exclusion. We are interested in the overall survival function of unemployment spells and the effect of education on the spell length.

Social exclusion is an unobserved variable that determines the probability of attrition. Sex is used as a stratification variable in the sampling design and as an auxiliary variable in the censoring model.

Table 1. The association between variables sex and social exclusion in populations 1, 2a, 2b and 3.

Population	Degree of association	ϕ	% of men excluded	% of women excluded
1	perfect	1	100	0
2a	strong	0.6	80	20
2b	weak	0.2	60	40
3	none	0	50	50

The 4 populations differ by the degree of association between variables sex and social exclusion according to Table 1. The phi coefficient ϕ measures association between 2 binary variables. The binary variables are considered positively associated if most of the data falls along the diagonal cells of a 2-by-2 frequency table. Value $\phi = 1$ corresponds to perfect positive association, value $\phi = -1$ to perfect negative association and value $\phi = 0$ to no association.

We consider a single spell analysis with one unemployment spell per individual (in an empirical analysis it could be e.g. the first spell beginning during the observation period). The unemployment spells were generated from the Weibull distribution, whose hazard function is

$$\lambda(t \mid a,b) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1}. \qquad (8)$$

The shape parameter a was set equal to 0.8 in each subpopulation. This corresponds to a decreasing hazard rate. The scale parameter b_1 of subpopulation 1, socially excluded men with low education, was set equal to 100, which corresponds a median duration of 63 days. The other scale parameters b_d, $d = 2,\ldots,8$, were chosen according to the hazard rates of Table 2.

The median duration of the unemployment spells as well as the effect of education on the hazard of spell completion are different among the excluded and among the non-excluded. The ratio of hazard rates (the hazard ratio or briefly hr) for the variable education is 3 among the non-excluded and 1.5 among the excluded. Censoring that depends on exclusion status thus biases both the estimates of survival function $S(t)$ and the estimate of the regression coefficient B.

Table 2. The hazard rates among the subpopulations. Median unemployment durations in parentheses.

	Excluded		Non-excluded	
	Low education	High education	Low education	High education
Men	h_1 (63)	$h_2 = 1.5h_1$ (38)	$h_3 = 2h_1$ (27)	$h_4 = 6h_1$ (7)
Women	$h_5 = h_1$ (63)	$h_6 = 1.5h_1$ (38)	$h_7 = 2h_1$ (27)	$h_8 = 6h_1$ (7)

5.2. Sampling design

We drew stratified simple random samples without replacement using sex as a stratification variable. Each sample is selected at a single time point, say day 0, at which there are $N = 10000$ individuals in the finite population. For each unit in the population is generated an unemployment spell according to relevant Weibull distribution. Unemployment spells start randomly during days $1,\ldots,30$. Each sample has 350 men and 250 women corresponding to inclusion probabilities of 0.07 and 0.05. This is a relatively simple sampling design where sample weights are however needed to produce design-unbiased population-level estimates. The sample weights are defined as the inverses of the inclusion probabilities. For each sample, an artificial 2-wave panel survey with attrition was conducted. The first wave interview is assumed to occur at day 30. It is assumed that there is no nonresponse at wave 1. Consequently, each unemployment spell is observed at least until day 30. We generated selective attrition by stratifying the samples according to exclusion status and drawing 80% samples among the non-excluded and 20% samples among the excluded. This corresponds to an attrition rate of 20% among the non-excluded and an attrition rate of 80% among the excluded. For the non-attriters, the observed duration is determined as $t = \min(T,600)$. For each sample, the IPC weights were constructed using sex as an auxiliary variable in the censoring model and weighted estimates of $S(t)$ and B were calculated. As the weights are time-varying the data had to be transformed into a counting process form (see e.g. Therneau and Grambsch, 2000, p. 68), where each unemployment spell is split into several intervals, the splitting points being defined by the times at which censoring occurs in the sample.

5.3. Results

The results from our simulation study are shown in Table 3. The number of replicate samples was 500. The true population parameters that are being estimated are shown in the first column. For both the design-weighted and IPC weighted estimators, the mean, standard deviation and percent bias are reported.

The percent bias of the design-weighted estimators shows how much selective attrition distorts the results. In population 1, the association between sex and social exclusion is perfect. This corresponds to a situation where the censoring mechanism is known and is, thus, an ideal situation for the IPC correction. Looking at the last column of Table 3, we see that the bias due to selective attrition has indeed almost vanished. A small positive bias remains in both $\hat{B}$ and $\hat{S}(t)$. As noted by Binder (1992), $\hat{B}$ is a design-consistent, but not a design-unbiased estimator of B. We are not aware of results concerning the design-based properties of the weighted Kaplan-Meier estimator.

In general, the bias of IPC weighted estimators grows as the association of sex and social exclusion gets weaker but is always less than the bias of design-weighted estimators. When sex and social exclusion are independent, the bias of IPC weighted Kaplan-Meier estimators is equal to that of design-weighted estimators. In that case there is thus no gain from using IPC weights in survival curve estimation.

Interestingly, the IPC weighted estimators of the hazard ratio $\exp(B)$ perform quite well relative to design-weighted estimators even when the association between sex and social exclusion is weak or when the variables are independent. This may be explained in the following way. Because of censoring, the IPC weights grow over time. This means that persons who remain a long time in the risk set and, therefore, are more likely to be excluded, get larger values of weights. This corrects the estimates in the right direction.

6. Discussion

We conducted a simulation study to investigate the performance of IPCW method in survival analysis based on complex survey data. If the censoring and event times are conditionally independent, given a set of auxiliary variables, then using this information in the construction of IPC weights can remove bias due to dependent censoring. However, in real-world situations, what is often observed are not the variables determining the censoring mechanism but some correlates of them. As a consequence, there may be residual dependency between censoring and event times even after the IPCW correction. Our simulation study shows that the IPCW method may be useful in survival analysis based on complex survey data even in such less-than-perfect real-world situations. Remarkably, there are gains from using the IPCW method in the estimation of the population regression coefficient even when the censoring mechanism is completely unknown. The development of design-based variance estimation methodology for IPCW Kaplan-Meier and Cox partial likelihood estimators remains an area where further research is needed (Lawless, 2003a).

Table 3. Results of a simulation study with 500 replications. $hr = \exp(B)$.

	Population parameters	Design weighted estimates mean	s.d.	% bias	IPC weighted estimates mean	s.d.	% bias
	$hr = 1.687$	1.977	0.213	17.2	1.710	0.235	1.4
	$S(1) = 0.940$	0.940	0.009	0.0	0.940	0.009	0.0
	$S(25) = 0.493$	0.477	0.022	-3.3	0.502	0.022	1.8
1	$S(50) = 0.333$	0.294	0.025	-11.8	0.342	0.027	2.6
	$S(75) = 0.248$	0.203	0.023	-18.2	0.253	0.027	2.3
	$S(100) = 0.191$	0.149	0.021	-22.3	0.197	0.027	2.8
	$S(125) = 0.147$	0.109	0.019	-26.1	0.151	0.026	2.6
	$S(150) = 0.116$	0.083	0.017	-28.3	0.119	0.024	2.2
	$hr = 1.663$	1.928	0.197	15.9	1.748	0.212	5.1
	$S(1) = 0.936$	0.935	0.010	0.0	0.935	0.010	0.0
	$S(25) = 0.492$	0.476	0.023	-3.2	0.484	0.023	-1.6
2a	$S(50) = 0.336$	0.296	0.026	-11.9	0.313	0.026	-7.0
	$S(75) = 0.242$	0.200	0.024	-17.2	0.215	0.024	-10.9
	$S(100) = 0.185$	0.145	0.021	-21.7	0.159	0.022	-13.9
	$S(125) = 0.142$	0.106	0.019	-25.4	0.119	0.021	-16.5
	$S(150) = 0.113$	0.082	0.017	-27.4	0.093	0.019	-17.7
	$hr = 1.655$	1.960	0.231	18.4	1.807	0.233	9.2
	$S(1) = 0.938$	0.939	0.009	0.1	0.939	0.009	0.1
	$S(25) = 0.504$	0.494	0.023	-2.0	0.495	0.023	-1.7
2b	$S(50) = 0.340$	0.300	0.025	-11.9	0.302	0.024	-11.1
	$S(75) = 0.246$	0.206	0.023	-16.3	0.207	0.023	-15.7
	$S(100) = 0.181$	0.143	0.021	-20.9	0.144	0.021	-20.4
	$S(125) = 0.138$	0.103	0.019	-25.4	0.104	0.019	-24.6
	$S(150) = 0.109$	0.079	0.017	-27.3	0.080	0.017	-26.4
	$hr = 1.720$	2.003	0.230	16.5	1.853	0.245	7.7
	$S(1) = 0.937$	0.938	0.009	0.1	0.938	0.009	0.1
	$S(25) = 0.495$	0.475	0.023	-3.9	0.476	0.023	-3.9
3	$S(50) = 0.332$	0.286	0.024	-13.9	0.286	0.024	-13.9
	$S(75) = 0.243$	0.195	0.024	-19.6	0.195	0.024	-19.6
	$S(100) = 0.182$	0.136	0.021	-24.7	0.136	0.021	-24.6
	$S(125) = 0.139$	0.099	0.018	-28.4	0.099	0.018	-28.3
	$S(150) = 0.110$	0.074	0.016	-32.8	0.074	0.016	-32.7

Acknowledgements

The authors would like to thank an anonymous referee for comments that greatly improved the readability of the paper.

REFERENCES

BINDER, D. (1992). Fitting Cox's Proportional Hazards Models from Survey Data. Biometrika 79, 1, 139—147.

CHAMBERS, R. and SKINNER, C. (2003). Analysis of Survey Data. Wiley.

COX, D. (1972). Regression Models and Life Tables. Journal of Royal Statistical Society B, 34, 187—220.

FOLSOM, R., LAVANGE, L. and WILLIAMS, R. (1989). A Probability Sampling Perspective on Panel Data Analysis. In Panel Surveys (D. Kasprzyk, G. Duncan, G. Kalton and M. Singh, eds), pp. 108—38. Wiley.

KALBFLEISCH, J. and PRENTICE, R. (1980). The Statistical Analysis of Failure Time Data. Wiley.

KAPLAN, E. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. Journal of American Statistical Association 53, 457—481.

LAWLESS, J. (2003a). Censoring and Weighting in Survival Estimation from Survey Data. SSC Annual Meeting, June 2003, Proceedings of the Survey Methods Section.

LAWLESS, J. (2003b). Event History Analysis and Longitudinal Surveys. In Analysis of Survey Data (R. Chambers and C. Skinner, eds), 221—243. Wiley.

PFEFFERMANN, D. (1993). The Role of Sampling Weights when Modeling Survey Data. International Statistical Review, 61, 317—337.

PFEFFERMANN, D. and SVERCHKOV, M. (2003). Fitting Generalized Linear Models under Informative Sampling. In Analysis of Survey Data (R. Chambers and C. Skinner, eds), 175—195. Wiley.

ROBERTS, G. and KOVACEVIC, M. (2007). Modelling Durations of Multiple Spells From Longitudinal Surveys. Survey Methodology, 33, 13—22.

ROBINS, J. (1993). Information Recovery and Bias Adjustment in Proportional Hazards Regression Analysis of Randomized Trials Using Surrogate Markers. In Proceedings of the Biopharmaceutical Section, American Statistical Association, 24—33. Alexandria, Virginia: American Statistical Association.

THERNEAU, T. and GRAMBSCH, P. (2000). Modeling Survival Data. Extending the Cox Model. Springer-Verlag.

VAN DER LAAN, M. and HUBBARD, A. (1997). Estimation with Interval Censored Data and Covariates. Lifetime Data Analysis 3, 77—91.

VAN DER LAAN, M. and ROBINS, J. (1998). Locally Efficient Estimation with Current Status Data and Time-Dependent Covariates. JASA, 93, 693—701.

WILLIAMS, R. (1995). Product-Limit Survival Functions with Correlated Survival Times. Lifetime Data Analysis, 1, 171—186.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 503—517

REGRESSION COMPOSITE ESTIMATION WITH APPLICATION TO THE FINNISH LABOUR FORCE SURVEY

Riku Salonen[1]

ABSTRACT

This paper examines regression composite estimation in a complex rotating panel design. Empirical results are based on quarterly data from the Finnish Labour Force Survey (LFS). The survey is repeated over time with partially overlapping samples. Currently, the Finnish LFS uses generalised regression (GREG) estimation and calibration techniques. Estimation for employment and unemployment are based on the current quarter's data. It is expected that estimation can be improved by using the rotating panel property, because employment and unemployment tend to be correlated over time. The new regression composite (RC) estimator studied here extends the standard GREG estimator by taking advantage of the temporal correlations. It is shown that the RC estimator outperforms the standard GREG estimator for estimates of both quarterly levels and quarter-to-quarter changes in employment and unemployment.

Key words and phrases: Complex rotating panel; Auxiliary information; Composite estimation.

1. Introduction

In repeated surveys with partially overlapping samples, utilisation of information collected in previous waves has proven advantageous. Level estimates based on cross-sectional data can be improved by using data from the previous waves because of the correlation due to the common samples. The resulting estimates of change and average over time can also be improved (e.g. Lent, Miller, Cantwell and Duff 1999).

The method of composite estimation uses the rotation pattern of the survey to improve efficiency of estimates. In this approach, the overlapping and non-

[1] Statistics Finland, riku.salonen@stat.fi

overlapping components of the sample are weighted differently. Traditional methods for composite estimation, known as K composite and AK composite estimators, have been discussed in recent literature (e.g. Fuller and Rao 2001). A weighting factor K ($0 < K < 1$) determines the weight in the weighted average of two estimators: (1) the current estimator (time t) and (2) the sum of the previous estimator (time $t-1$) and an estimator of the change since the previous time point. The weighting factor A ($0 < A < K < 1$) is the weight of adjustment term that determines the difference between estimates from the overlapping and non-overlapping components of the sample. The K composite estimator is a special case of the AK composite estimator with $A = 0$.

Although the traditional composite estimators lead to improved estimates, they suffer from a number of drawbacks such as inconsistency of estimates. Fuller (1990) and Lent, Miller, Cantwell and Duff (1999) introduced the method of the AK composite weighting estimator. This method eliminates the problem of inconsistent estimates. Since 1998, the United States Current Population Survey (CPS) uses the AK composite weighting estimator.

The regression-based composite estimation method was developed in Canada. The first variant of regression composite estimation was called level-driven modified regression (MR1) estimator (Singh and Merkouris 1995, Singh 1996). MR1 produced reduced variance estimates of level relative to standard generalised regression (GREG) estimation based on cross-sectional data. The second variant of regression composite estimation was described by Singh, Kennedy, Wu and Brisebois (1997). This method was called change-driven modified regression (MR2) estimator. It produced reduced variance estimates of change compared to the standard GREG estimator.

The regression composite (RC) estimator was suggested by Fuller and Rao (2001). It is a compromise between MR1 and MR2. RC estimation has also been studied by Singh, Kennedy and Wu (2001), Gambino, Kennedy and Singh (2001), Bell (2001), and Beaumont and Bocci (2005).

Exploitation of sample overlap over time to improve the efficiency of estimates can be done via calibration by using the RC estimator. This method extends the GREG estimator by using information from the previous wave in a similar manner as the standard GREG estimator uses auxiliary variables. The RC method uses correlation between labour force characteristics of two consecutive waves. Level estimates based on cross-sectional data can be improved by using past data because of the correlation due to the common samples. The resulting estimates of change and average over time can also be improved. A further advantage of the new approach is that it yields a single set of estimation weights, leading to internal consistency of estimates. Since 2000, the RC estimator has been successfully used in the Canadian LFS. (See Gambino, Kennedy and Singh 2001.)

The design of the Finnish LFS is a complex rotating panel. The survey is repeated over time with partially overlapping samples where there is a 3/5 sample

overlap between any two successive quarters. Currently in the estimation of quarterly figures for employment and unemployment, GREG estimation and calibration techniques are used (Särndal, Swensson and Wretman 1992, Deville and Särndal 1992). Certain register-based auxiliary information (e.g. sex, age group, region, register-based job-seeker status taken from an administrative register maintained by Ministry of Labour) are incorporated in the estimation procedure. The GREG estimator is based on the current quarter's data and does not use the rotating panel property of the study design.

The proposed RC estimator extends the GREG estimator by taking advantage of the correlations over time. RC estimation is expected to involve gains in efficiency for estimates of quarterly levels and quarter-to-quarter changes, because characteristics such as employment (especially employment by Standard Industrial Classification) and unemployment are correlated over time.

This paper examines regression composite estimation procedures that make use of sample information from previous waves and that can be implemented with a standard regression estimation program. Chapter 2 of the paper introduces the design of the Finnish LFS. Chapter 3 presents the current GREG estimator. Chapter 4 describes the RC estimator proposed by Fuller and Rao (2001). Chapter 5 gives empirical results based on quarterly data from the Finnish LFS. In this chapter, RC estimation results for employment and unemployment are compared with results from the current GREG estimator. Summary and discussion are given in Chapter 6.

2. Design of the Finnish LFS

The target population of the Finnish LFS is persons aged 15 to 74, including foreign workers, citizens temporarily abroad, members of the armed forces, non-resident citizens, and unsettled and institutional population. The LFS is a monthly survey of individuals selected by systematic sampling. For estimation purposes, the sampling design is approximated with a without-replacement simple random sampling design (SRSWOR). The sampling frame is based on the database of the total population maintained by Statistics Finland. The sample size is approximately 12,000 individuals each month divided into five waves and four or five reference weeks. The monthly sample is allocated so that the weekly sample sizes are equal in each wave. The reference quarters and years are groups of 13 or 52 consecutive weeks.

The survey is repeated over time with partially overlapping samples. Each person will be included five times during 15 months. The rotation pattern in the LFS can be described as follows 1-2-1-2-1-5-1-2-1 (see Djerf 2004). In the first month, an individual is in the panel in wave one and after a two-month break, he/she will be included in the interview in the second wave, and so on. The lag between the interviews is three months except for one occasion, when it is six months.

The design of the LFS ensures the independence of the monthly samples in each three-month period, i.e. a sample for a quarter consists of separate monthly samples. Each sampled person is included once per quarter. This simplifies the estimation of quarterly figures. In the LFS the sample size is 36,000 persons per quarter. There is dependence between successive quarters; the overlap from one quarter to the next is 3/5. There is also a 2/5 overlap between two consecutive years. The disadvantage of this rotating panel structure is that the annual average (of four quarters) will be estimated with a larger variance compared to independent samples.

3. Current estimation procedure

The Finnish LFS has been conducted by Statistics Finland since 1959. Over the years the estimation procedure has changed several times. The current estimation procedure was introduced in 1997. Register data on unemployment were used as auxiliary information at the estimation stage. The use of such auxiliary data significantly improved estimates on unemployment by reducing sampling errors and non-response bias (Djerf 1997). GREG estimation was used in this procedure.

Denote the finite population by $U = \{1,\ldots, k,\ldots, N\}$. A sample $s \subset U$ of size n is drawn by a sampling design $p(s)$ with inclusion probabilities π_k, $k \in U$. Under SRSWOR, the inclusion probabilities are $\pi_k = n/N$. The design weight of unit k is $a_k = 1/\pi_k = N/n$. Denote by y the variable of interest and by y_k its value for unit k.

In the Finnish LFS, post-stratification is used to improve the precision of estimation. The $H = 252$ post-strata are constructed by sex (2 classes), age group (6 groups) and region (21 regions). Let n_h be the number of sampled units in post-stratum h, so $\sum_{h=1}^{H} n_h = n$. At the population level, $\sum_{h=1}^{H} N_h = N$.

There is also missingness due to unit non-response. The weight adjusted for non-response is $d_k = 1/(\pi_k \hat{\theta}_k) = (N_h / n_h) \times (n_h / m_h) = N_h / m_h$ for element k in post-stratum h, where m_h is the number of responding units in post-stratum h and $\hat{\theta}_k = m_h / n_h$ is the estimated response probability for element k in post-stratum h. The weights d_k adjusted for non-response are calibrated using the available auxiliary information. The GREG estimator with linear fixed-effects assisting model is a special case of the calibration estimator (e.g. Särndal, Swensson and Wretman 1992).

As Deville and Särndal (1992 and 1993) show, the GREG estimator of a population total $t_y = \sum_U y_k$ can be given as $\hat{t}_{ygr} = \sum_r w_k^{gr} y_k$ where r refers to the respondent group and the calibrated weights are $w_k^{gr} = d_k g_k^{gr}$ with

$$g_k^{gr} = 1 + (\mathbf{t}_\mathbf{x} - \hat{\mathbf{t}}_\mathbf{x})' \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \mathbf{x}_k q_k . \quad (1)$$

The known auxiliary totals, called control totals, are $\mathbf{t}_\mathbf{x} = \left(t_{x1},...,t_{xj},... \ t_{xJ}\right)'$ and $\hat{\mathbf{t}}_\mathbf{x} = \left(\hat{t}_{x1},...,\hat{t}_{xj},... \ \hat{t}_{xJ}\right)'$ is a vector of estimates of the elements in $\mathbf{t}_\mathbf{x}$. The auxiliary information vector is defined as $\mathbf{x}_k = (x_{1k},...,x_{jk},...x_{Jk})'$ and q_k is a known constant (usually set equal to one). The calibration property assures that $\hat{\mathbf{t}}_\mathbf{x} = \sum_r w_k^{gr} \mathbf{x}_k = \sum_U \mathbf{x}_k = \mathbf{t}_\mathbf{x}$. In the Finnish LFS the auxiliary information vector is defined by four auxiliary variables taken from administrative registers: x_1 = sex (2 classes), x_2 = age (12 groups), x_3 = region (21 regions), x_4 = employment status in Ministry of Labour's job-seeker register (8 classes).

We used a linear distance function in the calibration procedure, available in CLAN, a program developed by Statistics Sweden for calibration and GREG estimation (Andersson and Nordberg 1998). Variance estimation in CLAN is based on GREG estimation. For variance estimation we need the residuals $e_k = y_k - \mathbf{x}_k' \hat{B}$, where

$$\hat{B} = \left(\sum_r \frac{x_k x_k' q_k}{\pi_k \hat{\theta}_k} \right)^{-1} \sum_r \frac{x_k y_k q_k}{\pi_k \hat{\theta}_k} .$$

The variance estimator of $\hat{t}_{ygr}$ under SRSWOR is given by

$$\hat{V}\left(\hat{t}_{ygr}\right) = \sum_{h=1}^{H} \frac{N_h^2}{m_h} \left(1 - \frac{m_h}{N_h}\right) \frac{1}{m_h - 1} \left[\sum_{r_h} \left(g_k^{gr} \times e_k \right)^2 - \frac{\left(\sum_{r_h} g_k^{gr} \times e_k \right)^2}{m_h} \right], \quad (2)$$

where r_h denotes the respondent set in post-stratum *h*.

4. Proposed estimation procedure

The current GREG estimator is based on the current quarter's data and does not use the rotation panel pattern. In RC estimation auxiliary data, known as composite auxiliary variables $\mathbf{z}_k$, are taken from the previous time period $t-1$. The composite auxiliary variables have random benchmarks determined by setting the weighted sum of variables $\mathbf{z}_k$ equal to the previous period's estimates. These estimated control totals are called composite control totals. Under a rotating

panel design, however, values for the composite auxiliary variables are known for the overlapping part of the sample. For the non-overlapping part the values are imputed.

In the Finnish LFS, the sample size is 36,000 persons per quarter. There is dependence between successive quarters; the overlap from a quarter to the next is 3/5. The part of the sample which is common for the current and previous quarters is referred to as the *matched*, i.e. overlap, sample. The remaining 2/5 part of the sample is known as the *unmatched*, i.e. non-overlap, sample.

In the RC estimation procedure for the LFS we used the following composite auxiliary data: labour force status (employed, unemployed) by sex/age group, labour force status by NUTS2 region and labour force status by industry. The corresponding composite auxiliary variables were defined as a linear combination of *MR*1 and *MR*2 as suggested by Fuller and Rao (2001). For the level-driven predictor *MR*1, data from the previous quarter were used for the matched sample, and mean imputation was used for the unmatched part. For the change-driven predictor *MR*2, carry backward imputation was used for the unmatched sample, and transformed values of the previous quarter data were used for the matched sample.

The composite auxiliary variables $\mathbf{z}_k$ were formulated as $\mathbf{z}_k = (1-\alpha)MR1_k + \alpha\, MR2_k$, $\alpha \in [0,1]$, where the choice of the coefficient α depends on the variable of interest and on the relative importance of level versus change (Gambino, Kennedy and Singh 2001). The level-driven and change-driven predictors are special cases corresponding to $\alpha = 0$ and $\alpha = 1$, respectively. The choices of the α value under this study are $\alpha = 0.3$, 0.5 and 0.7.

The level-driven predictor is given by

$$MR1_k = \begin{cases} \hat{t}_{t-1} / N_{t-1} & \text{if element } k \text{ belongs to the unmatched part of sample} \\ y_{t-1,k} & \text{if element } k \text{ belongs to the matched part of sample} \end{cases}$$

with $\hat{t}_{t-1}$ as an imputed value defined as the previous quarter's estimate of the total of the study variable and N_{t-1} is the corresponding population size, and $y_{t-1,k}$ refers to the observed value of the study variable for unit k at time point $t-1$.

The change-driven predictor is given by

$$MR2_k = \begin{cases} y_{tk} & \text{if element } k \text{ belongs to the unmatched part of sample} \\ y_{tk} + R^{-1}(y_{t-1,k} - y_{tk}) & \text{if element } k \text{ belongs to the matched part of sample} \end{cases}$$

where R is a ratio that adjusts the sample overlap from one quarter to the next (R = 3/5), and y_{tk} refers to the observed value of the study variable for unit k at time point t.

As Singh, Kennedy and Wu (2001) show, the RC estimator of $t_y = \sum_U y_k$ can be expressed in the form of $\hat{t}_{yrc} = \sum_r w_k^{rc} y_k$ where the calibrated RC weights w_k^{rc} are obtained in a similar manner as the GREG weights w_k^{gr}, except that the constraint $\sum_r w_k^{gr} \mathbf{x}_k = \mathbf{t}_\mathbf{x}$ is replaced by the constraints $\sum_r w_k^{rc} \mathbf{x}_k = \mathbf{t}_\mathbf{x}$ and $\sum_r w_k^{rc} \mathbf{z}_k = \hat{\mathbf{t}}_\mathbf{z}$. The vector of estimated composite control total $\hat{\mathbf{t}}_\mathbf{z}$ must be computed using the previous quarter's (quarter $t-1$) data. The RC weights w_k^{rc} are calibrated on the usual control totals $\mathbf{t}_\mathbf{x}$ given by $w_k^{rc} = d_k g_k^{rc}$, where g_k^{rc} has the same form as (1), with the exception that $\mathbf{x}_k$ and $\mathbf{t}_\mathbf{x}$ are replaced by $\left(\mathbf{x}_k', \mathbf{z}_k'\right)'$ and $\left(\mathbf{t}_\mathbf{x}', \hat{\mathbf{t}}_\mathbf{z}'\right)'$, respectively.

The approximate variance of $\hat{t}_{yrc}$ is calculated by using g_k^{rc} instead of g_k^{gr} in the GREG variance formula (2). Thus the variance of $\hat{t}_{yrc}$ is estimated by

$$\hat{V}\left(\hat{t}_{yrc}\right) = \sum_{h=1}^{H} \frac{N_h^2}{m_h}\left(1-\frac{m_h}{N_h}\right)\frac{1}{m_h-1}\left[\sum_{r_h}\left(g_k^{rc}\times e_k\right)^2 - \frac{\left(\sum_{r_h} g_k^{rc}\times e_k\right)^2}{m_h}\right]. \quad (3)$$

Here we have used the CLAN program for point and variance estimation.

5. Empirical results

This section summarises empirical results based on real data from the first and second quarters of the Finnish LFS in 2006. The RC method is used starting from the last quarter 2005, i.e. the time period from which the recursive process begins. The estimation results for employment and unemployment are presented for the current GREG estimator and for the proposed RC estimator. The choices of the α value under this study are α = 0.3, 0.5 and 0.7.

Relative efficiency (RE) evaluates the efficiency gains of the RC variance estimator (3) of $\hat{t}_{yrc}$ relative to the GREG variance estimator (2) of $\hat{t}_{ygr}$. RE can be formulated as $RE = \frac{\hat{V}\left(\hat{t}_{ygr}\right)\times 100}{\hat{V}\left(\hat{t}_{yrc}\right)}$ (Chen and Liu 2002). In this study, RE is introduced for the three different choices of the α value. A value of RE greater

than 100 indicates that the RC estimator is more efficient than the GREG estimator.

In Table 1, distribution of calibrated weights for GREG and RC estimators are given for the different values of the coefficient α. The calibrated weights are obtained by the CLAN program. The results show that the variation of the RC weights is smaller than that of the GREG weights. The variation of the RC weights is quite similar for the different α values.

Table 1. Distribution of calibrated weights for GREG and RC estimators for the 2nd quarter of 2006

Statistics for calibrated weights	GREG	RC $\alpha = 0.3$	RC $\alpha = 0.5$	RC $\alpha = 0.7$
Minimum	42.81	50.86	50.64	50.66
Maximum	416.29	227.13	223.40	221.87
Average	137.69	137.69	137.69	137.69
Median	135.12	137.00	137.06	137.15

Let us consider results for the level and change estimates of employment and unemployment for different population subgroups, by using the RE measure. The population subgroups of interest are sex, region and type of industry.

Table 2 gives RE results for level estimates for employment and unemployment with sex breakdown. RC estimation outperforms GREG estimation for all choices of α. The value of $\alpha = 0.7$ produces very efficient estimates at the quarterly level when compared with the other choices of α value. Efficiency gains appear to be largest for estimates for employment. For unemployment figures, all α values tend to produce quite similar efficiency.

Table 2. Relative efficiency (RE, %) of estimates for the level of employment and unemployment by sex, for different choices of the coefficient α

Level estimates		RE (%)		
Labour force status	Sex	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
Employed	Male	144.7	165.9	185.4
	Female	144.7	163.5	179.0
	Both sexes	151.4	165.0	173.8
Unemployed	Male	112.5	113.1	113.1
	Female	108.2	109.7	110.6
	Both sexes	104.4	105.1	105.4

Table 3 gives RE results for change estimates for employment and unemployment for males and females. Also here, RC estimation clearly outperforms GREG estimation, under all choices of α. For change estimates, the value of $\alpha = 0.7$ produces most efficient estimates; this especially holds for employment figures. Similarly as for the level estimates, efficiency gains are largest for estimates for employment. Also for change estimates, all α values tend to produce quite similar efficiency for the unemployment figures.

Table 3. Relative efficiency (RE, %) of estimates for quarter-to-quarter change of employment and unemployment by sex, for different choices of the coefficient α

Change estimates		RE (%)		
Labour force status	Sex	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
Employed	Male	145.3	167.5	188.1
	Female	148.7	169.7	187.8
	Both sexes	156.1	171.4	181.2
Unemployed	Male	115.2	116.5	116.8
	Female	109.8	110.5	110.8
	Both sexes	105.6	106.2	106.5

Let us turn to estimation results for labour force status by NUTS2 regions. There are four NUTS2 regions in Table 4 giving results for the level estimates. The pattern of efficiency figures is similar as in Table 2. Again, efficiency gains from RC estimation are large for estimates on employment, when compared with GREG estimation. Best results are for $\alpha = 0.7$. For unemployment, the choice of α value seems unimportant.

Table 4. Relative efficiency (RE, %) of estimates for the level of employment and unemployment by NUTS2 region, for different choices of the coefficient α

Level estimates			RE (%)		
Labour force status	NUTS2 region	Sample size	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
Employed	South Finland	8 803	145.4	165.4	182.6
	West Finland	4 579	141.2	161.8	177.9
	East Finland	2 134	140.8	157.5	168.5
	North Finland	1 998	142.2	159.2	173.3
Unemployed	South Finland	635	110.5	111.7	112.3
	West Finland	411	111.5	112.0	111.7
	East Finland	283	114.6	116.6	117.5
	North Finland	274	115.2	115.8	115.8

Table 5 indicates that the value of $\alpha = 0.7$ for employment produces very efficient estimates of quarter-to-quarter change when compared with the GREG estimator and other choices of α value. The best choice of α value for unemployment may be a value lower than $\alpha = 0.7$, especially for West Finland.

Table 5. Relative efficiency (RE, %) of estimates for quarter-to-quarter change of employment and unemployment by NUTS2 region, for different choices of the coefficient α

Change estimates			RE (%)		
Labour force status	NUTS2 region	Sample size	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
Employed	South Finland	8 803	146.8	167.8	186.0
	West Finland	4 579	143.1	164.1	181.3
	East Finland	2 134	142.7	163.0	179.8
	North Finland	1 998	145.5	165.2	181.6
Unemployed	South Finland	635	112.7	113.4	113.6
	West Finland	411	114.8	115.4	115.2
	East Finland	283	115.8	117.2	117.5
	North Finland	274	115.4	117.2	118.0

The results for employment by Standard Industrial Classification are found in Tables 6 and 7. Value $\alpha = 0.7$ produces very efficient estimates of quarterly level when compared with the GREG estimator and other choices of α value.

Table 6. Relative efficiency (RE, %) of estimates for the level of employment and unemployment by industrial classification, for different choices of the coefficient α

Level estimates		RE (%)		
NACE	Sample size	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
Agriculture	868	273.5	332.8	368.0
Manufacturing	3 385	299.7	383.7	449.4
Construction	1 126	254.6	301.5	327.9
Wholesale and retail trade	2 703	259.5	312.7	347.4
Transport, storage and communication	1 311	293.5	362.6	408.2
Financial intermediation	2 343	277.8	341.8	385.5
Public administration	5 956	253.6	304.1	337.5

Similar results can be seen for estimates of quarter-to-quarter change. The best α value for employment tends to be $\alpha = 0.7$. It produces very efficient estimates of quarter-to-quarter change when compared with the GREG estimator and other choices of α value.

Table 7. Relative efficiency (RE, %) of estimates for quarter-to-quarter change of employment and unemployment by industrial classification, for different choices of the coefficient α

Change estimates		RE (%)		
NACE	Sample size	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$
Agriculture	868	285.5	355.8	403.8
Manufacturing	3 385	309.9	402.6	478.0
Construction	1 126	246.7	291.0	318.7
Wholesale and retail trade	2 703	267.2	325.5	365.7
Transport, storage and communication	1 311	294.6	372.8	428.1
Financial intermediation	2 343	269.1	326.6	364.7
Public administration	5 956	261.0	318.0	357.6

6. Discussion and conclusions

In modern survey sampling, it is important to make the best use of the available auxiliary information so as to obtain as efficient estimators as possible (e.g. Estevao and Särndal 2002). In a Labour Force Survey, a rotating panel design can be effectively used to improve employment and unemployment estimates of both level and change by using data from the previous periods as

auxiliary information. The proposed RC estimator takes the advantage of correlations over time induced by sample overlap. This method extends the traditional GREG estimator in the sense that it uses data from previous wave as auxiliary information. Gambino, Kennedy and Singh (2001) showed that the RC estimator is usually efficient and stable, and often allows successful seasonal adjustment of the estimate series. Results from this study support well their results. Other advantages of the RC method are that if yields a single set of estimation weights, leading to internal consistency of estimates. RC estimation can be performed by using, with minor modification, standard software for GREG estimation, such as CLAN.

In this paper, empirical results for RC and GREG estimation are based on quarterly data from the Finnish LFS. Currently, the Finnish LFS uses the GREG estimator based on the current quarter's data. The rotating panel design of the LFS can be effectively used to improve efficiency of estimates for both quarterly level and quarter-to-quarter change. The estimation can be improved, because employment and unemployment are correlated over time. In the LFS, correlations for employment and unemployment between successive quarters are approximately 0.80 and 0.40, i.e. employment is highly correlated over time and unemployment is moderately correlated over time.

Empirical results are based on real data on the first and second quarters from the Finnish LFS in 2006. RC estimation results for employed and unemployed are presented for comparison with the current GREG estimator. The choices of α value under this study were $\alpha = 0.3$, 0.5 and 0.7. The RC estimator produced quarterly level and quarter-to-quarter change estimates that were usually more efficient than the estimates produced by the current GREG estimator. The results obtained with the different values of α indicated that the value $\alpha = 0.7$ is a reasonable choice for quarterly level and quarter-to-quarter change estimates.

For the variables that were included as composite control totals in the RC estimation procedure, there were substantial gains in efficiency for estimates of quarterly level and quarter-to-quarter change of employment figures. In particular, this holds for employment by Standard Industrial Classification. For most industries, gains of 200 to 300 per cent were typical. A reason for large efficiency improvement is the high correlation of employment over time. For unemployment estimates, the efficiency gains were modest, typically from 5 to 18 per cent. An explanation for this is that unemployment is only moderately correlated over time and the register data on unemployment (labour force status in Ministry of Labour's job-seeker register) are used as auxiliary information already at the GREG estimation stage. For variables that were not controlled, there were little or no efficiency gains from RC estimation, unless the variable in question was highly correlated with a composite auxiliary variable.

Chen and Liu (2002) reported a similar study using the Canadian LFS data for the time period between July 2000 and June 2001. The results were presented for the GREG estimator and the RC estimator with several α values, for both the

monthly level and month-to-month change estimates. The LFS in Canada is a continuous monthly survey of approximately 54,000 households selected using stratified multistage sampling. The households remain in the sample for six consecutive months. This means that the sample is split into six panels, and each month 1/6 of the panel is replaced after it has completed its six months' stay in the survey. Thus a 5/6 month-to-month sample overlap exists in any two consecutive months. The Canadian LFS uses the jackknife technique for variance estimation for the RC and GREG estimators (see Singh, Kennedy and Wu 2001). Chen and Liu (2002) reported that there were substantial gains in efficiency for variables that were added as composite control totals, such as employment by industry. For most industries, gains of to 10 to 50 per cent were typical for monthly level estimates. For estimates of month-to-month change, the efficiency gains for controlled variables were usually much greater. For unemployment estimates, the efficiency gains were usually from 5 to 18 per cent and minimal for variables that were not controlled. Their study showed that the value $\alpha = 0.67$ was a good compromise of monthly level and month-to-month estimates. Their results are quite similar as those of this study.

Bell (2001) has also compared the RC estimator with the value $\alpha = 0.7$ to the GREG estimator in the Australian LFS data for the time period from January 1993 to January 1999. The LFS is a monthly survey of approximately 30,000 households selected using multistage probability sample design. The sample is split into eight panels with each panel remaining in the survey for eight months and 1/8 of the sample being replaced each month. This results in a 7/8 month-to-month sample overlap. The Australian LFS uses the jackknife method for variance estimation for the RC and GREG estimators. With the Australian data, gains of to 30 to 140 per cent were typical for estimates of employment. For unemployment estimates, the efficiency gains usually ranged from 4 to 30 per cent.

This paper presented results on RC estimation for the Finnish LFS. The results are well comparable with results reported from other countries. Future studies on RC estimation for the Finnish LFS will extend the analysis to alternative variance estimators (e.g. jackknife and/or bootstrap) and other imputation methods (Beaumont 2005; Beaumont and Bocci 2005). An additional subject of future study will be the possible design bias of estimates and an optimal choice of the α value for quarterly level, quarter-to-quarter change and annual average of employment and unemployment.

REFERENCES

ANDERSSON, C. and NORDBERG, L. (1998): A User's Guide to CLAN 97. SCB, Stockholm.

BEAUMONT, J.-F. (2005). Calibrated imputation in surveys under a quasi-model-assisted approach, Journal of the Royal Statistical Society. Series B, 67, 445—458.

BEAUMONT, J.-F. and BOCCI, C. (2005). A Refinement of the Regression Composite Estimator in the Labour Force Survey for Change Estimates. SSC Annual Meeting, Proceedings of the Survey Methods Section, June 2005.

BELL, P. (2001). Comparison of Alternative Labour Force Survey Estimators. Survey Methodology, 27, 53—63.

CHEN, E.J. and LIU, T.P. (2002). Choices of Alpha Value in Regression Composite Estimation for the Canadian Labour Force Survey: Impacts and Evaluation. Methodology Branch Working Paper, HSMD-2002-005E, Statistics Canada.

DEVILLE J.-C. and SÄRNDAL C.E. (1992): Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376—382.

DEVILLE J.-C., SÄRNDAL C.E. and SAUTORY O. (1993): Generalized raking procedures in survey sampling. Journal of the American Statistical Association, 88, 1013—1020.

DJERF, K. (1997): Effects of Post-Stratification on the Estimates of the Finnish LFS. Journal of Official Statistics, 13, 29—39.

DJERF, K. (2004): Non-response in Time: A Time Series Analysis of the Finnish Labour Force Survey. Journal of Official Statistics, 20, 39—54.

ESTEVAO, V.M. and SÄRNDAL, C.E. (2002): The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling. Journal of Official Statistics, 18, 233—255.

FULLER, W.A. (1990). Analysis of repeated surveys. Survey Methodology, 16, 167—180.

FULLER, W.A. and RAO, J.N.K. (2001). A Regression Composite Estimator with Application to the Canadian Labour Force Survey. Survey Methodology, 27, 45—51.

GAMBINO, J., KENNEDY, B. and SINGH, M.P. (2001). Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation. Survey Methodology, 27, 65—74.

LENT, J., MILLER, S., CANTWELL, P. and DUFF, M. (1999). Effect of Composite Weights on Some Estimates From the Current Population Survey. Journal of Official Statistics, Vol. 15, No. 3, pp. 431—448.

SINGH, A.C., MERKOURIS, P. and WU, S. (1995). Composite Estimation by modified regression for repeated survey. ASA Proc., Surv. Res. Meth. Sec., 420—425.

SINGH, A.C. (1996). Combining information in survey sampling by modified regression. ASA Proc., Surv. Res. Meth. Sec., Vol. 1, 120—129.

SINGH, A.C., KENNEDY, B., WU, S. and BRISEBOIS, F. (1997). Composite Estimation for the Canadian Labour Force Survey . ASA Proc., Surv. Res. Meth. Sec., 300—305.

SINGH, A.C., KENNEDY, B. and WU, S. (2001). Regression Composite Estimation for the Canadian Labour Force Survey with a Rotating Panel Design. Survey Methodology, 27, 33—44.

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

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 519—524

CHALLENGES IN THE ESTIMATION AND QUALITY ASSESSMENT OF SERVICES PRODUCER PRICE INDICES

Markus Gintas Šova[1], John Wood[2], Ian Richardson[3]

ABSTRACT

The estimation of Services Producer Price Indices (SPPIs) faces challenges which are not encountered, or not encountered to the same extent, for goods Producer Price Indices. As a consequence, greater use is made of non-standard data collection methods, as in the use of model contract prices or unit value prices. Because of these challenges, estimating SPPIs and assessing the quality of SPPI estimates involves additional concerns: such as the effects of non-standard data collection methods; the suitability of indices for use in intermediate or final consumption; and the reliability of sampling frames for price collection. This paper describes how the UK Office for National Statistics is tackling the questions of estimation, measuring quality, setting quality standards and creating procedures to monitor and review the quality of SPPIs.

Key words: services producer price indices; estimation; quality assessment.

1. Introduction

Services Producer Price Indices (SPPIs) measure movements in the prices of services supplied by one business to another. For SPPI purposes, the definition of "business" here includes branches of central and local government. SPPIs are used by the Bank of England as an indicator of inflation for corporate services. They are used in the National Accounts as deflators for the Index of Services. Furthermore, Eurostat (the European Union's statistical agency) requires each EU member state to provide SPPI estimates for certain product classes. Eurostat has also set a timetable by which member states are expected to produce SPPI estimates for a range of additional product classes.

[1] Office for National Statistics, UK, e-mail: markus.sova@ons.gov.uk
[2] Office for National Statistics, UK, e-mail: john.wood@ons.gov.uk
[3] Office for National Statistics, UK, e-mail: ian.richardson@ons.gov.uk

The UK Office for National Statistics (ONS) has published SPPI estimates every quarter since 1996, originally under the name "Corporate Services Price Index" (CSPI). They are constructed as Laspeyres price indices (see Allen, 1975) which are chain-linked quinquennially. That is to say that every 5 years a new base period is introduced. This has the important effect of introducing new base period weights to reflect the changing importance of different services in the economy.

The nature of many service industries can make the estimation of their SPPIs much more challenging than the estimation of Producer Price Indices (PPIs) for goods. In the following section of this paper we examine the major issues and their causes, and we describe how ONS is tackling these challenges. We conclude with a discussion of ONS's plans to improve and monitor the quality of its SPPI estimates.

2. Specific issues affecting SPPI estimation

2.1. Classification of services

Internationally the classification of service provision is much less detailed than the classification of the production of goods. The recently revised European Classification of Products by Activity (CPA 2008) contains 124 pages of classifications. Of these, the production of goods covers 72 pages (not including agriculture and construction which cover a further 12 pages). In contrast, service provision covers only 31 pages (excluding wholesale and retail which cover a further 9 pages). This is in spite of the fact that service provision accounts for about 75% of the UK economy. To illustrate this phenomenon, group 56.3 "Beverage serving services" contains 1 class (56.30), which contains 1 category (56.30.1), which in turn contains 1 subcategory (56.30.10). As though to emphasise the point, each of these classification items is named "Beverage serving services". However, the corresponding group for goods, group 11.0 "Beverages" contains 7 classes, 16 categories and 18 subcategories. Such a lack of detail in the classification structure of service products can make it difficult to generate appropriate weights for items contributing to the index. If not resolved, the resulting SPPI is prone to bias. Consequently, ONS has developed its own in-house service product classification, through discussions with trade associations, to add the required level of detail. Unfortunately, because ONS's classification is not used by other National Statistics Institutes (NSIs), the international comparison of SPPIs is made more difficult.

2.2. Sampling issues

NSIs put considerable effort into maintaining business registers which are used as sampling frames for numerous surveys. In general, such registers give an indication of the broad activities in which establishments are involved, but without specifying the individual goods or services they produce, making them at best very inefficient sampling frames for goods PPIs and SPPIs. However, each EU member state is required to annually conduct a ProdCom *(Production Communautaire,* which is French for "Community Production") survey to produce national estimates of the production of goods, broken down into a highly detailed level of classification. The sample size for ONS's ProdCom survey is about 22000 reporting units (RUs). An RU is a sampling unit on ONS's Inter-Departmental Business Register (IDBR). Most RUs are whole businesses, however some large multi-site businesses are split into several RUs for data collection purposes. ONS uses its ProdCom sample as a sampling frame for PPIs for goods. Furthermore, the ProdCom data has turnover information which is used in the construction of PPI item weights. Unfortunately, there is no equivalent to ProdCom covering the service industries. Therefore ONS carries out a turnover survey every 5 years to fulfil the role of ProdCom for SPPIs. As the purpose of this turnover survey is solely to support SPPIs, it is on a much smaller scale than ProdCom, having a sample size of about 5000 RUs. Clearly, this can be limiting given that SPPI sample allocation is about 4500. The calculation of SPPI item weights is thus complicated by those businesses which contribute to the SPPI but are not sampled by the turnover survey. ONS has developed a method to construct such weights based on the UK ProdCom model devised by Chambers & Cruddas (1996). Such issues inevitably raise the question of how suboptimal the SPPI sample allocation is.

2.3. Product definitions

Ideally, every time a business supplies price data for an item which contributes to the SPPI, the data should be a price for a specific service with precisely specified requirements. These specifications should not change over time and should describe a service which was actually provided in the period of interest. This ensures that any differences in price data for a specific item are purely due to price movements, and not to quality changes in the service provided. Unfortunately, for many services this is not feasible. For example, in market research most pieces of work are unique. Consequently, non-standard methods may be required for price collection. Several such methods are described in the following paragraphs.

Quality adjustments are commonly used in estimating PPIs for goods in cases where a business has stopped production of one item, replacing it with a newer model with different specifications. Such circumstances occur frequently in the

production of computers and mobile telephones. The quality adjustment is calculated as the ratio of the price of the original item to the price of the new item. If there is at least one period where the prices of both items are available, then the quality adjustment can be calculated from real price data. If there is no such overlap, then the calculation of the quality adjustment will be at least partially subjective. Once the quality adjustment is calculated, it is applied every period to the price obtained for the new item, thereby removing the effect of the quality change. In cases where the precise service provided is different every period, a quality adjustment would have to be calculated every period. This could lead to the index movements being more a reflection of a statistician's judgement than of genuine price movements.

Some businesses provide a range of standard services, such as the translation of a one-page document from English into French, where the definition of a page refers to some precise number of letters and characters. The exact service may never be delivered in practice, due to document lengths never being precisely "one page". If the precise cost of each translation is proportional to its length, then a cost per unit length gives an automatically quality adjusted price for the service. It is necessary to periodically verify whether such services are still provided, as opposed to available but hardly ever required.

Hourly charge-out rates may be available for certain services, such as legal services. This method is relatively straightforward to implement, although if different grades of staff are involved in the service provision, then appropriate weights would need to be found in order to combine their hourly charge-out rates. A serious drawback of this method is that it does not take into account changes in efficiency. For example, suppose a service which once required 5 hours of work could now be completed in 3 hours due to the introduction of new technology and that the service provider does not change its hourly charge-out rate. The price of the service to the customer has fallen by 40%, but the index estimated by this method shows no price fall because the hourly charge-out-rate has not fallen.

For services such as computer programming, every contract has unique requirements. In such cases it may be possible to construct a model contract which somehow typifies the service provided. The model contract would need to be periodically reviewed to verify whether it still resembles actual contracts. This method tends to be resource intensive, both for the business and for the NSI.

Some services in some countries are regulated, such as telecommunications in the UK. It is not unusual for such regulators to collect a census of turnover and usage, possibly broken down by service product. This data allows the calculation of unit value prices, defined as the ratio of total turnover to total usage. These unit value prices can then be used to construct SPPIs. This method is easy to implement and the data may be available free of charge or at nominal cost. SPPIs constructed from unit value prices have zero standard error due to using data from a census of relevant businesses. However, the concept of quantity may be poorly defined. Furthermore, this method requires a level of homogeneity of service

products contributing to each unit value price in order to prevent potentially serious bias.

Even standard price collection procedures for SPPIs pose challenges for the estimation of standard errors. Although RUs can be selected at random with known inclusion probabilities, each business uses its own judgement in deciding for which specific service (or services) it will supply price data. There are thus no available inclusion probabilities for the specific services contributing to the index, making it impossible to apply design-based methods (comprehensively covered by Särndal *et al.*, 1992) for the estimation of standard errors. The use of hourly charge-out rates and model contracts further exacerbate this problem because the price data provided is not for a real service actually provided to a real client. ONS has recently produced estimated standard errors of SPPI movements using a model-based method originally developed for the estimation of standard errors of movement of its PPIs for goods (see Šova *et al.*, 2005, and Bucknall *et al.*, 2005).

2.4. Separation of business and retail components

For certain services it can be difficult to identify whether a client is a business. An example is passenger rail fares. Information from station ticket offices on the prices and specifications (*e.g.* seat class, destination, time of travel) of tickets sold does not distinguish between business and personal travel. If the distributions of travel patterns and ticket types are different for business and personal travel, and if these patterns experience different rates of inflation, then failure to correctly distinguish between business and retail transactions will bias the SPPI. This in turn will lead to biased estimates of Gross Domestic Product since SPPIs are used as deflators in the National Accounts. This issue can be addressed by appropriately weighting the prices of different specifications of ticket according to business travel patterns. However, much effort would be required to research the business travel patterns, and this research would need to be periodically repeated.

3. Plans to improve and monitor UK SPPI quality

The SPPI estimation issues described in this paper were identified in a quality assessment of ONS's live SPPI class indices which was conducted over 2006. Many of the assessment's recommendations are already in the process of being implemented. In particular, the recent availability of estimated standard errors of subcategory index movements has allowed an optimal sample allocation to be derived, and a plan has been put in place to move the current sample allocation towards optimality. Related to this is a review of those items whose weights are either very large, in which case they dominate their product subcategory index, or are very small, in which case they have a negligible impact on their product subcategory index. The specification descriptions of some items will be made

more precise in order to make quality changes more easily identifiable. Indices estimated using non-standard methods for price collection will be carefully monitored. The data flow for such indices will be reviewed where the use of spreadsheets or excessive typing is involved, in order to minimise the possibility of human error. It is hoped that at some point in the future resources will be made available for the ProdCom survey to be expanded to include service industries and service products. In addition to benefiting the sampling process, this will allow the introduction of annual chain-linking, making SPPIs more relevant to those services which are in rapid development. Finally, it will be necessary to periodically repeat the quality assessment in order to evaluate the effects of implemented changes, to cover those service product classes for which new SPPIs are being developed, and to identify any new issues which might have arisen in the meantime.

REFERENCES

ALLEN, R.D.G. (1975). Index Numbers in Theory and Practice. MacMillan.

BUCKNALL, R., SOVA, M., and WOOD, J (2005). Estimating Standard Errors of Movements in Producer Price Indices. Tenth GSS Methodology Conference, London, UK, June 2005. http://www.statistics.gov.uk/events/gss2005/agenda.asp

CHAMBERS, R., and CRUDDAS, M. (1996). Redesigning the ProdCom Inquiry. Office for National Statistics methodology report MQ021.

SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992) Model Assisted Survey Sampling. Springer-Verlag.

ŠOVA, M.G., WOOD, J., and BUCKNALL, R. (2005). Estimation of Laspeyres Producer Price Index Variances. Baltic-Nordic Workshop on Survey Sampling Theory and Methodology, conference digest, Vilnius, Lithuania, June 2005.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 525—537

ADMINISTRATIVE AND STATISTICAL REGISTERS IN BUSINESS STATISTICS OF UKRAINE

Olga A. Vasechko[1], Michel Grun-Rehomme[2]

ABSTRACT

The Ukraine is undergoing significant changes connected with fast internal development and increased participation in global markets. These changes call for improved business statistics; otherwise, public policies will bear increased risks. In particular, improvements are called for in the system of registers and sources of information used. This paper focuses on the role of administrative registers in business surveys. We analyze the statistical burden on Ukraine enterprises with respect to the dimensions of scale and dynamics. Approaches to systems of business registers are considered.

Key words: business statistics, operator, statistical burden, register, administrative file

1. Introduction

Modern business statistics should not only provide quality information on business conditions, but should also address the growing unwillingness of enterprises to spend time and resources for completing statistical questionnaires.

The spread of new kinds of business activity in the Ukraine and the occurrence of new types of businesses are creating a generation of top managers. These managers all too often, however, believe that the costs to complete the questionnaires exceed the benefits that might accrue to them in the future; failing to respond, however, creates future information problems for the enterprises. As a result, the Ukraine statistics office is now searching for new survey forms and methods to gather information. Guiding data collection policies of this effort are, on the one hand, reduction of response burden and the understanding of the

[1] Scientific and Technical Complex for Statistical Research; 3, Shota Rustaveli Str., Kyiv, Ukraine; E-Mail : O.Vasechko@ukrstat.gov.ua
[2] Université Paris 2, Ermes, UMR7181-CNRS ; 92 rue d'Assas, 75006 Paris, France ; E-Mail : grun@u-paris2.fr

necessity of co-operation with official statistics at the enterprises, and, on the other hand, the use of other, non-statistical sources of the information.

Other ministers outside the statistical agency are expected to endorse the effort to reduce and simplify responses in questionnaires for key statistics (e.g. business statistics). Better information from the statistical domain provides both useful market information for enterprises and better strategies for regulation. In this sense, the complex approach for integrating information from all surveys on business statistics into a unified whole is the most efficient and effective.

The paper begins with a discussion of the statistical burdens in administering business surveys in the Ukraine. Most notability, we find a problem with diversified manufactures. Given these problems, the paper proceeds to discuss three tools for reducing the information burdens, the importance of the statistical business registers with reference to the Ukraine, current conditions in and states organize principles. The paper concludes that different measures need to be explored to mitigate the statistical burden.

2. Statistical burden

Minimizing the statistical burden of and the development of administrative files to increase compliance and accuracy becomes new directions for official statistics of Ukraine.

The number of surveys forms the basis for the manager burden. It is the main cause for the negative attitude to official statistics. We find three types of principle problems in Ukraine business statistics:

- Surveys forms which incorrectly specify the kind of business activity or its legal form;
- Multiple kinds of activity which increases the number of requests for survey completion;
- Surveys of other (not statistical) official bodies (sometimes these surveys are as much as 50 % of the total number of surveys that enterprises receive).

To study this issue we use the information from the annual structural survey of the enterprises of Ukraine for 2000 and 2005. This allows us to find the average time for completing questionnaires, the change in completion time since the last survey period, the composition of time spent by economic activities (NACE), and the pattern of ownership for each of the enterprises. These surveys have been ongoing since 1999.

We find that the total time spent for completing all surveys categories of the enterprises in 2005 was 8,6 hours, a reduction of 1,4 percentage points from 2000. Except for E, J and P sections of NACE, the enterprises for all activities have reduced their time expensed. So we conclude that respondents better understand the format of the surveys, the questions, and the benefits of compliance. In

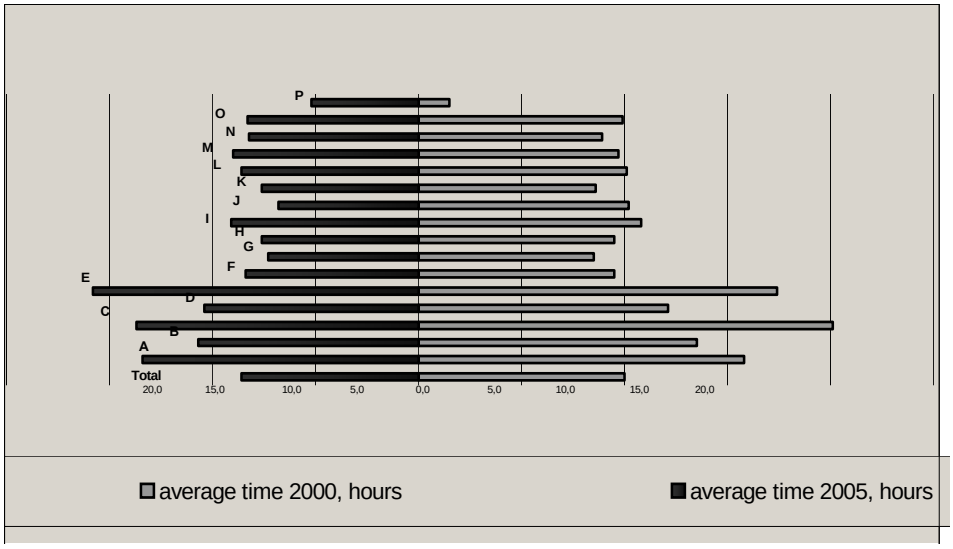
addition, there has been improvement in the work of regional statistical offices with the enterprises, reflecting the work done to update the registers.

Fluctuations of average completion time by economic activities for 2005 and 2000 are shown in Figure 1. They are the most significant for industrial enterprises (section C, D, E NACE), and agricultural ones (section A) and (section B).

The general expenses directly depend on a degree of complexity of the surveys which differ by business activities. The most problematic area is the situation for the industrial enterprises which are mostly diversified manufactures. These more diversified enterprises face the burden of answering many questions with more methodological issues. Concerning the agricultural enterprises, lower response rates for this category reflect the need of qualified experts to provide the primary data.

For different ownership, in 2005 survey expenses decreased strictly at the state enterprises, and less — at collective and private enterprises. It represents some subjective influence. In all countries a high level of subjective burden is inherent to big enterprises (Willeboordse, 1997). As such, small enterprises, which are mainly collective and individual, bear much more objective burden. So, on the average, small enterprises by their own estimation spent only 6,3 hours (8,0 hours — in 2000) for this questionnaire in 2005, which is essentially below average. And, in comparison with private enterprises the large state enterprises are more complex. As such, they need more time for filling out the questionnaire.

Figure 1: Average Time Spent by the Enterprises of Different Activity for the NACE Questionnaire for 2000 and 2005



3. Preconditions of administrative information sources used

The distribution of average indices of statistical burden on 12 base surveys, as measured by time, testifies to the essential expenses of the enterprises. So we need alternative information sources. Ukrainian business statistics use to address the burden on business Statistical Classification of Economic Activities (NACE), Business register, and General plan of accounting adapted to the international standards.

For many European countries it is logical to use the national classification of economic activities not only within the statistics but also within the other official bodies. In the Ukraine, NACE also has interdepartmental use. It gives the guarantee of correct realization of classifications and opportunities for interdepartmental data comparisons, through there is always some risk of incorrect coding of the enterprises to an activity because of administrative, tax, or social reasons that might bias an analysis.

The business register, even functioning ideally, cannot and should not exist independently from other registers. Multiple registers add information; administrative records create new opportunity for better information.

Thus, business statistics contains both the precondition for the use of administrative files and also the need of them. In the Ukraine, as well as in other countries, there is the need to develop a strategy for the systematic use of administrative sources. The use of the administrative sources is not a simple process, it has complexities that need to be considered and, in some cases, problems need to be solved. Taking into account experience of Ukraine, and statistical offices of such countries as France, Netherlands and Germany, we note the following:

1. As a rule, statistics using administrative declarations not destined for statistical purposes.
2. Low opportunities of integration with other data arise because of disagreements in definition of administrative and statistical units or in their identification codes.
3. Administrative declarations and the legal basis of administrative registers change and create problems at the forming of the time series.
4. Administrative data can contain errors because of possible misrepresentation by the respondent, for example, because of different taxation.
5. Sometimes, tax bodies accept declarations both as for the certain categories of tax bearers, and as a whole with the displaced fiscal year that does not coincide with a calendar year.
6. Always there is a problem of the control of answers' completeness in comparison with preliminary determined population.
7. There is a necessity of the maintenance of confidentiality, professional work and correction of the micro-files usually are carried out by experts

— statisticians on the basis of their own view of outliers and never transmitted to tax services.

Thus, administrative files can exist in the form of a database of the administrative register and in the databases of economic files which do not contain statistical observations.

4. Business register

The situation in Ukraine has developed in such a manner that the opportunity for efficiency of business statistics as whole appreciably depends on the presence of the statistical business register. The administrative register functions within the frames of the general state system and it is in effect interdepartmental. It is officially named as the Unified State Register of Enterprises and Organizations of Ukraine, but in short, we hereafter refer to it as the State Business Register.

The State Business Register is a database that comprises the administrative files concerning registration, location, and cessation of operators in the country. At the same time there is a set of problems which could make difficult to use the administrative register for statistical purposes. These problems are:

1. The state is interested in fuller database that covers all operators in the administrative register, through not, everybody, according to the Economic Code of Ukraine, is an enterprise or physical operator, and, therefore, not everyone should be included.
2. The State Business Register, which is interdepartmental, is not solely a statistical tool; and entering of any changes into it is carried out according to the law, instead of according to needs of statistics. It makes it conservative from the point of view of an opportunity of data actualization, although an official statistics always has sources for it.
3. The norms that define the order of operators' creation and state registration, and corresponding changes are not coordinated from the statistical point of view. As such, this complicates the formation of the target population. The changes completed in February, 2005 in the Economic Code of Ukraine are the example of it. According to them the structural (subsidiary) units of the enterprises are not the subjects of economic activities and, accordingly, are not a subject of the state registration.
4. The decision to transfer the register to another authority body or form a new administrative register may not be a fatal problem but it creates some problems for statistical systems.

Within the frames of business statistics, the statistical business register is a subsystem by means of which administrative files are transformed into a more suitable basis for the statistical data. The difference between the entering and processing of information for the business register depends on the degree of involvement of the national statistical office in the work to improve the

administrative registers. The experience of the Ukraine shows active work of the national statistical office on register improvement has allowed pulling together large amounts of administrative and statistical information for many observation units and now information for the registers mostly coincide with each other, that is, the information is more accurate.

We can control and reduce statistical burden using for each statistical unit in the register a special attribute. Other important opportunity of the register in this sense is its connection with base of metadata. The beginning of metadata base creation in business-statistics of Ukraine provides construction of such system, each variable in which has one universal (conceptual) basis and a mobile multiform attributive part. Such framework allows connecting among themselves the register and statistical observation through autonomy of each variable of metadata base. Thus, there is an opportunity of management of statistical burden not only by total number of surveys, but also by number of variables for each of them.

5. Business register quality

By virtue of the specificity of the demography of the enterprises, it is necessary to estimate the size of deviations and discrepancies that can take place. For the elimination and reduction of disagreements between ideal and expected target population, procedures have been implemented such as:

- Use of system of universal codes (numerous) which are for ever fixed to the enterprise, to trace transitions from one category of units to another,
- Integration of two or more data bases concerning the enterprises, and their careful comparison,
- Tracking demographic processes concerning each unit in time and saving records concerning inactive units with the attribute,
- Periodic sample survey for precision of the units' characteristics in the statistical register,
- The creation of links between statisticians and large enterprises.

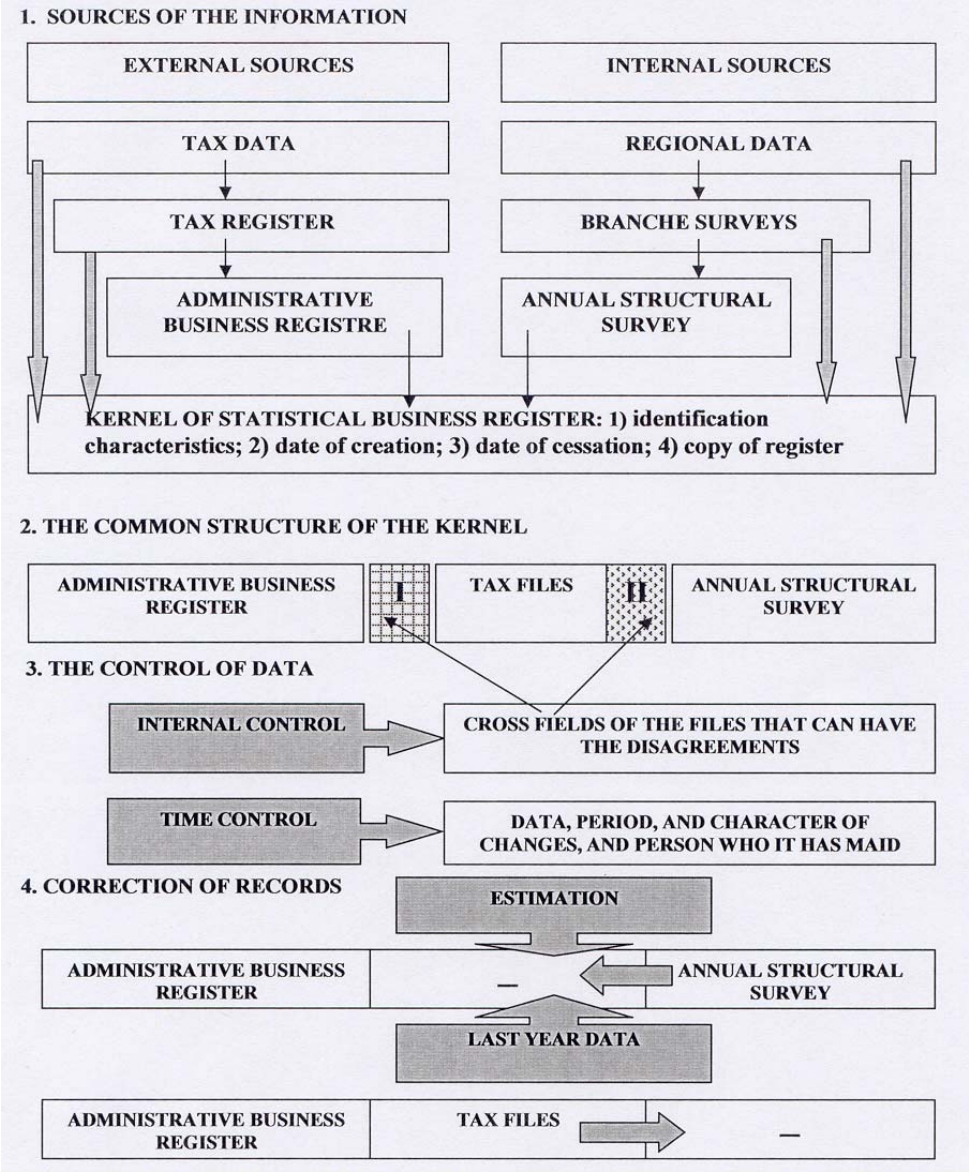
The question of how much each of the above procedures is implemented for the business register depends on depth of existing problems. For this it is important to have the block of the information which allows estimating a quality level of some sections in the statistical register. Thus at the level of the units it is necessary to record the last date of data actualization, date of modification, character, and the source of the actualization. For national statistical practice it is essentially important that the administrative register preserve information concerning the updating of records on the enterprises and their history which now is not stipulated.

The parity of different sources for actualization of the business register is complex. Indeed, as one source can be used for updating different blocks of the

information and on the contrary, the identical information can act from different sources.

The idea of the mechanism of the statistical register updating in Ukraine can be examined in simplified form with an example of four levels (See Figure 2).

Figure 2. Levels of the Statistical Business Register Actualization



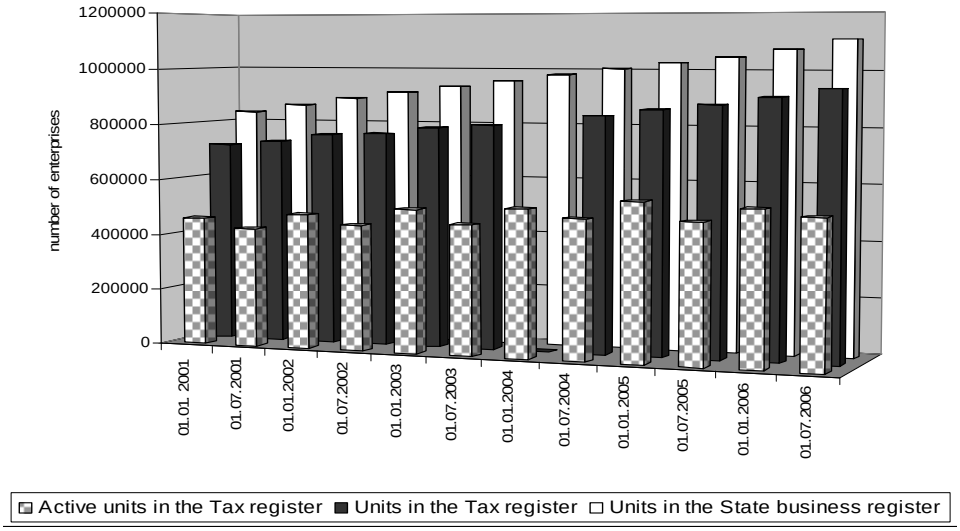
Unlike the administrative register, the statistical register is a more flexible tool which uses various data. It contains information from different sources, that is, to say, by double stream: directly or indirectly. The efficiency of this approach is achieved using cross fields of files. Kinds of the control are shared on: 1) internal control and 2) time control. The internal control provides finding-out of the size and the reason(s) of these disagreements. The time control fixes all external attributes made to any record.

The coordination of disagreements can be carried out in view of the lack of records. For example, if there is no tax data, information may be gained by using expert opinion to estimate taxes using corresponding information of the structural survey and tax data from the previous year. If there is no data in the structural survey due to non-responses of the last survey, then it is possible to use past tax files relative to information available.

Changes to the state register database, recorded with semi-annual periodicity, testifies to small reductions. In addition, during last two years, the level of disagreements between total of objects in it and number of subjects of managing which is given by tax administration has remained relatively stable, and has the number of active units in the tax database (See Figure 4) (State Statistics Committee of Ukraine, 2001—2006). Disagreements between data of the State Business Register and the Tax Administration can testify not only about the lacks of the State Business Register, but also about the problems in the tax registers. In addition: another, than in statistics, approach to definition of a registration unit, absence of data concerning structural items of the enterprises and hierarchical connections between them. On the other hand, the constancy of a level of these deviations speaks about the certain border of quality to overcome which the administrative register can not. It is the important finding which testifies necessity to create of statistical business register.

Tracking of changes of the register during a year enables us to analyze what might be expected at the beginning of the year and yield estimates of the real activity of the enterprises. Also, we find the number of active units of the register in the middle of year is always smaller in comparison with the beginning of the period (Figure 3).

Figure 3. Number of Operators According to Different Registers in the Ukraine



We have the same finding analyzing the economic activities of operators. There are essential disagreements in number of enterprises according to activities that provide different registers. The number actually active unit is obviously smaller for all activities, and especially for trade. This conclusion requires finding and using statistical methods for register quality monitoring and the target population.

Statistical method of quality monitoring is carried out using special sample surveys focusing on the precision of an available database. In the international practice such enterprises are surveyed in 6—9 months after registration. For Ukraine where the share of small enterprises with very short period of existence (one is high-), such surveys are most likely needed every quarter. There is a risk of not finding desired information on enterprises at the end of the year. This situation presents the likelihood of shadow sectors in the economy. This question is an actual one, especially taking into account correlation between the short period of existence of such enterprises and the large volume of turnover. This volume of turnover sometimes equals the activity level of large enterprises with significant industrial output.

Another method of register quality improving is connected with sample surveys; it was offered at the European Conference of Quality in Official Statistics (Grun-Rehomme, and Vasechko, 2004).

6. The register of operators — physical persons

Physical persons, who are engaged in business activity, now have essential influence on the development of small businesses and business as a whole. Except for the production of goods and services that are value adding, a significant number of these new workplace activities are an essential factor in the economic and social stability of the country (See Tab. 1, State Statistics Committee of Ukraine, 2001—2006).

Table 1. Physical Operators in the Economy of Ukraine (% to the total economy)

NACE	Number of Entities	Number of Employments	Turnover Volume
Total economy	82,2	23,8	5,5
A	34,0	2,8	2,6
B	82,9	15,4	6,2
C, D, E	54,7	3,7	0,8
F	29,0	3,4	2,2
G	91,4	68,8	9,4
H	56,7	19,2	12,7
I	90,7	18	4,2
K	42,6	6,3	2,9
M	45,7	4,2	3
N	74,8	17,6	7,5
O	83,9	34,3	10,7

There is a certain specificity of physical persons as an object of statistical observation. On the one hand, it consists in their growth at fast rates and their influence. On the other hand, there is an absence of detailed book keeping on them. There is also an absence of services which could prepare such information and provide it to statistical bodies. In some countries with developed markets, where institutions of business and physical persons have existed for a long time, the statistical register is a database with both legal persons, and operators — physical persons.

For Ukraine physical operators are a new phenomenon, and during formation of the State register there was no opportunity for discussion of a common register creation. This question has arisen recently. As such, statistical observation of physical persons in the Ukraine has only been conducted by official statistics since 2002.

Even with limited information, we find, nonetheless, that operators — physical persons are a very dynamic population. Annually the number of them has increased approximately 15 %, mid-annual number of workers has increased by about 18 %. In short, fast rates of trade and services development are evident.
There is the legal base for realization of enterprise activity by physical persons in Ukraine and its registration and realization of the state statistical observation over them.

The co-ordination among the register of legal persons and the register of physical persons is very important. They can work rather independently, but special tools of co-ordination of their actions should be stipulated to improve accuracy and increase the availability of information. It is predetermined by that despite the lack of specificity between each of these databases that their units have the essential common characteristics as a belonging to operators, the common principle of classification of activity, the common basic variable of business statistics – the turnover, the common contribution to statistical data synthesis of small business, and an opportunity for some entities to pass from one category in another.

7. The register of groups of the enterprises

The statistics of groups of the enterprises are the part of business-statistics and it is based on the common principles, concepts, normative documents and tools. Though business registers (of corporations) exist in many countries, a common problem exists in creating the register of enterprise groups. Invariably, it is necessary gain co-operation and avenues of mutual exchange between the corresponding ministries to develop this tool. Almost always, the main task consists in creating a legislative base to support such an effort, protocols for information exchange, and in the establishment of standardized procedures for units of observation unit.

According to the Law « About the holding companies » the holding company gets the status of the legal person from the date of its state registration in the Administrative Register of the Holdings of Ukraine which is an integral part of the Unified Administrative Register. The Register of the state corporate rights administered by the Fund of the state property of Ukraine may be used as additional source of the information.

It is necessary, however, that group unit is useful to enterprises; but at the same time, given the instability in the formation of groups, flexibility also needs to be part of the grouping process - as the group is the central unit of observation and analysis. Therefore, some limitations remain as to the information process.

8. Conclusion and acknowledgements

Thus, for improvement in the statistical production process in Ukraine, different measures need to be explored to mitigate the statistical burden. Good horizontal coordination between different authorities is needed to avoid collection of similar or even identical basic information from enterprises. The use of data for statistical purposes needs to follow strict confidentiality rules. Translation of similar user requirements into integrated surveys reduces compliance costs, increase information, and improves data accuracy. Clearly, the use of the system of business registers is beneficial. And, recognition of complex problems and a unified approach to their solution considerably improves results.

This paper describes part of the results of the research activity in official business statistics between the Centre of Econometric of University Paris 2 and the Scientific and Technical Complex for Statistical Research in Kyiv.

REFERENCES

EUROSTAT DOC. (1998): Legal texts relating to the European business statistical system

EUROSTAT DOC. (2002): Assessment of the Quality in Statistics. Draft Quality Measurement and Reporting Framework. 29.2002

GRANQUIST, L., and KOVAR, J.G. (1997): Editing in survey data: how much is enough?

THE SURVEY MANAGEMENT AND PROCESS QUALITY, Ed. Wiley, New York, 415—435

GRUN-REHOMME, M., and VASECHKO, O. (2004): Quality Measurement of a register for Structural Business Survey: Application to Ukrainian Data, European Conference on Quality and Methodology in Official Statistics, Germany, Mainz

INSEE Méthodes, Accounting standards, businesses and statistics (n.74, 1997)

INSEE Méthodes, Normes comptable, entreprises et statistiques (n.74—75, 1997)

MONOGRAPHS OF OFFICIAL STATISTICS (2003), Work session on statistical data confidentiality, part 3, Office for Official Publications of the European Communities, Luxemburg

RIVIERE, P. (1999): Qualité et statistique, Courrier des statistiques, 90, 47—58

STATE STATISTICS COMMITTEE OF UKRAINE, Statistical Yearbook of Ukraine (2001—2006), Kyiv

STATE STATISTICS COMMITTEE OF UKRAINE, The statistical bulletin of the State register of the enterprises and the organizations of Ukraine (2001—2006), Kyiv

THOMAS, R. (1996): Statistics as Organizational Products, Social Research Online, Vol. 1, 3, 13, http://www.socresonline.org.uk/socresonline/1/3/5.html

UNITED NATIONS, Links between Business Accounting and National Accounting (2000)

WILLEBOORDSE, A. (1997): Handbook on design and implementation of business surveys, Eurostat, Luxemburg

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 539—546

UNBIASED NONLINEAR ESTIMATORS OF A FINITE POPULATION TOTAL: DO THEY EXIST?

Jan Wretman[1]

ABSTRACT

Unbiased nonlinear estimators of a finite population total exist if and only if the sampling design is a nonunicluster design with strictly positive inclusion probabilities.

Key words: Survey sampling; Finite population; Unbiased estimator; Linear estimator; Nonlinear estimator; Unicluster design.

1. Introduction

Estimators of a population total that are suggested in sampling textbooks are mostly *linear* estimators, as this term is will be defined in Section 2 below. Also, most theoretical results concerning the foundations of finite population inference (such as minimum variance estimation and admissibility) deal with *linear* estimators; see, e.g., Cassel, Särndal, and Wretman (1977) and Chaudhuri and Vos (1988).

The question motivating the present paper is: Considering that *nonlinear* unbiased estimators of a population total are seldom or never mentioned, do such estimators exist at all?

The answer to be given is that *nonlinear* unbiased estimators of a population total exist if and only if the sampling design is a nonunicluster design with all inclusion probabilities strictly positive. The approach of the present paper is purely design-based. A preliminary version was presented at the symposium "Clinical Trials, Games or Power?" at the Department of Statistics, Lund University; see Wretman (1999).

[1] Department of Statistics, Stockholm University, S-106 91 Stockholm, Sweden, e-mail: jan.wretman@stat.su.se

2. Some terminology

We consider a *finite population* $U = \{1, 2, \ldots, k, \ldots, N\}$, consisting of N *elements* labelled $k = 1, 2, \ldots, N$. With each element $k \in U$ is associated a real number y_k, assumed to be completely unknown at the time when the survey is being planned. Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$. It is assumed that the population characteristic of interest is the *population total*,

$$t_{\mathbf{y}} = \sum_{k \in U} y_k$$

A *sample* s is a nonempty subset of U, and S is the set of all 2^N-1 such subsets. It is assumed that a sample will be selected using a specified *sampling design* p which gives each sample $s \in S$ the known probability $p(s)$ of being selected. For a given design p, let S_p denote the set of all $s \in S$ for which $p(s) > 0$. Thus,

$$\sum_{s \in S} p(s) = \sum_{s \in S_p} p(s) = 1$$

The probability that a specified element $k \in U$ will be selected, under a given design p, is given by the (first order) *inclusion probability*

$$\pi_k = \sum_{s \ni k} p(s)$$

where the summation is over all samples s that contain the element k. Thus, π_k is the probability of obtaining a sample containing the element k. The inclusion probabilities are assumed known for all elements in the population.

A design p is called a *unicluster* design if each element $k \in U$ is included in at most one sample $s \in S_p$, that is, if all samples in S_p are pairwise disjoint. An example of a unicluster design is one-stage cluster sampling, where the population is divided into a number of nonoverlapping clusters and a sample of clusters is selected, after which data are collected from all elements in the selected clusters. With this terminology, also a *census* (that is, a design p such that $p(U) = 1$ and $p(s) = 0$ for all $s \neq U$) is a special case of a unicluster design. A *nonunicluster* design is a design which is not a unicluster design.

After the sample s has been selected, the constants y_k will be observed for all $k \in s$, and the observed *sample data* will be used for estimating the unknown population total $t_{\mathbf{y}}$. An *estimator* $\hat{t}$ is a real function of s and $\mathbf{y}$, such that $\hat{t}(s, \mathbf{y})$ depends on $\mathbf{y}$ only through those y_k for which $k \in s$. The *expected value* of an estimator $\hat{t}$ is defined, for a given design p and a given population vector $\mathbf{y}$, as

$$E_p(\hat{t};\mathbf{y}) = \sum_{s\in S_p} p(s)\hat{t}(s,\mathbf{y})$$

An estimator $\hat{t}$ is said to be *unbiased* for $t_{\mathbf{y}}$ under a given design p, if $E_p(\hat{t};\mathbf{y}) = t_{\mathbf{y}}$ for any $\mathbf{y}\in R^N$. Following Godambe (1969), a *linear* estimator is defined as an estimator that can be written, for all $s\in S_p$ and all $\mathbf{y}\in R^N$, as

$$\hat{t}(s,\mathbf{y}) = w_{0s} + \sum_{k\in s} w_{ks} y_k$$

where the coefficients w_{0s} and w_{ks} may depend on s but not on $\mathbf{y}$. If $w_{0s} = 0$ for all $s\in S_p$, the estimator is said to be *linear homogeneous*. A *nonlinear* estimator is an estimator which cannot be written on the on the form above.

An example of a linear homogeneous estimator which is unbiased for $t_{\mathbf{y}}$ under any design p with $\pi_k > 0$ ($k = 1, 2,\ldots, N$) is the *Horvitz-Thompson* estimator $\hat{t}_{HT}$, defined as

$$\hat{t}_{HT}(s,\mathbf{y}) = \sum_{k\in s} \frac{y_k}{\pi_k}$$

for all $s\in S_p$ and all $\mathbf{y}\in R^N$. Here $w_{0s} = 0$ and $w_{ks} = 1/\pi_k$ for $k\in s$.

Another example of a linear homogeneous estimator is the well-known ratio estimator $\hat{t}_R$:

$$\hat{t}_R(s,\mathbf{y}) = t_{\mathbf{x}} \frac{\sum_{k\in s} y_k/\pi_k}{\sum_{k\in s} x_k/\pi_k}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ is a vector of values on an auxiliary variable, x, assumed to be known for the whole population, and $t_{\mathbf{x}} = \sum_{k\in U} x_k$. Although this ratio estimator contains a ratio of two Horvitz-Thompson estimators, it is still a *linear* estimator (linear in the y-values) according to the definition above. We have

$$w_{0s} = 0 \quad \text{and} \quad w_{ks} = \frac{t_{\mathbf{x}}}{\pi_k \sum_{\ell\in s} x_\ell/\pi_\ell} \quad \text{for } k\in s.$$

Still another example of a linear estimator, in the sense above, is the generalized regression estimator.

3. Result

Nonlinear estimators of a population total $t_{\mathbf{y}}$ are seldom (maybe never) considered in standard texts on survey sampling. One might wonder: Do unbiased nonlinear estimators of $t_{\mathbf{y}}$ exist at all? If yes, do they exist only for special designs? A kind of answer can be given as follows.

Result: For any design p, an unbiased nonlinear estimator of the population total $t_{\mathbf{y}}$ exists if and only if p is a nonunicluster design with inclusion probabilities $\pi_k > 0$ ($k = 1, 2, \ldots, N$).

Thus, for many sampling designs, nonlinear unbiased estimators of the population total do really exist, although they have not been discussed in the literature. A simple example will be given in Section 5. Note that in the present paper interest is focused on whether estimators exist that are both nonlinear and unbiased. It is not conjectured that such estimators might be more efficient than linear ones. However, nonlinear estimators that are *not* exactly unbiased have been discussed in he literature, for example, regression estimators based on nonlinear models.

4. Proof of the result

For convenience, we first prove a lemma from which the result in Section 3 above follows easily.

Lemma: For any unicluster design p with inclusion probabilities $\pi_k > 0$ ($k = 1, 2, \ldots, N$), an estimator $\hat{t}$ is unbiased for $t_{\mathbf{y}}$ if and only if it can be written on the form

$$\hat{t}(s, \mathbf{y}) = w_{0s} + \hat{t}_{HT}(s, \mathbf{y}) \tag{4.1}$$

for all $s \in S_p$ and all $\mathbf{y} \in R^N$, where w_{0s} does not depend on $\mathbf{y}$, and $\sum_{s \in S_p} p(s) w_{0s} = 0$.

Proof of the lemma: The ”if” part of the lemma is trivial and need not be proved here. Now, suppose $\hat{t}$ is unbiased for $t_{\mathbf{y}}$. Then we have to show that $\hat{t}$ can be written on the form (4.1) for all $s \in S_p$ and all $\mathbf{y} \in R^N$, where w_{0s} does not depend on $\mathbf{y}$, and where $\sum_{s \in S_p} p(s) w_{0s} = 0$.

For any any $s \in S_p$ and any $\mathbf{y} \in R^N$ we can write

$$\hat{t}(s,\mathbf{y}) = \underbrace{[\hat{t}(s,\mathbf{y}) - \hat{t}_{HT}(s,\mathbf{y})]}_{a(s,\mathbf{y})} + \hat{t}_{HT}(s,\mathbf{y}) = a(s,\mathbf{y}) + \hat{t}_{HT}(s,\mathbf{y})$$

We want to demonstrate that $a(s, \mathbf{y})$ can play the role of w_{0s}, which means that $a(s, \mathbf{y})$ does not depend on $\mathbf{y}$, and that $\sum_{s\in S_p} p(s)a(s,\mathbf{y}) = 0$.

From the fact that both $\hat{t}$ and $\hat{t}_{HT}$ are unbiased for $t_{\mathbf{y}}$ it follows that, for any $\mathbf{y}\in R^N$,

$$\sum_{s\in S_p} p(s)a(s,\mathbf{y}) = \sum_{s\in S_p} p(s)[\hat{t}(s,\mathbf{y}) - \hat{t}_{HT}(s,\mathbf{y})] = E_p(\hat{t};\mathbf{y}) - E_p(\hat{t}_{HT};\mathbf{y}) = 0 \quad (4.2)$$

It remains to see that $a(s, \mathbf{y})$ does not depend on $\mathbf{y}$. Suppose, on the contrary, that $a(s, \mathbf{y})$ *does* depend on $\mathbf{y}$. Then there must be at least one sample s_0 and two vectors $\mathbf{y}_1 = (y_{11}, y_{12}, \ldots, y_{1N})$ and $\mathbf{y}_2 = (y_{21}, y_{22}, \ldots, y_{2N})$ such that

$$a(s_0, \mathbf{y}_1) \neq a(s_0, \mathbf{y}_2) \quad (4.3)$$

Now, let $\mathbf{y}_1^* = (y_{11}^*, y_{12}^*, \ldots, y_{1N}^*)$ and $\mathbf{y}_2^* = (y_{21}^*, y_{22}^*, \ldots, y_{2N}^*)$ be two vectors defined by

$$y_{1k}^* = \begin{cases} y_{1k} & \text{for } k \in s_0 \\ 0 & \text{for } k \notin s_0 \end{cases}; \qquad \text{and} \qquad y_{2k}^* = \begin{cases} y_{2k} & \text{for } k \in s_0 \\ 0 & \text{for } k \notin s_0 \end{cases}$$

Because p is a unicluster design, and because $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$ have identical components as soon as $k\notin s_0$, it follows, using (4.2), that

$$\underbrace{\underbrace{\sum_{s\in S_p} p(s)a(s,\mathbf{y}_1^*)}_{=0} - \underbrace{\sum_{s\in S_p} p(s)a(s,\mathbf{y}_2^*)}_{=0}}_{=0} =$$

$$p(s_0)[a(s_0,\mathbf{y}_1^*) - a(s_0,\mathbf{y}_2^*)] + \sum_{\substack{s\in S_p \\ s\neq s_0}} p(s)\underbrace{[a(s,\mathbf{y}_1^*) - a(s,\mathbf{y}_2^*)]}_{=0}$$

Since $p(s_0) > 0$, we have that $a(s_0,\mathbf{y}_1^*) = a(s_0,\mathbf{y}_2^*)$, and since $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$ have the same components as $\mathbf{y}_1$ and $\mathbf{y}_2$ for all $k\in s_0$, it also follows that

$$a(s_0, \mathbf{y}_1) = a(s_0, \mathbf{y}_2) \quad (4.4)$$

From the assumption that $a(s, \mathbf{y})$ *does* depend on $\mathbf{y}$, we have thus arrived at both (4.3) and (4.4) which is a contradiction. We conclude that $a(s, \mathbf{y})$ *does not* depend on $\mathbf{y}$. Hence, under any unicluster design p, an unbiased estimator $\hat{t}$ of $t_{\mathbf{y}}$ can always be written as

$$\hat{t}(s,\mathbf{y}) = w_{0s} + \hat{t}_{HT}(s,\mathbf{y})$$

where w_{0s} does not depend on $\mathbf{y}$, and $\sum_{s\in S_p} p(s)w_{0s} = 0$, and we have seen that a possible choice of w_{0s} is

$$w_{0s} = \hat{t}(s,\mathbf{y}) - \hat{t}_{HT}(s,\mathbf{y})$$

Proof of the result in Section 3: First, suppose p is a nonunicluster design with all $\pi_k > 0$. Then we have to demonstrate that there is at least one unbiased nonlinear estimator of t_y. Since p is a nonunicluster design there are at least two nondisjoint and nonidentical samples, say s_1 and s_2, with $p(s_1) > 0$ and $p(s_2) > 0$. Suppose for simplicity that the population elements are labelled so that the element labelled $k=1$ is a member of both s_1 and s_2. Then the estimator $\hat{t}_0$, defined as follows, is an example of a nonlinear estimator which is unbiased for t_y under p:

$$\hat{t}_0(s,\mathbf{y}) = \begin{cases} \hat{t}_{HT}(s,\mathbf{y}) + p(s_2)\exp(y_1) & \text{for } s = s_1 \\ \hat{t}_{HT}(s,\mathbf{y}) - p(s_1)\exp(y_1) & \text{for } s = s_2 \\ \hat{t}_{HT}(s,\mathbf{y}) & \text{for } s \neq s_1 \text{ and } s \neq s_2 \end{cases}$$

Now, suppose that a nonlinear estimator exists, which is unbiased for t_y under a given design p with all $\pi_k > 0$. That p must then be a nonunicluster design follows from the preceding lemma, and the proof is completed.

The technique of this proof comes from Lanke (1973), who used the same technique for a different purpose (namely, for proving that a minimum variance unbiased estimator of t_y exists if and only if the sampling design is a unicluster design with all $\pi_k>0$).

5. An example of a nonlinear unbiased estimator

For illustrative purpose, a simple example of a nonlinear unbiased estimator will be given, constructed in accordance with the principle described in Section 4 above. A sample of fixed size, $n = 2$, is to be selected by simple random sampling from a population with $N = 4$ elements. The six possible samples and the associated values taken by the nonlinear estimator, $\hat{t}$, are given in the following table:

s	$p(s)$	$\hat{t}(s,\mathbf{y})$
{1, 2}	1/6	$2(y_1+y_2) + \exp(y_1)$
{1, 3}	1/6	$2(y_1+y_3) - \exp(y_1)$
{1, 4}	1/6	$2(y_1+y_4)$
{2, 3}	1/6	$2(y_2+y_3)$
{2, 4}	1/6	$2(y_2+y_4)$
{3, 4}	1/6	$2(y_3+y_4)$

It is easily seen from the table that $\hat{t}$, although nonlinear, is unbiased for $t_{\mathbf{y}}$.

6. Another way of expressing the result

Remembering that a sampling design admits an unbiased estimator of $t_{\mathbf{y}}$ if and only if all $\pi_k > 0$, we can also express the result above in the following way:

- If p is a design with $\pi_k = 0$ for at least one $k \in U$, then no unbiased estimator of $t_{\mathbf{y}}$ exists.
- If p is a *unicluster* design with all $\pi_k > 0$, then the class of unbiased estimators of $t_{\mathbf{y}}$ consists of exactly those estimators that can be written on the form (4.1). Thus, all unbiased estimators of $t_{\mathbf{y}}$ are *linear* estimators (and linear estimators that are closely related to the Horvitz-Thompson estimator).
- If p is a *nonunicluster* design with all $\pi_k > 0$, then both *linear* and *nonlinear* unbiased estimators of $t_{\mathbf{y}}$ exist.

REFERENCES

CASSEL, C.M., SÄRNDAL, C.E., and WRETMAN, J.H. (1977). *Foundations of Inference in Survey Sampling*. New York: Wiley. [Reprint Edition (1993). Malabar, FL: Krieger]

CHAUDHURI, A. and VOS, J.W.E. (1988). *Unified Theory and Strategies of Survey Sampling*. Amsterdam: North-Holland.

GODAMBE, V.P. (1969). Some Aspects of the Theoretical Developments in Survey Sampling. In N.L. Johnson and H. Smith (Eds.): *New Developments in Survey Sampling*. New York: Wiley Interscience.

LANKE, J. (1973). On UMV-Estimators in Survey Sampling. *Metrika* 20, 196—202.

WRETMAN, J. (1999). On the Existence of Nonlinear Estimators in Survey Sampling. In *Proceedings of the Symposium "Clinical Trials, Games or Power?" Lund, June 1999*. Department of Statistics, Lund University.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 547—561

ESTIMATION OF UNPAID WORK IN POLISH HOUSEHOLDS

Ilona Błaszczak-Przybycińska[1]

ABSTRACT

The paper presents empirical estimates of unpaid household production in Poland. The following home services, not included in macroeconomic accounts are evaluated: household upkeep, food preparation, making and care of textiles, child care, adult care and volunteer work. The data came from the Time Use Survey 2003-2004. Two approaches were employed: the market cost method and the alternative cost method. The total value of household production was equal to 30.0% of GNP.

Key words: Polish households; Estimation of unpaid work; Time use survey; Household work.

1. Introduction

The estimation of work done by household members to satisfy their own needs has its own history in Poland. The first attempts to estimate monetary value of this type of work were undertaken in the 1970s. An expanded estimation of household work was undertaken in the 1980s by the Central Statistical Office and the Polish Academy of Science (GUS and PAN). All this work was undertaken in the context of measuring the level of living.

According to some authors the monetary value of household work should be, as a matter of necessity, incorporated in the measure of the level of living. They have proposed to calculate "extended consumption" in which the bulk of private and public consumption was expanded to include the value of unpaid goods and services produced and consumed by households [Szczerbińska, 1980].

All these estimations were conducted in the centrally planned economy. It was difficult to make adequate empirical calculations in such circumstances, as the

1 The Warsaw School of Economics, Institute of Statistics and Demography, *e-mail*: iblasz@sgh.waw.pl

prices of goods and services used in the estimates were not based on market prices.

The last estimation of household work in Poland based on the Time Use Survey conducted by the Central Statistical Office captured one year in the 2003-2004 period, i.e. more than a decade after advent of a market economy. Two approaches were employed: the market cost method and the alternative cost method. The estimation based on the market cost method used average market wages for professions corresponding to different domestic activities. In the alternative cost method, the category of an average wage per hour was used.

Many empirical results, for instance the ratio of country household work value to Poland's GDP are comparable with similar estimates for other European countries.

2. Data source

The Time Use Survey 2003-2004 was the source of information about the time spent on household work. Just as in the previous time use surveys conducted in Poland, a two-stage sampling scheme was used. Census areas or combined census areas (if the minimum number of dwellings was not ensured) were the first stage sampling units and dwellings were the second stage sampling units. The sampling frame was provided by the BREC 99 system: a set of statistical regions and census areas designed for national census purposes but it is also used for sample surveys. There were days sampled for each dwelling. All days between June 1st, 2003 to May 31st, 2004 were represented (366 days). Each person kept records of the activities performed on two different days: one during the week and one during the weekend.

Six household types were distinguished by the main source of income: employees, employees-farmers, farmers, self-employed, retirees and invalid-pensioners and those living on unearned sources other than invalid-pension and retirement. Respondents were aged 15 and above. There were 20,264 persons aged 15 and above and 10,256 households in a sample. The share of individuals who refused to participate in the survey was 19.6 %. Survey results were generalised onto the entire population using the results of the 2002 census data (according to the socio-demographic structure of the population).

The time Use Survey 2003/2004 was harmonized with the European Time Use Survey [Błaszczak-Przybycińska, 2006]. Activities were registered in diaries in 10-minutes intervals. There was a list of more than 200 activities within ten groups.

As far as the estimation of household work is taken into account, the main group of activities was group "household and family care". However – according to the actual classification of activities in harmonized European Time Use Surveys — there are some activities within this group which must be excluded from the estimation (e.g. personal services as barber services which should rather be

treated as personal care activities). On the other hand, some activities outside this group were taken into account in the estimation process (e.g. household care travel, volunteer work done in the form of domestic work to benefit other households). Finally, 47 household activities within 5 groups were taken into account.

The second source of information was a sample survey on wages by profession conducted by the Central Statistical Office [GUS, 2003]. This survey provided information on average hourly wages for the specified professions. The survey was started in 1996 using a two stage sampling scheme. Enterprises are the first stage items and employees are the second stage items. In 2002 the first stage sample consisted of 22.9 thousand enterprises and 700 thousand employees. The hourly wages in the October 2002 survey were corrected by the wage rate to derive May 2004 wages.

The wages per hour for each type of home services were defined separately on the basis of the hourly wage for professions corresponding to different types of domestic work. This usually meant the average wage for several professions corresponding to the specified home activity.

3. Estimation methods

The estimation of household work was made using two approaches: the market cost method and the alternative cost method.

The main assumption in the market cost method is that household work done by household members could be done by a hired person, referred to as the "third party criterion", which distinguishes between productive and unproductive activities [Eurostat, 1999, p.7]. Hence each household job can be valued using the market cost of parallel services.

One way of estimating entails multiplying the average time spent on domestic work by the average hourly wage for professions corresponding to home activities. The total value of housework is then an effect of an aggregation of the values calculated for different types of activity.

In the empirical estimation for Poland, 38 detailed activities within four groups of household activities, such as household upkeep, food preparation, making and caring for textiles, child care and adult care were taken into account. Additionally, 9 domestic work activities to benefit other households in the form of volunteer work were added, according to Eurostat proposals [Eurostat, 1999].

All the work considered in the estimation is done in all or most households. The groups of household work correspond to the groups of services purchased by households on the market. Household members purchase meals (restaurants, staff canteens), clothing (tailoring, cleaning, repairing), care (children's day care, care for the elderly) and the provision and maintenance of housing. Households also need to perform ancillary activities, such as transportation, shopping, gardening, pet care, planning and organizing things, managing finances etc. They are performed in connection with one or more principal activities [Eurostat, 1999, p.21].

The results of household work estimation were broken down by sex, activity on the labour market, family status by number of children, maritual status of respondents including the age of household members, place of living, the number of household members and the level of education.

The first stage of the estimation entailed calculating the average time per day for each of the 47 specified activities. The average time of the *a-th* activity duration in the *j-th* activity group for all days of the week was calculated separately for all selected groups of respondents:

$$ {}_F\bar{t}^{\,z}{}_{laj} = \frac{\sum_{i=1}^{n_1} {}_F t^{z}{}_{ilaj}}{n_F} \tag{1} $$

$$ {}_M\bar{t}^{\,z}{}_{laj} = \frac{\sum_{i=1}^{n_2} {}_M t^{z}{}_{ilaj}}{n_M} \tag{2} $$

where:

- ${}_Ft^{z}{}_{ilaj}$ – duration of the *a-th* activity in the *j-th* group for the *i-th* woman from the *l-th* class in the *z-th* day of the week,
- n_F – the number of women in a subsample,
- ${}_Mt^{z}{}_{ilaj}$ – duration of the *a-th* activity in the *j-th* group for the *i-th* man from the *l-th* class in the *z-th* day of the week,
- n_M –the number of men in a subsample,
- z – the day of the week; z = 1,2,3, where: 1 – Monday-Friday, 2 – Saturday, 3 – Sunday,
- j – group of domestic activities, j = 1,2,3,4,5.

All activities performed by household members to satisfy their own needs are characterized by weekly seasonality. Hence the time spent in a week on domestic work was calculated as the weighted average using the days of the week. The weight of 5/7 was used for weekdays from Monday to Friday and 1/7 for Saturday and Sunday. The weighted average time was calculated separately for men and women:

$$ {}_F\bar{t}_{laj} = [\frac{5}{7}\,{}_F\bar{t}^{1}_{laj} + \frac{1}{7}({}_F\bar{t}^{2}_{laj} + {}_F\bar{t}^{3}_{laj})] \times 7 \tag{3} $$

$$ {}_M\bar{t}_{laj} = [\frac{5}{7}\,{}_M\bar{t}^{1}_{laj} + \frac{1}{7}({}_M\bar{t}^{2}_{laj} + {}_M\bar{t}^{3}_{laj})] \times 7 \tag{4} $$

where:

- ${}_F\bar{t}_{laj}$, ${}_M\bar{t}_{laj}$ - the average week time duration of the *a-th* activity in the *j-th* group for the *i-th* men and women from the *l-th* class,
- ${}_F\bar{t}^1_{laj}$, ${}_M\bar{t}^1_{laj}$ - the average duration of the *a-th* activity in the *j-th* group for men and women from the *l-th* class (weekdays from Monday to Friday),
- ${}_F\bar{t}^2_{laj}$, ${}_M\bar{t}^2_{laj}$ - the average duration of the *a-th* activity in the *j-th* group for men and women from the *l-th* class (on Saturdays),
- ${}_F\bar{t}^3_{laj}$, ${}_M\bar{t}^3_{laj}$ - the average duration of the *a-th* activity in the *j-th* group for men and women from the *l-th* class (on Sundays).

The average weekly value of housework per person from the *l-th* class selected according to socio-demographic characteristics was calculated separately for men and women:

$$ {}_F h_l = \sum_{j=1}^{4}\sum_{a=1}^{n_a} {}_F\bar{t}_{laj} s_{aj} \tag{5} $$

$$ {}_M h_l = \sum_{j=1}^{4}\sum_{a=1}^{n_a} {}_M\bar{t}_{laj} s_{aj} \tag{6} $$

where:

- ${}_Fh_l$ – weekly housework value for a woman from the *l-th* class,
- ${}_Mh_l$ - weekly housework value for a man from the *l-th* class,
- ${}_F\bar{t}_{laj}$ - average duration per week of the *a-th* activity in the *j-th* group for women from the *l-th* class,
- ${}_M\bar{t}_{laj}$ - average duration per week of the *a-th* activity in the *j-th* group for men from the *l-th* class,
- s_{aj} – hourly wage calculated for the *a-th* activity in the *j-th* group.

The average monthly value of domestic work for men and women in May 2004 was calculated by multiplying the weekly value by the number of weeks in a calendar year (52) and dividing the result by 12.

$$ {}_F H_l = \frac{52}{12}\sum_{j=1}^{4}\sum_{a=1}^{n_a} {}_F\bar{t}_{laj} s_{aj} \tag{7} $$

$$ {}_M H_{li} = \frac{52}{12}\sum_{j=1}^{4}\sum_{a=1}^{n_a} {}_M\bar{t}_{laj} s_{aj} \tag{8} $$

where:

- ${}_{F}H_l$ – monthly housework value for a woman from the *l-th* class,
- ${}_{M}H_l$ – monthly housework value for a man from the *l-th* class.

The monthly values of domestic work for men and women of different characteristics formed the basis for calculating the country's annual value. It was weighted according to population structure. Finally, the annual value of work was calculated using the monthly wage rate.

In contrast to the market cost method calculation, in which each type of household work was estimated using an individual wage, the alternative cost method used one wage level for all types of domestic work. The gross wage per hour for men and women was applied. Hence the final results reflected ratios of the average time spent on domestic work.

The average weekly value of housework for men and women can be disaggregated by job and was calculated as follows:

$$ {}_{F}h_l = \sum_{j=1}^{4} \sum_{a=1}^{n_a} {}_{F}\bar{t}_{laj} s_F \tag{9}$$

$$ {}_{M}h_l = \sum_{j=1}^{4} \sum_{a=1}^{n_a} {}_{M}\bar{t}_{laj} s_M \tag{10}$$

where:

- ${}_{F}h_l$ – weekly housework value for a woman from the *l-th* class,
- ${}_{M}h_l$ - weekly housework value for a man from the *l-th* class,
- ${}_{F}\bar{t}_{laj}$ - average duration per week of the *a-th* activity in the *j-th* group for women from the *l-th* class,
- ${}_{M}\bar{t}_{laj}$ - average duration per week of the *a-th* activity in the *j-th* group for men from the *l-th* class,
- s_F – hourly wage for women,
- s_F – hourly wage for men.

4. Empirical results

The average weekly value of household work in Poland in May 2004 was 230 PLN per capita which yielded one thousand PLN per month (gross). This accounted for as much as 42% of the average monthly gross wage in Poland in May 2004 (54% for women and 31% for men). The annual value of household work per person calculated using the monthly wage rate was 12.4 thousand PLN. The monthly value of housework calculated for a woman (1,266 PLN) was 74% higher than the value calculated for a man (728 PLN).

The structure of the total value of household work by type of activity is presented in Table 1. The highest share was observed for food preparation: it accounted for 52% of the total value for women and 41% for men.

Table 1. The structure of the total value of household work by type of activity.

No	Groups of domestic work	Structure (in %)		
		Total	Females	Males
1.	Household upkeep	20,5	15,1	32,0
2.	Food preparation	47,9	51.9	41.3
3.	Making and caring for textiles	5,8	7.1	2.4
4.	Child care and adult care	25.8	25.9	24.3
5.	**Total**	**100.0**	**100.0**	**100.0**

Source: Błaszczak-Przybycińska (2005), p. 90.

Detailed information about all the specified domestic activities is presented in Table 2. As far as meal preparation is taken into account, the value of work done by women was more than double the value of work done by men. There were jobs done primarily by women (e.g. meal preparation, washing dishes) but there were also jobs for which the estimated value of work done by men was higher (e.g. gardening). The high value of work connected to food preparation stems from all the activities in this group being done very frequently and practically in all households.

Table 2. The value of household work in May 2004 estimated by market cost method (in PLN)

No	Specification	Total	Females			Males		
			Total	Employed	Unemployed	Total	Employed	Unemployed
1.	**Household upkeep**	**204.58**	**191.19**	**160.07**	**203.45**	**233.22**	**192.66**	**279.72**
1.1	House construction and renovation	4.85	—	—	—	13.17	13.87	11.79
1.2	Repair of dwelling	17.98	4.33	5.03	4.33	31.63	28.73	33.76
1.3	Making. repairing and maintenance of equipment	5.42	—	—	—	11.74	11.74	15.38
1.4	Other construction and repairs	4.33	—	—	—	8.62	8.62	7.89
1.5	Supply of heating and water	25.96	12.65	8.67	12.65	39.26	31.94	54.56
1.6	Cleaning dwelling	78.78	121.29	101.88	132.30	35.19	22.06	52.00

1.7	Cleaning yard. snow removal	7.45	4.03	4.03	6.89	14.91	8.02	17.77
1.8	Other household upkeep	0.52	3.16	—	3.16	0.52	0.52	3.16
1.9	Travel related to household care	10.53	10.53	6.33	10.53	14.78	9.84	18.29
1.10	Vehicle maintenance	19.89	—	—	—	40.82	36.83	37.83
1.11	Various arragements	28.86	35.19	34.15	33.58	22.58	20.50	27.30
2.	**Food preparation**	**479.05**	**656.33**	**570.05**	**705.08**	**300.30**	**226.16**	**393.34**
2.1	Meal preparation	198.03	300.65	276.21	316.72	88.18	70.98	110.93
2.2	Baking	6.02	10.96	11.53	10.44	0.56	—	0.56
2.3	Preserving	7.15	13.17	10.96	12.61	3.29	0.56	3.29
2.4	Other food management	—	—	—	—	—	0.56	—
2.5	Shopping	86.49	105.78	97.11	110.41	67.17	57.20	80.47
2.6	Gardening	41.43	40.30	24.74	49.49	45.46	29.94	66.78
2.7	Tending domestic animals	8.10	7.45	—	8.71	8.71	4.38	16.81
2.8	Dish washing	55.21	89.48	78.52	97.59	18.07	11.92	26.65
2.9	Travel related to shopping and services	76.61	88.53	70.98	99.10	68.86	50.61	87.84
3.	**Making and caring for textiles**	**58.07**	**89.96**	**84.24**	**96.46**	**17.81**	**14.47**	**24.01**
3.1	Producing textiles	8.62	14.86	8.62	17.25	—	—	—
3.2	Commercial and administrative services	10.66	7.32	7.32	10.01	14.00	10.66	17.33
3.3	Other shopping and services	—	—	—	—	—	—	—
3.4	Other making and caring for textiles	2.60	3.08	3.08	3.08	—	—	—
3.5	Laundry	23.79	43.29	42.38	45.20	3.34	2.86	6.20
3.6	Ironing. Starching	12.39	21.41	22.84	20.93	0.48	0.95	0.48
4.	**Child care and adult care**	**258.31**	**328.25**	**343.94**	**325.43**	**176.63**	**206.57**	**120.73**
4.1	Physical care and supervision	76.18	122.29	125.32	120.29	31.07	40.08	14.04
4.2	Teaching children	30.07	48.40	60.10	38.39	11.70	20.02	8.36
4.3	Reading. playing and talking with children	99.67	105.73	111.76	111.76	83.07	104.22	42.29
4.4	Accompaying children	2.69	2.69	5.89	2.69	2.69	2.69	—
4.5	Other child care	—	—	—	—	—	—	—
4.6	Adult care	4.16	4.16	4.16	4.77	4.16	0.61	4.16
4.7	Caring for pets	7.50	7.50	4.29	7.50	7.50	4.29	7.50
4.8	Walking the dog and other pets	22.97	22.45	16.55	25.65	24.57	17.64	33.15
4.9	Other gardening and pet care	—	—	—	—	—	—	—
4.10	Transporting a child	7.67	10.83	10.83	10.18	3.21	7.67	3.21
4.11	Transporting an adult family member	3.21	—	—	—	4.46	7.67	3.81

4.12	Household management	4.20	4.20	5.03	4.20	4.20	1.69	4.20
5.	**Voluntary work**	**70.03**	**63.53**	**39.30**	**91.48**	**67.43**	**48.53**	**101.27**
5.1	Food management	4.46	5.03	6.15	7.84	0.56	0.56	1.13
5.2	Household upkeep	3.68	6.28	3.68	6.28	3.68	3.16	6.85
5.3	Gardening and pet care	3.34	0.56	0.56	3.34	3.90	3.90	7.19
5.4	Construction and repairs	15.17	—	—	—	31.11	22.32	39.09
5.5	Shopping and services	—	—	—	3.51	—	—	0.69
5.6	Help in employment and farming	7.19	3.60	3.60	6.59	12.00	9.01	14.39
5.7	Child care	31.68	43.55	20.80	59.41	14.86	8.93	26.74
5.8	Adult care	4.51	4.51	4.51	4.51	1.30	0.65	4.51
5.9	Other specified and unspecified informal help	—	—	—	—	—	—	0.69
6.1	**Group 1—4 together**	**1000.00**	**1265.72**	**1158.30**	**1330.42**	**727.96**	**639.86**	**817.79**
6.2	**Group 1—5 together**	**1070.03**	**1329.25**	**1197.60**	**1421.90**	**795.38**	**688.39**	**919.06**

A relatively high value of work was also noted in child and adult care. It constituted 26% of the total value of domestic work done by women and 24% by men.

More than 90% of all women and 70% of all men were engaged in domestic work, notwithstanding the day of the week. Activities in the "child or adult care" group were performed relatively infrequently (by about 40% of the women and 30% of the men).

The value of domestic work on household upkeep accounted for about 1/5 of the total housework value (32% for men and 15% for women). It was the only group of the four specified groups of domestic work in which the absolute level of monetary value was higher for men than women. This group included activities performed solely by men (e.g. house construction, renovation, making, repairing and maintaining equipment) but there were also jobs in which the value of work performed was higher for women (cleaning dwelling). The high value of domestic work done by men in the "household upkeep" group resulted from two factors: first, on average, men spent more time on activities in this group than women and second, activities traditionally performed by men have been ascribed relatively high hourly wages.

The lowest value of work emerged for the "making and caring for textiles" group. Although the monthly value for women was five times higher than the monthly value for men, the absolute value of work for women was very low. This may stem from household textile production being unprofitable coupled with extensive access to market services of this kind.

The estimates of household work were broken down by the socio-economic characteristics which have an impact on the estimated value. The selected results of the estimation are presented in Table 3.

Table 3. Monthly value of household work in Poland in May 2004 according to selected socio-economic characteristics (market cost method).

Specification	Women		Men	
	Value of housework in PLN	Value of the specified group stated as a percentage of total value for women (%)	Value of housework in PLN	Value of the specified group stated as a percentage of total value for men (%)
TOTAL	**1265.72**	**100.0**	**727.96**	**100.0**
by activity on the labor market				
• employed	1158.30	91.5	639.86	87.9
• unemployed	1330.42	105.1	817.79	112.3
by family type (age and presence of children)				
living with parents				
• aged under 25	489.36	38.7	321.92	44.2
• aged 25-44	711.84	56.2	498.94	68.5
in couples, without children under 18				
• aged under 45	960.66	75.9	593.28	81.5
• aged 45-64	1247.91	98.6	781.65	107.4
• aged 65 and above	1265.59	100.0	851.89	117.0
not in couples, without children under 18				
• aged under 45	674.61	53.3	501.67	68.9
• aged 45-64	1145.73	90.5	913.81	125.5
• aged 65 and above	1021.28	80.7	957.67	131.6
with children				
• in couples, children aged 0-6	2507.48	198.1	1152.49	158.3
• in couples, children aged 7-17	1464.54	115.7	762.97	104.8
• single parent	1944.45	153.6	915.72	125.8
by family type (number of children)				
• a couple without children	1157.65	91.5	809.12	111.1
• a couple with one child	1450.06	114.6	842.23	115.7
• a couple with 2 children	1543.84	122.0	782.73	107.5
• a couple with 3+ children	1658.80	131.1	668.89	91.9
by type of residence				
• urban	1201.20	94.9	734.20	100.9
• rural	1351.96	106.8	711.23	97.7
by level of education				
• tertiary	1231.49	97.3	818.22	112.4
• secondary	1287.69	101.7	763.49	104.9
• elementary vocational	1574.34	124.4	747.89	102.7
• primary	1171.56	92.6	687.22	94.4
• primary not completed	1008.02	79.6	667.94	91.8

Variables differentiating the value of household work include sex and activity on the labour market. Also the presence of small children in a household has a

strong impact on the estimated value. The value of domestic work estimated for a woman with children aged 0-6 was 2500 PLN and it was almost two times higher than for women on average (the exchange rate on May 31st, 2004 was: 1 USD=3.80 PLN and 1 EUR=4.65 PLN). The value of household work calculated for a woman living in a couple, with children aged 7-17 was considerably lower (1500 PLN). The value of work done by a single mother fell within the two figures given above (1900 PLN). These ratios were similar for men although the absolute monetary values attached to the domestic work done by men were significantly lower.

Obviously, the high value of domestic work in households with children aged 0-6 years can be observed mainly in the child care group of activities.

According to some authors, the value of domestic work in the child care group may be overestimated. The relatively long time spent on child care may be the main reason for overestimation when using the market cost method calculation. A portion of this time is only passive care. On the other hand, many parallel domestic jobs are done simultaneously and are not subject to separate valuation.

The following relationship has been observed in households without children aged under 18: for both men and women, the monetary value of household work increased with age. At the same time, the value of work done by women not living in couples was lower in all age groups compared to the value of domestic work done by women living in couples. The opposite was observed for men – for all age groups the value of domestic work was lower for those living in couples. This may reveal some information about the extent of domestic work done by women for the benefit of their spouses.

The number of children in a household has an impact on the value of domestic work done by women. It increases with the number of children irrespective of their ages, though the increase is not very intensive. The work done by men decreases with the number of children. This may be explained by their more intensive work on the labour market.

The results of domestic work valuation were also presented by the type of residence. The highest values were derived for women living in rural areas when compared with urban areas. For both types of residence, the value of domestic work done by women grew with the number of household members, with the exception of households consisting of 6 or more members. Among men, the highest value of work was recorded for those living in towns but the difference was not high.

Some interesting relations in housework values can be pointed out for respondents grouped by level of education. The total value of domestic work done by women fell with the level of education (elementary vocational, secondary and tertiary level of education). This relationship was observed on jobs connected to household upkeep, food preparation, making and caring for textiles. This may be explained partly by better educated women being more active on the labour market.

A different situation was noted for child and adult care. In this field the value of work done by women with a university degree was higher than for those with secondary education. Women with an elementary vocational education recorded the highest value in this category, which can be partly explained by them having the largest number of children in their households compared to other groups of respondents.

The figures for men were totally different. The total value of domestic work done by men rose with the level of education, especially in the area of child care. This may be explained by changes in social customs and the overall trend of men doing more of the housework.

The market cost method was used also to estimate the value of volunteer work done for the benefit of other households. Its average monthly value was about 70 PLN, accounting for 7% of the value of work done in one's own household. The value calculated for men and women not working on the labour market was more than double the figure for people working on the market. Child care consumed the highest amount of time spent by women in total on volunteer work (69%). This may be explained by the extensive amount of care provided to grandchildren living under separate roofs. Men spent the most time on construction and repairs. Men who do not work on the labour market also had a high share of work on child and adult care.

The highest absolute value of volunteer work applied to women living in childless couples. If the age of respondents is taken into account, the highest value of volunteer work was observed among men and women in couples aged 45—64. This shows the importance of grandparents active on the labour market engaged in helping their adult children to provide care to their grandchildren.

The total value of household work in Polish households (excluding volunteer work) was 265,749 million PLN, amounting to 30.0% of GNP.

The estimation of household work based on the alternative cost method was based on the average hourly wage for men and women. The absolute values of household work were much higher compared with the results derived using the market cost method (Table 4). This stems from using a higher average hourly wage than the wages by profession used in the market cost method calculation.

Table 4. Monthly values of household work per capita in May 2004

No	Group of works	Value of household work (in PLN)					
		Market cost method			Alternative cost method		
		Total	**Females**	**Males**	**Total**	**Females**	**Males**
1.	Household upkeep	204.58	191.19	233.22	316.42	308.75	347.36
2.	Food preparation	479.05	656.33	300.30	770.81	1003.69	497.94
3.	Making and caring for textiles	58.07	89.96	17.81	107.99	161.85	29.12
4.	Child care and adult care	258.31	328.25	176.63	235.65	273.56	182.69
5.	Volunteer work	70.03	63.53	67.43	85.45	69.51	91.35
6.	**Groups 1-4 together**	**1000.00**	**1265.72**	**727.96**	**1430.87**	**1747.85**	**1057.12**
7.	**Groups 1-5 together**	**1070.03**	**1329.25**	**795.38**	**1516.32**	**1817.36**	**1148.46**

On the basis of the valuation of household work using both input methods, an attempt was undertaken to assess the accuracy of the valuation. The approximate method was used (Kordos, 1960). In order to estimate the relative error at least two valuations are needed. The final estimation result (x_{fin}) is calculated as the mean of all the estimates:

$$x_{fin} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{11}$$

The maximum and minimum values are defined (x_{max}, x_{min}) for all the estimates (x_i) and the range is defined as the level of accuracy:

$$\Delta = x_{\max} - x_{\min} \tag{12}$$

The absolute assessment error is calculated as the level of accuracy divided by the number of evaluations:

$$s_n = \frac{\Delta}{n}. \tag{13}$$

The relative error is defined as the percentage share of absolute assessment error (s_n) in the mean of all the estimates (x_{fin}). Although this method does not allow one to define estimation accuracy as precisely as in the sampling method, it is nevertheless still useful in these types of intermediary assessments.

The achieved accuracy depended on the type of domestic work. For women, the minimum relative error was derived for household management (2.3%) while the largest one was for teaching children (31%). For the total household work done by women, the relative error was the lowest for volunteer work (4.5%) while the highest was for making and caring for textiles (28.6%). For men, the minimum relative error was derived for child care (0.1%) while the highest was for laundry (35.6%). For the total work done by men the relative error was the lowest

for child and adult care (1.7%). It was equal to 15.1% for volunteer work and 19.7% for household upkeep. The highest error level emerged for making and caring for textiles and food preparation (24.1% and 24.8%, respectively).

5. Conclusions

The problems connected with household production estimation are very current in the context of the Satellite Account of Household Production recommended by Eurostat.

If the harmonized time use survey is used in the estimation of household work using input methods some changes in time use surveys should be considered. More precise definitions of activities within household work should be applied and all activities connected with household work should be gathered within a single group excluding those activities not focused on satisfying a household's own needs. It would be very important to expand the items connected to child and adult care, considering recent demographic changes in Europe. Apart from this, research is needed on the scope of household work estimation in the context of decreasing leisure time as a consequence of undertaking household work.

REFERENCES

ALIAGA CH.,WINQVIST K. (2003), How Men and Women Spend Their Time, *Statistics in Focus*, Theme 3—12/2003, Eurostat.

BŁASZCZAK-PRZYBYCIŃSKA I. (2005), Estimation of Housework on the Basis of Time Use Survey Data, in: CENTRAL STATISTICAL OFFICE, Time Use Survey 1st July 2003—31st May 2004, *Statistical Studies and Analyses*, Warsaw (in Polish).

BŁASZCZAK-PRZYBYCIŃSKA I. (2006), Methodology and Empirical Results of Time Use Surveys in Poland, *Statistics in Transition*, Vol. 7, No. 6, pp. 1345—1360.

EUROSTAT (2003), Household Production and Consumption. Proposal for a Methodology of Household Satellite Accounts, *Task force report for Eurostat*, Unit E1, Luxembourg.

EUROSTAT (1999), Proposal for a Satellite Account of Household Production, Eurostat Working Papers, 9/1999/A4/11, Luxembourg.

GERSHUNY J. (1995), Time Budget Research in Europe, *Statistics in Transition*, Vol. 2, No. 4, pp. 529—551.

GUS (2003), Structure of Wages and Salaries by Occupation in October 2002, *Information and Statistical Papers*, Warsaw.

GUS (2005), Time Use Survey 1st July 2003—31st May 2004, *Statistical Studies and Analyses*, Warsaw.

HARVEY A. (1995), Emerging Needs for Time Use Data, *Statistics in Transition*, Vol. 2, No. 4, pp. 513—515.

KORDOS J. (1960), An Attempt of Accuracy Assessment of Estimates, *Statistical News*, No. 3. pp. 24—27 (in Polish).

KORDOS J. (1988), Time Use Surveys in Poland, *Statistical Journal of the United Nations, Economic Commission for Europe*, Vol. 5, No. 2, pp. 159—168.

KORDOS J. (1998), Social Statistics in Poland and its Harmonization with the European Union Standards, *Statistics in Transition*, Vol. 3, No. 4, pp. 617—639.

RYDENSTAM K. (1995), The Harmonized European Time Use Surveys, *Statistics in Transition*, Vol. 2, No. 4, pp. 553—581.

SZCZERBIŃSKA L. (1980), Monetary Valuation of Household Work, GUS, Warszawa 1980 (in Polish).

SŁABY T. (1998), Conclusions from the Polish Time Use Pilot Survey 1996, *Statistics in Transition*, Vol. 3 No. 4, GUS, Warsaw, pp. 743—756.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 563—576

SOME ASPECTS OF POST-ENUMERATION SURVEYS IN POLAND[1]

Jan Kordos[2]

Formerly Warsaw School of Economics

ABSTRACT

Central Statistical Office of Poland has started the preparation for the Census of Agriculture and the Census of Population and Housing to be carried out in 2010 and 2011 respectively. This paper is concerned with the evaluation of census quality methods presented in the *Recommendations for the 2010 Censuses of Population and Housing* (2006). Out of several methods for evaluating the results of a census, one method is selected, i.e. *a post-enumeration survey* (PES). First purposes of PES are recounted and some methodological issues discussed. Next post-enumeration surveys in Poland are presented and selected methodological issues considered. In conclusion, some recommendations for the 2011 Census of Population and Housing are put forward.

Key words: Post-enumeration survey; Census of Population; Census of Agriculture; Coverage error; Content error; Data collection; Demographic analysis.

1. Introduction

In Poland as well as in other countries a population census is the most extensive and expensive data-collection investigation. With vast amounts of resources spent, there is usually tremendous pressure on census takers to ensure that census results are accurate. As a result of a massive nature of the census operation, it is inevitable that some inaccuracies arise from deficiencies, including errors of coverage and response. This, however, does not diminish the importance

[1] This paper has been presented at the XXVI Conference on Multivariate Statistical Analysis and the VI Conference on Statistics in Social and Economic Practice, Łódź, Poland, 5-7 November 2007

[2] E-mail: jan1kor2@aster.pl

of the census as long as users understand the limitations of the data and the errors do not affect the major uses of the data.

Several methods are available to evaluate censuses, including (Conference of European Statisticians, 2006; Kordos, 1988):

- demographic analysis,
- comparison of census results with data from other sources, and
- matching census responses with responses from interviews conducted during a post-enumeration survey (PES).

Basically, a PES is an independent survey that replicates a census. The survey results are compared with census results, permitting estimates to be made of coverage and content errors. Coverage errors refer to people missed in the census or erroneously included, whereas content errors evaluate response quality of selected questions. The PES allows census organizations to uncover deficiencies in the methodology of the census and make adjustments for future censuses. PES results can also be used to adjust census results, although this is as likely to be a political decision as a technical one.

In Poland to evaluate quality of census data some post-census enumeration surveys were used. However, some attempts were also made to apply demographic analysis (Nowak,1998; Strzelecki et al, 2002). Here we focus only on some aspects of PES. First some features of evaluation of census quality and cooperation of different experts involved at various levels of census programme are recounted.

2. Evaluation of Census Quality

As is stressed in the *Recommendations for the 2010 Censuses of Population and Housing* (2006), *a census quality assuran*ce regime comprises a wide variety of mechanisms and processes acting at various levels throughout the census programme. An important technique applicable in many census operations is *statistical quality control* (Duncan, 1986; Hald, 1981). It primarily addresses accuracy, although depending on the operation it may also address other elements of quality. Coverage is a critical element of accuracy. It has a direct influence on the quality of population counts and an indirect impact on the quality of all other data produced by the census. Thus, the coverage concerns should be taken into consideration in the design and implementation of most census activities and their quality assurance programmes.

It should be stressed that instructions and training on dwelling coverage for staff engaged in dwelling listing and enumeration must be clear, explicit and easy to understand. The target population must be well-defined and related instructions and questions for both interviewers and respondents need to be carefully developed and thoroughly tested. Clarity and simplicity of instructions concerning place of residence for enumeration is vital to help ensure people are enumerated exactly once and at the correct location. This is particularly important in

minimizing overcoverage. Questionnaires should include guidance or questions to assist with situations where it may be unclear whether certain persons should be included or not. Special procedures should be developed for population groups difficult to enumerate (for example remote areas, collectives or group quarters, persons with literacy or language difficulties). Processing procedures should be developed with a view to minimizing the risk of erroneously cancelling, losing or artificially creating households. A well-prepared publicity campaign can play an important role in promoting census awareness and response, thus helping minimizing coverage error. All of these steps, along with appropriate training, supervisory checks and quality assurance approaches during operations will help minimize coverage error. Nonetheless some coverage error is unavoidable. Hence it is important to measure, analyze and report on coverage error. This is best done via an independent *post-enumeration survey* of a sample of census areas or via *a Reverse Record Check* methodology (Dauphin and Canamucio, 1993;Hogan, 1988; Kordos, 1988). Results of coverage studies provide an important evaluation of the current census and can also provide valuable guidance for the next census. Results in conjunction with the census counts themselves are a critical input for population estimation programmes. Analysis of census results vis-à-vis demographic projections of the population from the previous census can also be informative (*demographic analysis*).

3. Cooperation of different experts

Quality concerns need to receive appropriate attention during design, implementation and assessment. Very important is the cooperation of different experts:

- *Subject matter experts* will bring knowledge of content, client needs, relevance and coherence.
- *Statistical methodologists* bring their expertise on statistical methods and data quality trade-offs, especially with respect *to accuracy, timeliness and cost.*
- *Operations experts* bring experience in operational methods, and concerns for practicality, efficiency, field staff, respondents and operational quality assurance and control.
- The *systems experts* bring knowledge of technology standards and tools that will help facilitate achievement of quality, particularly in the timeliness and accuracy dimensions. In collaboration with subject matter experts,
- *Dissemination experts* will bring a focus to accessibility and interpretability.

Poor cooperation among different experts may influence the quality of census data significantly.

4. Evaluation methods of census errors

A number of methods are available to estimate the coverage and content errors of censuses. These include (Hogan, 1992; Hogan and Wolter, 1988; Kordos, 1988; Whitford and Banda, 2001; Woltman et al, 1988):

a) Quality control techniques such as internal consistency checks;
b) Comparisons of results with other data sources including previous censuses, current household surveys, and/or administrative records;
c) Record-checking, in which individual census records are matched against alternative sources and specific data items are checked for accuracy;
d) Some evaluations analyze, interpret, and synthesize the effectiveness of census components and their impact on data quality or census coverage;
e) Post-enumeration surveys are used to estimate census coverage error;
f) Post-census surveys designed to measure content error are usually known as re-interview surveys; and
g) Ethnographic and social network methods provide a way to study the effects of mobility on census coverage or to measure census coverage of specific subpopulations.

Other evaluation methods are also used. These include:

- Surveys to determine customer satisfaction with data collection instruments or questionnaire assistance; and
- Focus group interviews to learn how or why respondents behave in a certain way.

We focus only on a post-enumeration survey.

5. Purpose of a post-census enumeration survey

While a number of methods have been developed to evaluate census data, for many countries the PES seems to be ideal owing to the scarcity of appropriate data to facilitate the effective use of other methods. The lack or incompleteness of registration systems and absence of regular population and demographic surveys contributes to the lack of use or limited use of other methods of census evaluation. In general, a number of countries have relied primarily on PES methodology to evaluate the census undercount (Dauphin and Canamucio, 1993; Whitford and Banda, 2001)

Post-enumeration surveys are an accepted census self-evaluation tool. Typically, a PES is an independent survey that replicates a census. The survey and the census results are then compared (matched). The results of the comparison are used to measure the coverage and/or errors in content of the census. Estimates of net coverage, the number of people omitted in the census, the number erroneously enumerated and content error rates for specific questions are typical products of a PES.

Additionally, these estimates can be broken down further into their component parts. One can design a survey so that reliable estimates of undercount or overcount can be obtained for the entire census, for geographic areas of interest in the census, and for any of a host of demographic characteristics, such as age, race and sex, for which one might desire census coverage statistics.

The survey results also enable one to *uncover census methodologies* or *operations* that, when implemented, produced less than desirable results. Suppose, for instance, that a high census omission rate was observed in rural areas. One might then use specific PES results to examine whether the rural errors were due to the omission of whole housing units. If so, this might well imply an incomplete census frame and cause one to re-examine the methodology for building an address list in rural areas.

PES results can be used to *adjust census results*. Using a carefully designed survey, under- or overcounts can be converted into *adjustment factors* and the census population increased or decreased accordingly by these factors. Later in this paper, we will discuss post-stratification and the need to ensure homogeneity within each adjustment cell.

In addition, censuses are used for many other purposes, such as updating population estimates; developing and updating sampling frames; correcting and updating population registers and the establishment and updating of key components of the Geographic Information System (GIS).

These many uses suggest that there is a need to use an objective method for assessing coverage and content errors as a crucial step for concluding a census operation. *Quality assurance alone, introduced at various stages of census operations, cannot ensure a complete evaluation of the qualitative and quantitative accuracy of census data.*

To sum up, post-enumeration surveys have many useful purposes. They basically inform users regarding the quality of the census data. As stated earlier, providing limitations of published census data increases the confidence of informed users in such data. On the other hand, there are distinct limitations and constraints in managing and implementing the evaluation survey, which are presented below.

6. Problems and constraints associated with post-census enumeration surveys

Although a PES can be an important component of a census programme and can contribute to the process of building confidence in the census results, a poorly designed and executed survey can inflict considerable damage to census legitimacy. We list below some of the problems and constraints associated with post-enumeration surveys (Dauphin et al,1993; Whitford et al, 2001).

- Planning and management of a PES, ideally, have to be undertaken by a staff that is separate from the census staff. This is not usually the case in many countries;
- The design of the survey-especially the matching step — is relatively complex. For example, in the United States planners continue to find design flaws in the matching system. However, as corporate experience grows, these flaws become more and more minor;
- The PES interview itself is demanding. Usually it incorporates questions to determine if the respondent should "really" be counted at the residence in question. Also, the PES interview usually transpires after the census interview, at which point the respondent may feel overburdened and not be as forthcoming with accurate information;
- Past failures in some countries in conducting post-enumeration surveys discourage such countries and others from conducting PES's in the subsequent rounds;
- Some of the countries, which have conducted PES's, have not used the results to adjust population census figures. In such cases questions have been raised about the rationale for conducting PES's;
- In some countries, census planners feel it is enough to institute good-quality assurance procedures at various stages of census activities; therefore, they see no need for a PES.

7. Design and methodological issues

Considering whether a PES is worth it or not leads immediately to some decisions that have to be made regarding the design of the survey. These decisions revolve around what goals one has for the survey and what answers best suit the individual situation in which the survey will be conducted.

We will assume in this paper that the goal of the PES interview is to establish carefully who lived in the subject housing unit on the day the census was officially taken. In the next step we match the results from the interview to appropriate census forms in a well-defined area around that subject housing unit.

Two other design decisions have to be made: *What is the primary sampling unit for the survey*? and *What is the definition of cases to be included in the survey*? This leads into our next topic, the sampling frame of the survey.

Frames

A popular choice for a sampling frame is to use an area sample for the coverage measurement survey. The primary sampling unit can be the census enumeration area (CEA) or a block. CEAs are land areas surrounded by visible geographic features such as roads and streams. The frame, therefore, consists of creating a universe of CEAs in the country and dividing those into sets of CEAs (or clusters of CEAs) that can be interviewed by a single interviewer within the allotted time.

Another option is to use a survey that is already in place that is being taken around the time of the census. This has the large advantage of using an existing organization to manage the PES. It also has several disadvantages ((Whitford et al, 2001):

- The existing survey may not be large enough, and supplementing it may be as complex as creating a specially designed survey;
- Procedures may have to be augmented with the result that the quality of the existing survey and the PES suffers;
- The ultimate sampling units may not lend themselves to being an efficient erroneous enumeration sample, where duplication, coding errors and so forth need to be easily visible.

Sample design

We mentioned above the option of using CEAs (or blocks) clusters as a frame for the survey. One might want to design the sample in several stages by first choosing a group of these clusters and then optimizing the sample by sub-sampling. For instance, the housing unit totals from a previous census might be used to choose the initial sample of clusters; then, after the addresses of all of the housing units in the sample clusters are listed, the CEAs clusters might be divided into small, medium and large sampling strata.

Listing

For the CEAs clusters in which interviewers are to enumerate, interviewers need maps to find each subject housing unit and a listing of all the housing units in the CEAs cluster. The listing operation is done independently of any census activity. Not only are the addresses of existing housing units listed, but inquiries are made at commercial structures and other structures to ensure that no people live in them. One option is to give enumerators blank maps upon which they put a numbered spot representing each living quarter or potential living quarters in the cluster.

The listing needs to be of a high quality. It must ensure not only that the correct CEAs are listed but also that a complete list is obtained within the cluster. So a quality assurance plan needs to be created to ensure correct listings — for

instance, to ensure that commercial structures have, indeed, been checked to see if they contain residences.

Interviewing

The interview approach is currently the common method used in PES. The questionnaire asks about all people who reside at the sample address on census day and asks questions to ensure that the respondents should have been counted at this address. Subsequently it searches for them at that address and in the search area surrounding it.

Obviously, since people move, it is most efficient if the PES interview can occur as soon after census day as practical. Getting information about out-movers (those who move out of the sample address between census day and the PES interview) is usually difficult.

Matching

After data capture is completed for the PES interviews and the census data prepared for each PES cluster, the next step is to match the two. One approach is to accomplish this in two steps: computer matching, (that is, makes the easiest matches), followed by clerical matching of the remaining non-matches and possible matches as determined by the computer process.

Of course, matching can be completed manually. The process involves clerks first gathering materials to facilitate matching in a particular cluster. The materials include:

- Address lists from both the census and the PES;
- Census forms for the cluster;
- PES interview results for the cluster; and
- Maps for the cluster from the census and the PES.
- Gathering materials for a cluster is indeed a cumbersome part of the matcher's job.

Regardless of how it is done, the basic process of matching remains the same: comparing people's names and demographic characteristics between forms — the census form and the PES interview form.

Reconciliation

PES interviews ask a battery of questions to determine if a person should have actually been counted at the particular housing unit on census day. If whole households of sample housing units have been not matched to anyone in the census- sample, they would not have been asked the residence questions. A field follow-up operation is needed to determine if the unmatched people in the sample unit were or were not erroneously enumerated.

Additionally, in a coverage measurement survey follow-up operation, census-sample interviews with unmatched people that had been completed by

proxy respondents can be followed up. Some research has indicated that PES interviews by proxy respondents need this additional attention to ensure accuracy.

After the follow-up operation, forms are received in the processing office and their final match status is coded—probably by the same people who did the earlier matching. This completes the operations for the PES, and we move on to estimating the under- or overcount.

Estimation

Estimation method is an important step in data quality assessment. In this case a dual-system estimation method should be applied. In order to calculate the dual-system estimate for each post-stratum, one must know how many people were counted in the PES only, the census only, and in both. Missing data make this impossible to determine exactly, either because a person is not assigned to a unique poststrata, or because the person is not assigned to a specific dual-system estimation cell (Hogan, 1992; Hogan and Wolter, 1988; Woltman at al, 1988).

Missing data

Inevitably, in any survey, missing data are encountered. Interviewers, however stubborn, cannot obtain every answer to every question and, in fact, given time constraints in surveys, cannot interview every household. During estimation for coverage measurement surveys, we must account for missing PES-sample data and missing census-sample data.

Missing data can be separated into categories with different approaches taken for each. For instance, missing data can be divided into three types:

- Entire households that were unable to be interviewed in the census sample. With this type of missing data, planners can take the approach of redistributing the sampling weight assigned to each of these households to other households living in similar-type dwellings interviewed in the same block;
- Missing demographic characteristics data. These data are to be used in post-stratification (see next section). When they are missing, substitute data can be imputed in their place using a “hot deck” procedure. This procedure chooses data from a completed case that are very similar to the case with the missing data;
- Unresolved match status or residence status. In the census sample there are cases in which, even though they have undergone a reconciliation interview, match status and/or residence status cannot be resolved. Match status is whether a person matches a census enumeration or not, and residence status is whether or not a person actually should have been counted in the census at the subject residence. In the PES sample similar missing data are encountered when there isn’t enough information to determine if someone is correctly enumerated in the census. Cases

without match or residence status can be assigned a probability of matching and/or a probability of being a census day resident based on all the information collected about them and cases with similar characteristics.

Post-stratification

The object of post-stratification is to include in each dual-system estimate the people who have similar capture probabilities in the census. For instance, young people are usually more mobile than the elderly and so more difficult to count. To mix young and old in one dual-system estimate would lead to a bias in the estimate. However, to have separate post-strata for age groups and separate estimates for each and then to add the estimates across the age groups avoids this bias problem.

On the other hand, having too many post-strata so that each one does not receive a large enough sample will increase variance of the estimates. Post-stratification is a balancing act that has to be carefully thought out.

Some examples of coverage measurement survey post-stratification variables are, for instance: age, sex, tenure (whether one owned or rented his/her home), degree of urban/city and type of enumeration area of the country and mail return rate of census forms.

The following topics should be carefully considered: a) frames, b) sample design and sample size, c) listing, d) interviewing, e) matching, f) reconciliation, g) missing data, h) post-stratification, i) estimation, j) report preparation.

8. Post-enumeration surveys in Poland

Some kind of census data checking were used for all population censuses carried out in Poland. First population censuses carried out in 1921, 1931, 1950, 1960 and 1970 used rather very small sample selected on non-sampling basis (GUS, 1998; Kordos, 1988; Strzelecki et al, 2002). First PES on sampling basis was applied in 1978 for the 1978 Population Census (Zasepa, 1993). Quite reasonable size of samples and sampling designs were used for PES in 1988 (Nowak, 1998), for the Micro-census 1995 and the 2002 General Population and Housing Census (GUS, 1996; Szabłowski et al, 1996).

Here focus is put on PES used to evaluate the 1995 Microcenus of population and to the 2002 Population Census (Szablowski et al., 1996), where new data quality measures have been applied.

8.1. Post-enumeration surveys for the 1995 Micro-census and the 2002 Census of Population in Poland — new measures for data evaluation

New measures for data quality evaluation were used for last two censuses:
1. The 1995 Micro-Census of Population,
2. The 2002 General Census of Population.

The 1995 Micro-Census and PES

In May 1995 a large-scale sample survey (micro-census) of the population and housing was carried out. This was the third micro-census; the two pervious ones were conducted in 1974 and 1984. The 1995 Micro-Census covered 5 per cent of population, i.e. nearly 600 thousand households.

After the Micro-census, a post-census enumeration survey was carried out. Out of the surveyed sample for the Micro-census, 1176 census enumeration areas (CEA's) were selected, and in the selected CEA every other dwelling was chosen for checking purposes. Altogether, 12.6 thousand dwellings were checked. The survey studied coverage error and content error. Some results were published in GUS (1996) without any analysis.

The 2002 Population and Housing Census and PES

A PES was conducted three weeks after the main census. A primary sampling unit was Census Enumeration Area (CEA). Out of all 177,591 CEAs for PCES 903 CEAs were selected using stratified sampling design by region with proportional allocation. Altogether 60,029 dwellings were selected. 27 census items were checked. As by now no PES results were published.

New measures for census data evaluation

Starting from 1995 new two measures of data quality evaluation have been applied (Szablowski et al., 1996):
a) fitness index , and
b) relative error of averages.

Fitness index

Measurement error is characterised as the difference between the observed value of a variable and the true, but unobserved, value of that variable. As a measure of data quality an "*Index of consistency*" was applied (Szablowski et al, 1996.). This index takes may be estimated as follows:

$$IZ = 1 - \frac{s_d^2}{s_x^2 + s_y^2} \qquad (1)$$

where: $d_i = x_i - y_i$

s_d^2 – stands for a variance of d,

s_x^2 – variance of variable x,

s_y^2 – variance of y.

With assumption of special model of errors' generating this index is generalization of well known in statistical literature *inconsistency index* introduced by Pritzk and Hanson (1962).

Relative error of averages

Relative error of averages is estimated from the simple formula:

$$BWZ = \frac{\bar{x}(P) - \bar{x}(C)}{\bar{x}(P)} 100\% \qquad (2)$$

where $\bar{x}(P)$ and $\bar{x}(C)$ stand for average of characteristic in question in the census and control survey respectively.

No results of the 2002 Population Census quality assessment were published and no conclusions from the 2002 PES drawn. There are no publications from other post-enumeration surveys conducted previously. It is not clear why these post-enumeration surveys were carried out.

9. Concluding remarks

Taking into account the international recommendations (Conference of European Statisticians, 2006; United Nations, 1998) and international practice in this field one may conclude that post-enumeration surveys in the 2010 Census of Agriculture and the 2011 Census of Population and Housing should be carried out. However, post-enumeration surveys are worth conducting if they are carefully planned and function within operational and statistical constraints. Cooperation of the different kind of experts involved in preparation, implementation, processing and publication of population census is very important for the quality of census results. While independence between the census and the PES is a fundamental requirement, in practice operational independence seems to suffice because it is not possible to make all the various aspects of the census and PES operations mutually exclusive.

Since there is no error-free census, there is a need to continue to consider PES's as part of census programmes. For the PES to be useful in measuring coverage and content errors, it must be well planned and implemented. We therefore suggest that efforts should be made to:

a) evaluate the PES carried out after the 2002 Population Census and Housing as a first step for the next census preparation;

b) develop good area frames, with well-defined and mutually exclusive enumeration areas;
c) design plausible probability samples to facilitate objective generalization of PES results to relevant domains;
d) consider application of dual estimation system;
e) prepare a programme for checking quality of registers if they are to be used in the census operation;
f) consider application of small area estimation methods;
g) adopt efficient but realistic matching rules;
h) harmonize definitions and concepts used in both the census and the PES;
i) ensure that items included in the PES for matching purposes are relevant and useful;
j) involve well-trained and qualified field staff;
k) train key staff, involved in the design of PES samples, in survey sampling methods;
l) carry out pre-tests for the PES process and field reconciliation;
m) allocate adequate funds to the PES within the framework of the census;
n) keep the PES as simple as possible and stick to objectives that are attainable, and
o) publish all methodology of the PES;

Planning and management of a PES, ideally, have to be undertaken by a staff that is separate from the census staff. The survey must be independent of the census. In the survey's sample areas, census results *must* not be biased by the implementation of the PES.

REFERENCES

CONFERENCE OF EUROPEAN STATISTICIANS (2006), Recommendations for the 2010 Censuses of Population and Housing , Eurostat, United Nations, New York and Geneva.

DAUPHIN, M, , and CANAMUCIO, A. (1993). Design and implementation of post-enumeration survey: developing country example. Washington: *International Statistical Programs Center, US Bureau of the Census.*

DUNCAN, A.J. (1986), Quality Control and Industrial Statistics. Fifth edition. R.D. Irwin Inc., Illinois

EUROSTAT (2000b), *Standard Quality Report*. Eurostat Working Group on Assessment of Quality in Statistics, Luxembourg, April 4—5.

GUS (1998), Methodology and Organisation of Microcensuses (in Polish), *Statystyka w Praktyce*, Warszawa.

GUS (1996), An Index of Consistency as a measure of data quality – on the basis of Post-enumeration Survey for the Microcensus 1995 (In Polish)). *Aneks Metodyczny*, GUS, Warszawa. GUS (2003),

HALD, A. (1981), *Statistical Theory of Sampling Inspection by Attributes*. Academic Press, New York.

HOGAN, H. (1992), The 1990 Post-Enumeration Survey: An Overview, *The American Statistician, Vol. 46, No. 4, pp. 261—269.*

HOGAN, H. and WOLTER, K. (1988), Measuring Accuracy in a Post-Enumeration Survey, *Survey Methodology*, vol.14, No. 1, pp. 99—116.

KORDOS, J. (1988), Quality of Statistical Data (In Polish), PWE, Warsaw , 244 pages.

NOWAK, L. (1998), Quality of Census Data, In: Tendencies of Changes in Structure of Population, Households and Families in 1998-1995 (in Polish). GUS, Warsaw, pp. 22—31.

STRZELECKI, Z. and TOCZYNSKI, T. (Eds) (2002), Population Censuses of the Polish Republic (in Polish), *Polish Demographic Society, Central Statistical Office*, Warszawa.

SZABLOWSKI, J., WESOLOWSKI, J. and WIECZORKOWSKI, R. (1996), Index of Fitting as a Measure of Data Quality — on basis of the Post-enumeration Survey of Microcensus 1995, (In Polish)., "*Wiadomosci Statystyczne*", No. 4. pp. 43—49.

UNITED NATIONS (1998).*Principles and recommendations for Population and Housing Censuses. Revision I,* Sales No. E.98.XVII.8.

WHITFORD, D. C. and BANDA, J. (2001), *Post-enumeration Surveys: area they worth it?* Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects, UN Statistics Division, New York, 7—10 August 2001.

WOLTMAN, H., Albertti, N., and MORIARITY, C. (1988), Sample Design for the 1990 Census Post-Enumeration Survey, *Proceedings of the Survey Research Methods Section, American Statistical Association, pp. 529—533.*

ZASĘPA, R. (1993), Use of Sampling Methods in Population Censuses in Poland, *Statistics in Transition*, vol. 1, Number 1, June 1993, pp. 69—78.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 577—591

ON BASIC NOTIONS OF NONPARAMETRIC BAYESIAN INFERENCE

Marek Męczarski

ABSTRACT

In this paper basic concepts of nonparametric approach to Bayesian statistical analysis are presented, in particular random probability measures and the Dirichlet process. Their fundamental properties and simple statistical applications are shown.

Key words: Bayesian statistics; stochastic process; random probability measure; Dirichlet process.

1. Introduction

Let us recall a general Bayesian statistical model: let $\mathcal{X}$ be a sample space, $\mathcal{M}$ — a σ-algebra of events, $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ — a parametric family of probability distributions. We introduce a σ-algebra $\mathcal{F}$ of events in the parameter space Θ and we treat the parameter value θ as a realization of a random variable with values in Θ. A probability distribution on the space $(\Theta, \mathcal{F})$ is called a prior distribution.

Observe that in such a way we choose at random a probability distribution from the family $\mathcal{P}$ via a random parameter. Nonparametric approach in Bayesian statistics consists in direct random choice of the sample probability distribution, i. e. in considering probability distributions as random elements called *random probability distributions* (*random probability measures*). This means for example that probabilities of random events are random variables which have to satisfy the definition of probability. Such an approach implies quite complex structure of dependence of these variables and therefore requires theoretical foundations which ensure consistency and coherence of definitions and properties. These foundations are provided by the theory of stochastic processes. We will define a random probability measure as a particular stochastic process.

Considerations on nonparametric Bayesian methods were begun by Ferguson (1973), who introduced a notion of the Dirichlet process. This knowledge was

quickly extended by Antoniak (1974), Doksum (1974) and again by Ferguson (1974). They considered other nonparametric prior distributions and gave further information on the Dirichlet process, e. g. construction methods. Some books appeared later; Schervish (1995) addressed this subject in some sections and the first monograph book in this area was published by Ghosh and Ramamoorthi (2003). A number of concepts and ideas in random measures and the Dirichlet processes in a nonstatistical context may be found in Kingman (1993).

This paper is due first of all to Ferguson (1973, 1974), Antoniak (1974) and Doksum (1974), and a little also to Ghosh and Ramamoorthi (2003).

2. Random probability measures

Definition 2.1. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space, where the class $\mathcal{A}$ of sets is a σ-algebra of Borel subsets of the space $\mathcal{X}$. Let $(\Omega, \mathcal{S}, \Lambda)$ be a probability space. A *random probability measure* (RPM) on $(\mathcal{X}, \mathcal{A})$ is any stochastic process $\{P(A) : A \in \mathcal{A}\}$ on the probability space $(\Omega, \mathcal{S}, \Lambda)$ with the following properties:

(i) $(\forall A \in \mathcal{A})$ $P(A)$ is a random variable on $(\Omega, \mathcal{S}, \Lambda)$ with the range $[0,1]$, i. e. a function $\omega \mapsto P_\omega(A)$ for $\omega \in \Omega$;

(ii) $P(\mathcal{X}) = 1$ almost surely, i. e. $\Lambda(\{\omega \in \Omega : P_\omega(\mathcal{X}) = 1\}) = 1$;

(iii) almost sure countable additivity, i. e. for any countable family of pairwise disjoint sets $\{A_i\} \subset \mathcal{A}$ it holds $P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$ almost surely.

Then P is a well defined stochastic process; in particular it satisfies the Kolmogorov consistency conditions (see e. g. Borovkov, 1972), but the role of the time variable t (or the space variable in case of a random field) is played by a set argument A from the σ-field $\mathcal{A}$. In such a way, for any given $A \in \mathcal{A}$ we obtain a random variable $P(A)$ — the process value and for any fixed $\omega \in \Omega$ we obtain a "usual" (non-random) probability distribution $P_\omega(\cdot)$, which is a path of the process P. If we denote the set of all the probability distributions on $(\mathcal{X}, \mathcal{A})$ by $\mathcal{M} = \mathcal{M}(\mathcal{X}, \mathcal{A})$ then the RPM P is a random variable $P : \Omega \to \mathcal{M}(\mathcal{X}, \mathcal{A})$. We omit the problem of defining random events and the Borel σ-field in the space of measures $\mathcal{M}(\mathcal{X}, \mathcal{A})$ (see Ghosh and Ramamoorthi, 2003). The numerical value $P_\omega(A)$ of the probability of a given event $A \in \mathcal{A}$ for a fixed $\omega \in \Omega$ is a state of the process P.

The distribution of RPM P as a stochastic process is entirely defined by all the finite-dimensional distributions of vectors of the process values; in this case by all the finite-dimensional distributions of the random vectors of the following form: $(P(A_1), P(A_2), ..., P(A_m))$, where the sets $A_1, A_2, ..., A_m$ form a partition of

the space $\mathcal{X}$. Let us denote the distribution generated by all such finite-dimensional distributions by $\mathbf{P}_\Lambda$ and the corresponding expectation by $\mathbf{E}_\Lambda$. The distribution $\mathbf{P}_\Lambda$ is related to the probability Λ from Definition 2.1 in the sense that any m-dimensional distribution Π_m of P is a probability on the Borel measurable space $([0,1]^m, \mathcal{B}([0,1]^m))$ defined for $C \in \mathcal{B}([0,1]^m)$ by the following formula:

$$\Pi_m(C) = \Lambda(\{\omega \in \Omega : (P_\omega(A_1), P_\omega(A_2), ..., P_\omega(A_m)) \in C\}).$$

Theorem 2.1. Let $\rho(A) = \mathbf{E}_\Lambda P(A)$, $A \in \mathcal{A}$. Then

(a) ρ is a probability measure (distribution) on $(\mathcal{X}, \mathcal{A})$;

(b) $(\forall A \in \mathcal{A})$ $P(A) = 0$ almost surely if and only if $\rho(A) = 0$.

Proof is due to Doksum (1974) and is a straightforward consequence of the properties of RPM and probability. To prove countable additivity of ρ we have to notice uniform integrability of the family of random variables $\{P(A) : A \in \mathcal{A}\}$.

The statement in the conclusion (b) "$P(A) = 0$ almost surely" means that $\Lambda(\{\omega \in \Omega : P_\omega(A) = 0\}) = 1$ and such a set of ω's may depend on A. That is why we cannot deduce mutual absolute continuity of the RPM P and the probability distribution ρ.

Example 2.1. We may define a RPM P in such a way that for any $A \in \mathcal{A}$ the variable $P(A)$ has the beta distribution: $P(A) \sim Beta(\alpha(A), \alpha(A'))$, where α is an arbitrarily fixed finite measure.

Example 2.2. Let us consider three simple examples to illustrate the above concepts:

(1) Let X have a distribution P_θ with a density f_θ, where θ is a random variable and it has a prior distribution with a density π. So we consider a parametric Bayesian model. In the parametric model P_θ is a probability distribution for any value of the parameter θ, so the conditions of Definition 2.1 are satisfied.

For an arbitrary $A \in \mathcal{A}$ we have

$$P_\theta(A) = \int_A f_\theta(x)dx,$$

$$E_\pi P_\theta(A) = \int_\Theta \left(\int_A f_\theta(x)dx \right) \pi(\theta)d\theta = \int_A \left(\int_\Theta f_\theta(x)\pi(\theta)d\theta \right) dx =$$

$$= \int_A m_\pi(x)dx = P_{m_\pi}(A),$$

that is we obtain the marginal probability of the event A. Therefore $\rho = P_{m_\pi}$ is the marginal distribution of the data. It can be used to introduce prior information to the statistical model as well (see e. g. Betrò et al., 1994).

(2) Let $\Omega = \{\omega_1, \omega_2\}$; let $(\Omega, 2^\Omega, \lambda)$ be a probability space, $(\mathcal{X}, \mathcal{A})$ — an arbitrary measurable space, P — a RPM. Then

$$P(A) = P_\omega(A) = \begin{cases} P_{\omega_1}(A), & \omega = \omega_1, \\ P_{\omega_2}(A), & \omega = \omega_2. \end{cases}$$

Let $\lambda(\omega_i) = \lambda_i$, $i = 1,2$. We can write

$$\rho(A) = \lambda_1 P_{\omega_1}(A) + \lambda_2 P_{\omega_2}(A).$$

Similarly to the example (1) we can see that the probability measure ρ is the marginal distribution of the data A.

(3) Let $\mathcal{X} = \{x_1, x_2\}$, $\mathcal{A} = 2^\mathcal{X}$ and let $(\Omega, \mathcal{S}, \lambda)$ be an arbitrary probability space and P -- a RPM on $(\mathcal{X}, \mathcal{A})$. Thus we have for any $\omega \in \Omega$

$$P_\omega(\{x_1\}) = p(\omega), \quad P_\omega(\{x_2\}) = 1 - p(\omega),$$

where p is a random variable on [0,1] or more precisely on the space $(\Omega, \mathcal{S}, \lambda)$ with the range [0,1]. Defining a RPM is equivalent to defining the random variable p. Finite-dimensional distributions of the form $(P(A_1), P(A_2), ..., P(A_m))$ are entirely defined by p and the following equalities hold:

$$\mathbf{E}_\lambda P(\{x_1\}) = \mathbf{E}_\lambda p = \rho(\{x_1\}), \quad \mathbf{E}_\lambda P(\{x_2\}) = 1 - \mathbf{E}_\lambda p = \rho(\{x_2\}).$$

In such a way we have defined the distribution $\rho = \mathbf{E}_\lambda P$. The RPM P can be called a random two-point distribution.

Now let us extend the above to a finite sample space $\mathcal{X} = \{x_1, x_2, ..., x_k\}$. Again let $\mathcal{A} = 2^\mathcal{X}$ and P be a RPM on $(\mathcal{X}, \mathcal{A})$. Thus for any $\omega \in \Omega$ we have $P_\omega(\{x_i\}) = p_i(\omega)$, where $\sum_{i=1}^{k} p_i(\omega) = 1$ and we can write $\sum_{i=1}^{k-1} p_i(\omega) \le 1$. Therefore defining a RPM on $(\mathcal{X}, \mathcal{A})$ is equivalent to defining the random vector $\mathbf{p} = (p_1, p_2, ..., p_{k-1})$ on the set

$$J_{k-1} = \left\{ (r_1, r_2, ..., r_{k-1}) \in [0,1]^{k-1} : \sum_{i=1}^{k-1} r_i \le 1 \right\}.$$

We have also $\mathbf{E}_\lambda P(\{x_i\}) = \mathbf{E}_\lambda p_i = \rho_i = \rho(\{x_i\})$, where ρ is a probability on $(\mathcal{X}, \mathcal{A})$ and $\sum_{i=1}^{k-1} \rho_i \leq 1$.

3. Sampling from random probability distributions

Definition 3.1. Let P be a RPM on a measurable space $(\mathcal{X}, \mathcal{A})$. We say that a sequence $X_1, X_2, ..., X_n$ of random variables on a probability space $(\Omega, \mathcal{S}, \Lambda)$ with their values in $(\mathcal{X}, \mathcal{A})$ is an *n-dimensional random sample* from the RPM P, if

$$(\forall m \in \mathbf{N}) \quad (\forall A_1, A_2, ..., A_m, C_1, C_2, ..., C_n \in \mathcal{A})$$

$$\mathbf{P}_\Lambda(X_1 \in C_1, X_2 \in C_2, ..., X_n \in C_n \mid P(C_1), P(C_2), ..., P(C_n), P(A_1), P(A_2), ... \\ ..., P(A_m)) = P(C_1)P(C_2)...P(C_n)$$

almost surely.

The definition means that under the data values $P(C_1), P(C_2), ..., P(C_n)$, which are random variables, the events $\{X_i \in C_i\}$, $1 = 1,2,...,n$, are independent from the "rest" of the process and they are mutually independent.

Corollaries. (1) $\mathbf{P}_\Lambda(X_j \in C_j \mid P(C_1), P(C_2), ..., P(C_n)) = P(C_j)$ almost surely for any $j = 1,2,...,n$.

(2) If $X \sim P$, i. e. X is a one-element random sample from P (a random variable with the distribution P), then $\mathbf{P}_\Lambda(X \in A) = \rho(A)$.

Proof. We have $\mathbf{P}_\Lambda(X \in A \mid P(A)) = P(A)$ almost surely and

$$\mathbf{P}_\Lambda(X \in A) = \mathbf{E}_\Lambda \mathbf{P}_\Lambda(X \in A \mid P(A)) = \mathbf{E}_\Lambda P(A) = \rho(A).$$

4. The Dirichlet process

Let us introduce the notation $Dir(a_1, a_2, ..., a_k)$ for a k-dimensional Dirichlet distribution with the parameters $a_1, a_2, ..., a_k$, i. e. with the following probability density function:

$$f_{a_1, a_2, ..., a_k}(y_1, y_2, ..., y_k) = \frac{\Gamma(a_1 + a_2 + ... + a_k)}{\Gamma(a_1)\Gamma(a_2)...\Gamma(a_k)} y_1^{a_1 - 1} y_2^{a_2 - 1} ... y_k^{a_k - 1},$$

where $a_i > -1,\ y_i > 0,\ i = 1,2,...,k$ and $\sum_{i=1}^{k} y_i = 1$. Because of the last equality this distribution can be actually considered as a $(k-1)$-dimensional distribution with the density

$$f_{a_1,a_2,...,a_k}(y_1, y_2,..., y_{k-1}) = \frac{\Gamma(a_1 + a_2 + ... + a_k)}{\Gamma(a_1)\Gamma(a_2)...\Gamma(a_k)} y_1^{a_1-1} y_2^{a_2-1} ... \left(1 - \sum_{i=1}^{k-1} y_i\right)^{a_k-1},$$

where $\sum_{i=1}^{k-1} y_i \le 1$. So this is a multidimensional generalization of a beta distribution corresponding to the case $k = 2$. We give some properties of the Dirichlet distribution in the form of two lemmas given below.

Lemma 4.1. For the Dirichlet distribution $Dir(a_1, a_2,..., a_k)$ we have

$$\int_0^{z_1} ... \int_0^{z_k} y_j f_{a_1,a_2,...,a_k}(y_1, y_2,..., y_k) dy_1 dy_2 ... dy_k =$$

$$= \frac{a_j}{a_1 + a_2 + ... + a_k} F_D(z_1, z_2,..., z_k \mid a_1,..., a_j + 1,..., a_k),$$

where $F_D(z_1, z_2,..., z_k \mid a_1, a_2,..., a_k)$ is a cumulative distribution function of the distribution $Dir(a_1, a_2,..., a_k)$, i. e.

$$F_D(z_1, z_2,..., z_k \mid a_1, a_2,..., a_k) = P(Y_1 \le z_1, Y_2 \le z_2,..., Y_k \le z_k)$$

for $(Y_1, Y_2,..., Y_k) \sim Dir(a_1, a_2,..., a_k)$.

Proof consists in easy direct computation. Because of the definition of the Dirichlet distribution the above integral is actually $(k-1)$-dimensional.

The next lemma gives a property of marginal distributions and distributions of sums of components of a Dirichlet random vector.

Lemma 4.2. For $(Y_1, Y_2,..., Y_k) \sim Dir(a_1, a_2,..., a_k)$ and $l \le k$, $r_l = k$ there holds

$$\left(\sum_{1}^{r_1} Y_i, \sum_{r_1+1}^{r_2} Y_i,..., \sum_{r_{l-1}+1}^{r_l} Y_i\right) \sim Dir\left(\sum_{1}^{r_1} a_i, \sum_{r_1+1}^{r_2} a_i,..., \sum_{r_{l-1}+1}^{r_l} a_i\right).$$

Now we can define a RPM called a Dirichlet process.

Definition 4.1. Let α be a finite measure on the space $(\mathcal{X}, \mathcal{A})$. We say that a RPM P is a *Dirichlet process* on $(\mathcal{X}, \mathcal{A})$ with the parameter α if for any $k \in \mathbf{N}$ and for any measurable partition $(B_1, B_2,..., B_k)$ of the set $\mathcal{X}$, i. e.

$B_1, B_2, ..., B_k \in \mathcal{A}$, $B_1 \cup B_2 \cup ... \cup B_k = \mathcal{X}$, $B_i \cap B_j = \emptyset$, $i \neq j$, $i, j = 1,2,...,k$, we have

$$(P(B_1), P(B_2), ..., P(B_k)) \sim Dir(\alpha(B_1), \alpha(B_2), ..., \alpha(B_k)).$$

The measure α is often written in the form $\alpha = n_0 P_0$, where $n_0 = \alpha(\mathcal{X})$ and $P_0 = \frac{\alpha}{n_0}$. Then P_0 is a probability distribution. So the phrase "P is a Dirichlet process with the parameter α" or "with the parameters n_0, P_0" can be written in the form $P \sim PD(\alpha)$ or $P \sim PD(n_0, P_0)$. For the latter parametrization n_0 is called the *concentration parameter* and P_0 — a *location parameter*. A larger value of n_0 means that the RPM P is more concentrated around the distribution P_0.

Example 4.1. In case $k = 2$ we have the partition of $\mathcal{X}$ in the form (B, B') and the distributions $P(B) \sim Beta(\alpha(B), \alpha(B')) = Beta(n_0 P_0(B), n_0(1 - P_0(B)))$. This implies

$$\mathbf{E}_\Lambda P = P_0 = \frac{\alpha}{\alpha(\mathcal{X})}, \quad \mathbf{D}^2_\Lambda P = \frac{P_0(1-P_0)}{n_0 + 1} = \frac{\alpha(B)\alpha(B')}{(\alpha(\mathcal{X}))^2(\alpha(\mathcal{X}) + 1)}.$$

In the latter formula we can observe the relation between the concentration and n_0.

The joint probability distribution of the Dirichlet process and the one-element random sample has a property as follows.

Theorem 4.1. For $P \sim PD(\alpha)$ on the space $(\mathcal{X}, \mathcal{A})$, $X \sim P$ - a one-element random sample, a measurable partition $(B_1, B_2, ..., B_k)$ of the space $\mathcal{X}$, an arbitrary set $A \in \mathcal{A}$ the following equality holds

$$\mathbf{P}_\Lambda(X \in A, P(B_1) \le y_1, P(B_2) \le y_2, ..., P(B_k) \le y_k) =$$
$$= \sum_{j=1}^{k} \frac{\alpha(A \cap B_j)}{\alpha(\mathcal{X})} F_D(y_1, y_2, ..., y_k \mid \alpha(B_1), ..., \alpha(B_j) + 1, ..., \alpha(B_k)).$$

Proof. Let $B_{j1} = B_j \cap A$, $B_{j0} = B_j \cap A'$, $j = 1,2,...,k$. Let $Y_{jr} = P(B_{jr})$ for $j = 1,2,...,k$, $r = 1,2$. Then the definition of sampling from a RPM yields

$\mathbf{P}_\Lambda(X \in A \mid Y_{jr}, j = 1,2,...,k, r = 1,2) = P(A) = \sum_{j=1}^{k} Y_{j1}$ almost surely. This follows for any $y_{jr} \in [0,1]$, $j = 1,2,...,k$, $r = 1,2$ that

$$\mathbf{P}_\Lambda(X \in A, Y_{jr} \le y_{jr}, j = 1,2,...,k, \quad r = 1,2) =$$

$$= \int\limits_{\left\{\substack{z_{jr}\le y_{jr},\\ j=1,\ldots,k,\ r=1,2}\right\}} \mathbf{P}_\Lambda(X \in A \mid Y_{jr} = z_{jr},\ j=1,2,\ldots,k,\quad r=1,2)\,dF_D^{(k,j)} =$$

$$= \sum_{j=1}^{k} \frac{\alpha(B_{j1})}{\alpha(\mathcal{X})} F_D(y_{10},\ldots,y_{k0},y_{11},\ldots,y_{k1} \mid \alpha(B_{10}),\ldots,\alpha(B_{j0})+1,\ldots,\alpha(B_{k0}),$$
$$\alpha(B_{11}),\ldots,\alpha(B_{j1})+1,\ldots,\alpha(B_{k1})),$$

where

$$F_D^{(k,j)} = F_D(z_{10},\ldots,z_{k0},z_{11},\ldots,z_{k1} \mid \alpha(B_{10}),\ldots,\alpha(B_{j0})+1,\ldots,\alpha(B_{k0}),$$
$$\alpha(B_{11}),\ldots,\alpha(B_{j1})+1,\ldots,\alpha(B_{k1})),$$

what is a consequence of Lemma 4.1. Since $\alpha(B_{j0}) + \alpha(B_{j1}) = \alpha(B_j)$ and $Y_{j0} + Y_{j1} = P(B_j)$, we obtain the conclusion from Lemma 4.2.

Another property was proved by Ferguson (1973) by using a particular construction of the Dirichlet process.

Theorem 4.2. Let $P \sim PD(\alpha)$ on the space $(\mathcal{X}, \mathcal{A})$, $Z : \mathcal{X} \to \mathbf{R}$ be an $\mathcal{A}$-measurable function. If $\int |Z|\,dP_0 < \infty$, then $\int |Z|\,dP < \infty$ almost surely and $\mathbf{E}_\Lambda \int Z dP = \int Z d(\mathbf{E}_\Lambda P) = \int Z dP_0$.

The above theorem implies e. g. that if k moments of P_0 are finite, then k moments of P are finite almost surely. In general, it reflects close relationship between the Dirichlet process P and the probability distribution P_0 (Ferguson, 1973).

5. The distribution of the Dirichlet process under a random sample – the posterior distribution

Let δ_x denote the following set function:

$$\delta_x(A) = \mathbf{1}_A(x) = \begin{cases} 1, x \in A, \\ 0, x \notin A \end{cases}$$

for any $A \in \mathcal{A}$, $x \in \mathcal{X}$. Let $P \sim PD(n_0, P_0)$ and let $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)$ be a random sample from P.

Theorem 5.1. The conditional distribution of the Dirichlet process P under the variables $X_1, X_2, \ldots, X_n$ is also a Dirichlet process

$$PD\left(n_0 + n, \frac{n_0}{n_0 + n} P_0 + \frac{1}{n_0 + n} \sum_{i=1}^{n} \delta_{X_i} \right),$$

what has for $n_0 P_0 = \alpha$ the form $PD\left(\alpha + \sum_{i=1}^{n} \delta_{X_i} \right)$.

Remark. The following equality holds:

$$\mathbf{E}_\Lambda (P \mid \mathbf{X}_n) = \frac{\alpha(\mathcal{X})\mathbf{E}_\Lambda P + n\hat{P}_n}{n + \alpha(\mathcal{X})} = \frac{n_0 P_0 + n\hat{P}_n}{n + n_0},$$

where $\hat{P}_n$ denotes the empirical distribution derived from the sample $\mathbf{X}_n$. Thus $n_0 = \alpha(\mathcal{X})$ may be interpreted as the level of belief in prior knowledge expressed with the measure $\alpha = n_0 P_0$ or with the parameters n_0, P_0. Taking the context into consideration we may even interpret this as "prior sample size" (Antoniak, 1974).

Proof of Theorem 5.1 may be made by induction in the sample size (Ferguson, 1973). We will show it for $n = 1$, since the inductive step is rather technical.

For analogy and some intuition let us remark first that in the parametric case we have the following equalities for prior, sample, joint, posterior and marginal densities:

$$f(x \mid \theta)\pi(\theta) = p(x, \theta) = \pi(\theta \mid x) m_\pi(x)$$

and consequently we have for $A \in \mathcal{A}$

$$\int_A P_\pi(\theta \le t \mid x) m_\pi(x) dx = P(X \in A, \theta \le t), \qquad (*)$$

where P is in $(*)$ a joint probability distribution of (X, θ) for the parametric case.

The idea of the proof consists in analogies with the formula $(*)$. Since we know the form of the marginal distribution of the data in the Bayesian nonparametric model with the Dirichlet process prior analogous to m_π on the left hand side of $(*)$, we will find the joint distribution of (X, P), analogous to the right hand side of $(*)$, in the form of an integral representation, which will give us the form of the posterior distribution analogous to the term $P_\pi(\theta \le t \mid x)$ on the left hand side of $(*)$.

So let $(B_1, B_2, ..., B_k)$ be a measurable partition of the space $\mathcal{X}$, $A \in \mathcal{A}$, $X \sim P$. Then

$$\int_A F_D(y_1, y_2, ..., y_k \mid \alpha(B_1) + \delta_x(B_1), \alpha(B_2) + \delta_x(B_2), ..., \alpha(B_k) + \delta_x(B_k)) \times$$
$$\times (\alpha(\mathcal{X}))^{-1} d\alpha(x) =$$
$$= \sum_{j=1}^{k} \frac{\alpha(A \cap B_j)}{\alpha(\mathcal{X})} F_D(y_1, y_2, ..., y_k \mid \alpha(B_1), ..., \alpha(B_j) + 1, ..., \alpha(B_k)).$$

By Theorem 4.1 the last expression equals to

$$\mathbf{P}_\Lambda(X \in A, P(B_1) \le y_1, P(B_2) \le y_2, ..., P(B_k) \le y_k),$$

but this is the joint distribution of X and of the k-dimensional vector of the values of the process P. Thus since $\frac{\alpha}{\alpha(\mathcal{X})}$ is the marginal distribution of the observation X, then

$$F_D(y_1, y_2, ..., y_k \mid \alpha(B_1) + \delta_x(B_1), \alpha(B_2) + \delta_x(B_2), ..., \alpha(B_k) + \delta_x(B_k))$$

is actually the posterior cumulative distribution function.

6. Simple applications

The examples below are due to Ferguson (1973).

6.1. Estimation of a distribution with quadratic loss

Let $\mathcal{X} = \mathbf{R}$, $\mathcal{A} = \mathcal{B}(\mathbf{R})$ — the Borel σ-field in $\mathbf{R}$. Let $P \sim PD(\alpha)$, $F(t) = P((-\infty, t])$. Further, let $P_0 = \frac{\alpha}{\alpha(\mathbf{R})}$, $n_0 = \alpha(\mathbf{R})$, F_0 — a cdf of P_0, $\hat{F}$ — an estimator of the cdf F. The loss function is defined as follows:

$$L(F, \hat{F}) = \int_R (F(t) - \hat{F}(t))^2 dW(t),$$

where W is a weight function (a cdf of a finite measure) on $\mathbf{R}$. Then $F(t) \overset{D}{\approx} Beta(n_0 F_0(t), n_0(1 - F_0(t)))$. Let the finite sequence of random variables $\mathbf{X}_n = (X_1, X_2, ..., X_n)$ be a random sample from a Dirichlet process P. Minimization of the expectation $\mathbf{E}_\Lambda(L(F, \hat{F}) \mid \mathbf{X}_n)$ with respect to $\hat{F}$ results in the following Bayesian estimator of the cdf F:

$$\hat{F}_n^B(t) = \mathbf{E}_\Lambda(F(t) \mid \mathbf{X}_n) = \frac{n_0}{n + n_0} F_0(t) + \frac{n}{n + n_0} \hat{F}_n(t)$$

(where $\hat{F}_n$ denotes the empirical cdf with respect to the sample $\mathbf{X}_n$), since the conditional (posterior) distribution of the random variable $F(t)$ under $\mathbf{X}_n$ has the form $Beta(n_0 F_0(t) + n\hat{F}_n(t), n_0(1-F_0(t)) + n(1-\hat{F}_n(t)))$.

6.2. Estimation of a median

As before, let $\mathcal{X} = \mathbf{R}$, $\mathcal{A} = \mathcal{B}(\mathbf{R})$ and let P be a RPM on $(\mathbf{R}, \mathcal{B}(\mathbf{R}))$. It can be seen that if $P \sim PD(\alpha)$, then the median of the random distribution P is well defined (unique) almost surely. Certainly, multiple medians may be located only in α-zero-measure interval: there exist at most countably many such intervals (because they correspond to discontinuities of the distribution) and the probability that such an interval J is an interval of medians equals to $\mathbf{P}_\Lambda\left(F(t) = \frac{1}{2}\right)$ for any $t \in \operatorname{int} J$. But since $F(t)$ has a beta distribution, this probability is zero. Thus for a Dirichlet process the median m is a well defined random variable.

Let us define the absolute deviation loss $L(m, \hat{m}) = |m - \hat{m}|$. Then the minimizer with respect to $\hat{m}$ is any median of the distribution $P^{(m)}$ of the random variable m. If the underlying distribution is a Dirichlet process then any median of the distribution of the random variable m is a median of the expectation of the process and vice versa:

$$\operatorname{med} P^{(m)} = \operatorname{med}(\mathbf{E}_\Lambda P) = \operatorname{med} P_0 .$$

We prove this as follows: a number t is a median of the distribution of the variable m if and only if $\mathbf{P}_\Lambda(m < t) \le \frac{1}{2} \le \mathbf{P}_\Lambda(m \le t)$. By the definition of a median we can write equivalently

$$\mathbf{P}_\Lambda\left(F(t) < \tfrac{1}{2}\right) \le \tfrac{1}{2} \le \mathbf{P}_\Lambda\left(F(t) \le \tfrac{1}{2}\right).$$

But $F(t)$ has the distribution $Beta(n_0 F_0(t), n_0(1-F_0(t)))$ and its median is nondecreasing in t and equal to $\frac{1}{2}$ if and only if it is the median of P_0. Therefore the Bayesian estimator of the median m under a a sample $\mathbf{X}_n$ denoted by $\hat{m}^B(\mathbf{X}_n)$ is the median of the distribution of $\mathbf{E}_\Lambda(P \mid \mathbf{X}_n)$ which has the cdf $\hat{F}_n^B$: $\hat{m}^B(\mathbf{X}_n) = \operatorname{med} \hat{F}_n^B$.

6.3. General remarks

Contemporary applications of Bayesian nonparametric statistical methods go beyond simple problems of statistical inference. Examples can be seen in e. g.

Ghosh and Ramamoorthi (2003); see also a recent application in Niemiro (2006). The last well attended international meeting, *Fourth Workshop on Bayesian Nonparametrics: Methodology, Theory and Applications* was held in 2005 at the University of Rome "La Sapienza". Recently Bayesian nonparametric inference has become a frequently used technique of statistical research and it has also remained an area of theoretical investigations which are still wanted very much.

7. Discreteness of the Dirichlet process

When making use of the Dirichlet process we exploited the fact that for any fixed $A \in \mathcal{A}$ the process value $P(A)$ is a random variable with a known distribution. From the point of view of the above theory almost sure discreteness of the Dirichlet process seems to be a little unexpected. We mean that for almost all $\omega \in \Omega$ the distribution P_ω, which is a path of the process P, is discrete. There exist a number of proofs for this property. In Ferguson (1973) it was a consequence of a construction of the Dirichlet process and some more equivalent constructions were given elsewhere. The direct proof was given by Blackwell (1973). We will show a proof based on Basu and Tiwari (1982), following Ghosh and Ramamoorthi (2003) and giving more detailed argument.

Theorem 7.1. Let $\mathcal{X}$ be a Polish space (metric, separable and complete), $\mathcal{A} = \mathcal{B}(\mathcal{X})$ — a σ-field of Borel subsets of $\mathcal{X}$, $(\Omega, \mathcal{S}, \Lambda)$ — an underlying probability space introduced in Definition 2.1, P — a Dirichlet process $DP(\alpha)$ on the space $(\mathcal{X}, \mathcal{A})$. Let $\mathcal{M} = \mathcal{M}(\mathcal{X}, \mathcal{A})$ be the set of all the probability distributions on $(\mathcal{X}, \mathcal{A})$. Then $\mathbf{P}_\Lambda(\{Q \in \mathcal{M} : Q \text{ is discrete}\}) = 1$.

Proof. We will show that any observed result $x \in \mathcal{X}$ of a random variable $X \sim P$ has positive probability what means that the Dirichlet process is discrete in the given sense. Let

$$H = \{(Q, x) \in \mathcal{M} \times \mathcal{X} : Q(\{x\}) > 0\}.$$

A probability distribution Q is discrete if and only if $\sum_{\{x:(Q,x)\in H\}} Q(\{x\}) = 1$.

The set H is measurable (is a random event) with respect to the product Borel σ-field $\mathcal{B}(\mathcal{M} \times \mathcal{X})$. We need to show that the function which transforms $(Q, x) \in \mathcal{M} \times \mathcal{X}$ to $Q(\{x\}) \in [0,1]$ is measurable (is a random variable). Let us consider the class $\mathcal{F}$ of all measurable sets $F \subset \mathcal{X} \times \mathcal{X}$ such that for

$$F^x \overset{def}{=} \{y \in \mathcal{X} : (x, y) \in F\}$$

the mapping $(Q, x) \mapsto Q(F^x)$ is measurable. It contains all the Borel sets of the form $B_1 \times B_2$ which form so-called π-system and it is a λ-system (for π and λ-systems, see Schervish, 1995, or Ghosh and Ramamoorthi, 2003), so

$$\mathcal{F} \supset \sigma(\{B_1 \times B_2 : B_i \in \mathcal{B}(\mathcal{X}),\ i = 1,2\}).$$

But $\mathcal{F} \subset \mathcal{B}(\mathcal{X} \times \mathcal{X})$, what follows that $\mathcal{F} = \mathcal{B}(\mathcal{X} \times \mathcal{X})$. In particular we may take $F = \{(x, x) \in \mathcal{X} \times \mathcal{X}\}$, from which we have $F^x = \{x\}$. Thus we obtain $H = \{(Q, x) \in \mathcal{M} \times \mathcal{X} : Q(F^x) > 0\}$ as a measurable set.

Now let

$$H_x = \{Q \in \mathcal{M} : Q(x) > 0\}$$

and

$$H_Q = \{x \in \mathcal{X} : Q(x) > 0\}.$$

By the above proof of measurability we have $H_x \in \mathcal{B}(\mathcal{M})$ and $H_Q \in \mathcal{B}(\mathcal{X}) = \mathcal{A}$. A probability distribution Q is discrete if and only if $Q(H_Q) = 1$.

Let $X \sim P$ and $P \sim DP(\alpha)$. Then

$$\mathbf{P}_\Lambda(H) = \mathbf{P}_\Lambda((P, X) \in H) = \Lambda(\{\omega \in \Omega : P_\omega(\{X(\omega)\}) > 0\})$$

but also

$$\begin{aligned}\mathbf{P}_\Lambda(H) &= \mathbf{E}_\Lambda \mathbf{P}_\Lambda(H \mid X) = \mathbf{E}_\Lambda \mathbf{P}_\Lambda((P, X) \in H \mid X) = \\ &= \mathbf{E}_\Lambda \mathbf{P}_\Lambda(P \in H_X \mid X) = \mathbf{E}_\Lambda \mathbf{P}_\Lambda(P(\{X\}) > 0 \mid X).\end{aligned}$$

We remind that the conditional distribution of P under X is $DP(\alpha + \delta_x)$. Therefore $P(\{X\})$ has the conditional distribution under $X = x$ of the form $Beta(\alpha(\{x\}) + 1, \alpha(\mathcal{X} - \{x\}))$, so it is positive $\mathbf{P}_\Lambda$-almost surely. From the above we can write

$$\mathbf{P}_\Lambda(H) = \mathbf{E}_\Lambda \mathbf{P}_\Lambda(P(\{X\}) > 0 \mid X) = \mathbf{E}_\Lambda \mathbf{P}_\Lambda(P_{\alpha + \delta_X}(X) > 0) = 1.$$

Now, further

$$\begin{aligned}1 = \mathbf{P}_\Lambda(H) &= \mathbf{E}_\Lambda \mathbf{P}_\Lambda(H \mid P) = \mathbf{E}_\Lambda \mathbf{P}_\Lambda((P, X) \in H \mid P) = \\ &= \mathbf{E}_\Lambda \mathbf{P}_\Lambda(X \in H_P \mid P) = \mathbf{E}_\Lambda \mathbf{P}_\Lambda(H_P \mid P) = \mathbf{E}_\Lambda P(H_P),\end{aligned}$$

what follows $P(H_P) = 1$ $\mathbf{P}_\Lambda$-almost surely.

The discreteness of the Dirichlet process need not to be more troublesome in applications than for example the discreteness of an empirical distribution. To see this let us consider the following property of the Dirichlet process (Ferguson, 1973).

Theorem 7.2. Let $P \sim DP(\alpha)$ on $(\mathcal{X}, \mathcal{A})$ and let Q be a fixed probability distribution on $(\mathcal{X}, \mathcal{A})$ absolutely continous with respect to the measure α, i. e. $(\forall A \in \boldsymbol{A})\, \alpha(A) = 0 \Rightarrow Q(A) = 0$. Then for any $m \in \mathbf{N}$, any sets $A_1, A_2, ..., A_m \in \boldsymbol{A}$ and any $\varepsilon > 0$ it holds

$$\mathbf{P}_\Lambda (| P(A_i) - Q(A_i) | < \varepsilon,\ i = 1,2,...,m) > 0.$$

The conclusion of Theorem 7.2 means, among others, that with positive probability the random distribution P is arbitrarily close in the given sense to any probability distribution Q which is absolutely continuous with respect to the measure α, including the case of continuous measures α and Q. That is why we claim that the discreteness of the Dirichlet process does not „disturb". But it can make impossible to establish a nonparametric Bayesian model with prescribed properties if continuity is desired. Therefore research on others RPM started quite early, beginning from mixtures of Dirichlet processes (Antoniak, 1974) and more general classes of processes (Doksum, 1974).

Acknowledgments

The research was supported by the Ministry of Science and Higher Education program 03/S/0023/05 at Warsaw School of Economics.

REFERENCES

C. ANTONIAK (1974), Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics* **2**, pp.1152—1174.

D. BASU and R. C. TIWARI (1982), A note on the Dirichlet process. In*: Statistics and Probability: Essays in Honor of C. R. Rao*, pp. 89—103, North Holland, Amsterdam 1982.

B.BETRÒ, M. MĘCZARSKI and F. RUGGERI (1994), Robust Bayesian analysis under generalized moments conditions. *Journal of Statistical Planning and Inference* **41**, pp.257—266.

D. BLACKWELL (1973), Discreteness of Ferguson selections. *Annals of Statistics* **1**, pp. 356—358.

A. A. BOROWKOW (1972), *A Course in Probability Theory*. Nauka, Moskva.

K. DOKSUM (1974), Tailfree and neutral random probabilities and their posterior distributions. *Annals of Probability* **2**, pp. 183—201.

T. FERGUSON (1973), A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, pp. 209—230.

T. FERGUSON (1974), Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, pp. 615—629.

J. K. GHOSH and R. V. RAMAMOORTHI (2003), *Bayesian Nonparametrics*. Springer Verlag, New York.

J. F. C. KINGMAN (1993), *Poisson Processes*. Oxford University Press.

W. NIEMIRO (2006), Bayesian prediction with an asymmetric criterion in a nonparametric model of insurance. *Statistics* **40**, pp. 353—363.

M. SCHERVISH (1995), *Theory of Statistics*. Springer Verlag, New York.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 593—599

ASYMPTOTIC TESTS FOR RECEIVER OPERATING CHARACTERISTIC CURVES

Agnieszka Rossa[*]

ABSTRACT

In the paper two significance tests for Receiver Operating Characteristic Curves (*ROC*) are proposed. Both tests are based on an asymptotic χ^2 distribution of test statistics.

Key words*:* sensitivity, specificity, *ROC* curves, goodness-of-fit test, homogeneity test.

1. Notation

Suppose a medical diagnostic test is used to detect the presence of a disease. Denote by π_0, π_1 a disease group and a control group, respectively. Let X be a continuous random variable representing the test result. We will assume the following classification rule: an individual is classified to the population π_0 if the test result X exceeds a fixed threshold x, and to the population π_1, otherwise.

Let us consider diagnostic test results of individuals randomly drawn from the populations π_0, π_1, respectively. Both random variables will be hereafter denoted by

$$C = X|\pi_0, \quad Z = X|\pi_1$$

Let F and G be cumulative distribution functions (*CDF*'s) of C and Z, respectively.

[*] Dept. of Statistical Methods, University of Łódź, Rewolucji 1905, nr 41, 90-214 Łódź, Poland; e-mail: agrossa@uni.lodz.pl

2. The *ROC* Curve

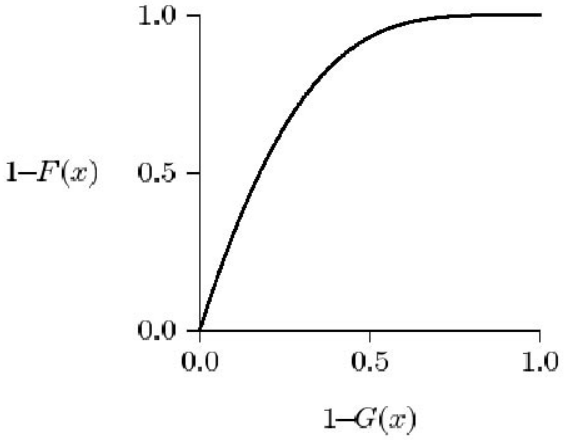
The *ROC* curve (Receiver Operating Characteristic curve, see: Green, Swets 1966; Lloyd 1998) is a plot of $p_0 = 1 - F(x)$ against $p_1 = 1 - G(x)$ as x varies over the support of X. In biomedical context p_0 is termed *sensitivity*, and $1 - p_1$ is termed *specificity*. Thus, p_0 is a probability $P\left(X > x | \pi_0\right)$ of a true positive diagnosis and p_1 is a probability $P\left(X > x | \pi_1\right)$ of a false positive diagnosis.

It can be shown that *ROC* depends on F and G *via* the formula

$$ROC(v) = 1 - F\left(G^{-1}(1-v)\right), \quad v \in [0,1]. \tag{1}$$

Indeed, let $v = 1 - G(x)$, then $G(x) = 1 - v$, and $x(v) = G^{-1}(1-v)$. Thus, for $v \in [0,1]$ we have $ROC(v) = 1 - F\left(x(v)\right)$ what leads to (1).

Figure 1. An example of a *ROC* curve



Estimation of *ROC(v)* is usually based on replacing the unknown functions F and G by their empirical counterparts, say F_m and G_n, defined as follows

$$F_m(x) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(C_i \le x), \tag{2}$$

$$G_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Z_i \le x), \tag{3}$$

where $\mathbf{1}(\cdot)$ denotes an indicator function, and

$$C_1, C_2, \ldots, C_m, \qquad Z_1, Z_2, \ldots, Z_n \tag{4}$$

are independent random samples drawn from the populations π_0, π_1, respectively.

3. Some properties of *ROC* curves

The *ROC* curve summarizes the separation between the distributions *F* and *G* in two populations π_0 and π_1. The higher is a *ROC* curve, the greater is the prediction accuracy of *X*. If the *ROC* curve lies on the diagonal of the unit space then there is no difference in distributions of *X* in the populations π_0 and π_1.

We will show that *ROC*, defined by (1), can be treated as a *CDF* of the variable $W = 1 - G(C)$. We have, for $v \in [0,1]$

$$P(W < v) = P(1 - G(C) < v) = P(G(C) > 1 - v) = P(C > G^{-1}(1 - v)) =$$
$$= 1 - P(C \le G^{-1}(1 - v)) = 1 - F\left(G^{-1}(1 - v)\right) = ROC(v).$$

Unfortunately, it is usually impossible to observe *G*(*C*) without any parametric assumptions concerning the unknown distribution function *G*. However, if we replace the unknown function *G* by its empirical counterpart G_n, defined in (3), then we obtain a fully observable random variable $G_n(C)$.

First, we will find the probability distribution function of $G_n(C)$. It can be seen that $G_n(C)$ takes values from the following finite set

$$\left\{0, \frac{1}{n}, \ldots, \frac{n-1}{n}, 1\right\}.$$

Let *R* be a *CDF* of *G*(*C*). Let also denote by *r* and *f* density functions of *G*(*C*) and *C*, respectively. Note that for any $x \in [0,1]$ the following equalities hold

$$R(x) = F\left(G^{-1}(x)\right), \qquad r(x) = f\left(G^{-1}(x)\right)\left[G^{-1}(x)\right]'. \tag{5}$$

Thus, for a fixed integer $i \in \{0,1,\ldots,n\}$ we have

$$P\left(G_n(C) = \frac{i}{n}\right) = \binom{n}{i} \int_{-\infty}^{\infty} G^i(x)\left[1 - G(x)\right]^{n-i} f(x)dx.$$

Denoting by $y = G(x)$ we obtain

$$x = G^{-1}(y), \quad dx = \left[G^{-1}(y)\right]' dy,$$

and hence

$$P\left(G_n(C)=\frac{i}{n}\right)=\binom{n}{i}\int_0^1 y^i(1-y)^{n-i}\,f\left(G^{-1}(y)\right)\left[G^{-1}(y)\right]'dy.$$

This result together with (5) lead to the probability distribution function of $G_n(C)$

$$P\left(G_n(C)=\frac{i}{n}\right)=\binom{n}{i}\int_0^1 y^i(1-y)^{n-i}r(y)dy, \qquad i\in\{0,1,\ldots,n\}. \tag{6}$$

4. The goodness-of-fit and homogeneity tests for *ROC* curves

The properties of $G_n(C)$ presented in previous section will be used to define two test statistics for testing two separate null hypotheses. The first one assumes that a *ROC* curve lies on the diagonal, that is

$$H_0: \quad \underset{v\in[0,1]}{\forall} ROC(v)=v, \tag{7}$$

against the general alternative $H_1: \quad \sim H_0$.

The second null hypothesis assumes that *ROC* curves for two diagnostic variables, say X_1 and X_2, are equal

$$H_0': \quad \underset{v\in[0,1]}{\forall} ROC_1(v)=ROC_2(v), \tag{8}$$

against the alternative $H_1': \quad \sim H_0'$.

4.1. Testing the null hypothesis H_0

Note that if the null hypothesis (7) is true, then $r(y)=1$ for any $y\in[0,1]$ and (6) reduces to

$$P\left(G_n(C)=\frac{i}{n}\middle|H_0\right)=\binom{n}{i}\int_0^1 y^i(1-y)^{n-i}dy, \qquad i\in\{0,1,\ldots,n\}. \tag{9}$$

Let $B(\alpha,\beta)$ denote a beta function with parameters α,β, that is

$$B(\alpha,\beta)=\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}dy. \tag{10}$$

It is well-known that the following equalities hold

$$B(\alpha,\beta)=\Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta), \qquad \Gamma(n+1)=n! \qquad n\in\mathbf{N}, \tag{11}$$

where $\Gamma(\cdot)$ denotes a gamma function.

Thus, from (9)-(11) we have

$$\begin{aligned} P\left(G_n(C)=\frac{i}{n}\middle|H_0\right)&=\binom{n}{i}\int_0^1 y^i(1-y)^{n-i}dy=\binom{n}{i}B(i+1,n-i+1)= \\ &=\binom{n}{i}\frac{\Gamma(i+1)\Gamma(n-i+1)}{\Gamma(n+2)}=\binom{n}{i}\frac{i!(n-i)!}{(n+1)!}=\frac{1}{n+1}. \end{aligned} \tag{12}$$

Let us consider a random sequence

$$G_n(C_1),G_n(C_2),\ldots,G_n(C_m) \tag{13}$$

derived by a transformation of sequences (4).

For testing the null hypothesis (7) we propose the following test statistic

$$Z_1=\sum_{i=0}^{n}\frac{(m_i-mp)^2}{mp}, \tag{14}$$

where *m* is a size of the sample (13), $p=1/(n+1)$ represents the theoretical probability (12) that $G_n(C)=i/n$ (under H_0), and m_i stands for an empirical number of observations in the sequence (13) which are equal to i/n.

Under H_0 the distribution of the test statistic Z_1 tends to the χ^2 distribution with *n* degrees of freedom if $m\to\infty$.

4.2. Testing the null hypothesis H_0'

Let X_1, X_2 be two diagnostic variables, and π_0,π_1 be two populations of individuals under consideration.

According to the notation assumed let us introduce the following random variables

$$C_1=X_1|\pi_0, \quad C_2=X_2|\pi_0, \quad Z_1=X_1|\pi_1, \quad Z_2=X_2|\pi_1.$$

We will consider *m* independent copies of C_1, *k* independent copies of C_2, and *n* independent copies of both Z_1 and Z_2, that is

$$C_{11},C_{12},\ldots,C_{1m}, \qquad C_{21},C_{22},\ldots,C_{2k}. \tag{15}$$

$$Z_{11}, Z_{12}, \ldots, Z_{1n}, \qquad Z_{21}, Z_{22}, \ldots, Z_{2n}. \tag{16}$$

In other words, we will assume that four independent random samples are observed: two random samples (15) drawn from the population π_0 and two random samples (16) drawn from the population π_1. Let us also consider transformations of (15)-(16) performed by the analogy to (13)

$$G_{1n}(C_{11}), G_{1n}(C_{12}), \ldots . G_{1n}(C_{1m}), \tag{17}$$

$$G_{2n}(C_{21}), G_{2n}(C_{22}), \ldots . G_{2n}(C_{2k}), \tag{18}$$

where G_{1n}, G_{2n} represent empirical distribution functions (see the formula (3)) derived from the sequences $Z_{11}, Z_{12}, \ldots, Z_{1n}$ and $Z_{21}, Z_{22}, \ldots, Z_{2n}$, respectively.

If the null hypothesis (8) is true, then all the variables $G_{1n}(C_{1j}), \quad j = 1,2,\ldots,m$ and $G_{2n}(C_{2j}), \quad j = 1,2,\ldots,k$ given in (17)-(18) are iid with a common probability distribution function expressed by (6).

Let us assume the following notation

$$\begin{aligned} N_{1i} &= \sum_{j=1}^{m} \mathbf{1}\left(G_{1n}(C_{1j}) = \frac{i}{n} \right), \quad i = 0,1,\ldots,n, \\ N_{2i} &= \sum_{j=1}^{k} \mathbf{1}\left(G_{2n}(C_{2j}) = \frac{i}{n} \right), \; i = 0,1,\ldots,n, \\ N_{\bullet i} &= \sum_{s=1}^{2} N_{si}, \quad N_{1\bullet} = \sum_{i=0}^{n} N_{1i} = m, \quad N_{2\bullet} = \sum_{i=0}^{n} N_{2i} = k, \; N = m + k. \end{aligned} \tag{19}$$

All the statistics defined in (19) are summarized in Table 1.

Table 1. A contingency table

i/n	$G_{1n}(C_1)$	$G_{2n}(C_2)$	$\sum$
0	N_{10}	N_{20}	$N_{\bullet 0}$
$1/n$	N_{11}	N_{21}	$N_{\bullet 1}$
…	…	…	…
$(n-1)/n$	$N_{1\,n-1}$	$N_{2\,n-1}$	$N_{\bullet n-1}$
1	N_{1n}	N_{2n}	$N_{\bullet n}$
$\sum$	$N_{1\bullet} = m$	$N_{2\bullet} = k$	$N = m + k$

For testing the null hypothesis (8) we propose the test statistic of the form

$$Z_2 = N\sum_{i=0}^{n} \frac{\left(N_{1i} - \dfrac{N_{1\bullet}N_{\bullet i}}{N}\right)^2}{N_{1\bullet}N_{\bullet i}} + N\sum_{i=0}^{n} \frac{\left(N_{2i} - \dfrac{N_{2\bullet}N_{\bullet i}}{N}\right)^2}{N_{2\bullet}N_{\bullet i}}. \qquad (20)$$

Under H_0' the statistic Z_2 tends in distribution to the χ^2 distribution with *n* degrees of freedom if $N \to \infty$.

REFERENCES

GREEN D.M., SWETS J.A., (1966), *Signal Detection Theory and Psychophysics*, New York.

LLOYD C.J., (1998), Using Smoothed Receiver Operating Characteristic Curves to Summarize and Compare Diagnostic Systems, *JASA* 93, pp. 1356—1364.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 601—609

Report from the Small Area Estimation 2007 Conference (SAE2007)

Pisa, Italy, 3—5 September 2007

The SAE 2007 Conference was held at the Faculty of Economics of the University of Pisa in Tuscany in central Italy. The conference was organized by the International Association of Survey Statisticians (IASS) in co-operation with the University of Pisa and the Regione Toscana. The Scientific Committee consisted of: Luigi Biggeri (Chairman), Ray Chambers, J.N.K. Rao, Danny Pfeffermann, Jean Opsomer, Daniela Cocchi, Andrea Giommi, Monica Pratesi, Piero Manfredi, Nicola Torelli, Silvia Biffignandi, Risto Lehtonen, Timo Alanko, Philip Clarke, Dan Hedlin, Jan Kordos, Li-Chun Zhang, Domingo Morales and Montserrat Herrador.

The conference participants, representing universities, statistical agencies and other institutions, came from 23 countries. During the Conference more than 70 papers were presented at plenary, specialized and contributed sessions. The largest number of participants (except Italy) arrived from the USA (12 people), Poland (11) and the UK (9). The purpose of this Conference was to stimulate research both in theoretical and methodological developments in SAE and related fields, and practical applications of SAE methods, including their potential usage in various research areas.

Topics discussed during the Conference included:

- model-assisted and model-dependent estimation for domains;
- sample design and weighting in the context of SAE;
- design-based versus model-based approaches for SAE;
- parametric versus non parametric small area estimation techniques;
- combining hierarchical models with autocorrelation, state-space models, spatial econometrics, spatial epidemiology and disease mapping;
- case studies of SAE in official statistics and agricultural, environmental, health statistics, marketing research, poverty mapping; model-misspecification;
- robustness and diagnostics.

The list of papers is presented in the annex.

The keynote lecture, given by J.N.K. Rao presented methodological developments in SAE. The invited papers were presented by **R. Chambers** on robust MSE estimation for linear predictors for domains, by **D. Cocchi** on

Bayesian least squares approximations in SAE, by **W. Fuller** on small area prediction subject to a restriction, by **J. Jiang** on fence method for SAE, by **D. Morales** on small area estimation of totals of unemployed and employed people in the Spanish labour force survey, by **D. Pfeffermann, Benedicte Terryn and Fernando Moura** on Bayesian small area estimation of literacy under a two part random effects model.

Contributed and specialized sessions covered broad area of topics. They are listed in the Annex.

As a part of the conference program, ***the full-day course on SAE techniques and tools*** intended for statistics practitioners was prepared by **P. Lahiri**. It provided an introduction to some important concepts in SAE and outlined various approaches for estimating various small area parameters. The detailed topics included standard design-based methods, various traditional indirect methods and the state-of-the-art small-area estimation methods that use mixed models. Data analyses using several empirical examples were also presented.

On the first day of the Conference a ***Panel Discussion on small area estimation in official statistics*** was organized by **Monica Pratesi** with **Paola Baldi** as chairman (Tuscany Region). **N. Torelli** (University of Trieste**), D. Hedlin** (Statistics Sweden), **S. Rubin-Bleuer** (Statistics Canada) and **J. Kordos** (Warsaw School of Economics) were participants of the panel.

N. Torelli focused on the problem of the gap between theoretical developments in SAE and their current use for the production of official statistics. The increasing demand expressed by local government bodies in Italy does not meet the estimates produced by ISTAT for the expected level of territorial aggregation. In N. Torelli's opinion using SAE methods implies technical problems that are very often beyond the complexity that can be managed by local statistical offices (complex models involved, lack of standard software and non standard tools for evaluating the model). He pointed out the solution: promotion of effective cooperation between users (local authorities), researchers in NSIs and universities. He also presented two examples of such cooperation. The Istat and CISIS (Board of the statistical offices of the Italian regions) project aimed at the construction of a system for the computation of small area estimates and ABS (Australian Bureau of Statistician) publication of a Guide To SAE.

D. Hedlin started with the defining the present position of SAE within the area of statistical research development. In his opinion SAE is somewhere in the middle of its development and because of that the choice of method is often subjective. He declared some principles to be agreed by statistical offices: good theoretical foundations of the SAE methodology, application of common measures of the quality of estimators, cooperation between different types of users especially in the field of methodology and the quality of the data. He also discussed three other issues of common use of SAE methodology in statistical offices: quality and communication, register-based statistics and resource allocation. The first one is connected with the fear of misuse of produced small

area estimates (sometimes associated with great uncertainty), in particular as part of official statistics. The problems with register-based statistics arise when the register data suffer from lack of relevance in relation to the goals of the survey. The last one addresses the issue of sampling and non-sampling errors. The conclusion is that NSIs should consider reducing sample sizes below the limit where direct estimation is sustainable and instead spend resources on reducing and estimating non-sampling errors.

S. Rubin-Bleuer presented experience of Statistics Canada in SAE. Statistics Canada tries to identify small area needs at the planning stage of a survey and allocate the sample to them. At the estimation stage primary consideration is given to model assisted estimators. Only if these estimators built with all the available auxiliary data cannot be used, the Agency produces SAEs. Before using model based methods, assumed models undergo careful evaluation for consistency and accuracy, not only internally, through extensive model diagnostics and simulation, but by comparison of estimates with sources of local knowledge. Methods with built-in benchmarking are preferable. Problems of measuring accuracy of small area estimates, availability of auxiliary data for business surveys and limitations connected with auxiliary data in person based files were also discussed. Work towards a user-friendly general software package for generating SAEs is ongoing at Statistics Canada and such a product is already in use for some of projects.

J. Kordos gave the overview of the Polish experiences with survey methodology, use of administrative registers and use of models in producing estimates before 1989. He also pointed out importance of four events in SAE research in Poland after 1989: *International Scientific Conference on Small Area Statistics and Survey Designs* held in Warsaw in 1992; *International Conference on Small Area Estimation held in Riga*, Latvia, in 1999, the *EURAREA project* (2001-2004) and *Conference in Finland* (SAE2005). The last part of J. Kordos' speech was devoted to attempts of application of SAE methods in estimation of selected characteristics covering the following fields: employment and unemployment, broken down by region and powiat (county), small business, agriculture figures, broken down by region and powiat using agricultural sample surveys and agricultural census data.

During the Conference the idea of **European Working Group on Small Area Estimation** (EWORSAE) was put forward and some very important decisions to establish it were made. The web page of EWORSAE has been created by Tomasz Jurkiewicz from University of Gdansk, Poland (http://www.sae.wzr.pl). Its aim is to build and to maintain a network of researchers and statisticians in the particular field of SAE. The working group is opened to all people working in the SAE field. Its main activities include: organization of SAE conferences, workshops, courses and preparation of common research project proposals. The coordinator of the EWORSAE **is Domingo Morales** (d.morales@umh.es) and the *First Council* consists of: *Bart*

Buelens (Statistics Netherlands), Philip Clarke (Statistics UK (ONS)), Kari Djerf (Statistics Finland), Stefano Falorsi (Statistics Italy (ISTAT)), Elżbieta Gołata (University of Economics in Poznań, Poland), Dan Hedlin (Statistics Sweden), Montserrat Herrador (Statistics Spain (INE)), Risto Lehtonen (University of Helsinki, Finland), Domingo Morales (University Miguel Hernández de Elche, Spain), Monica Pratesi (University of Pisa, Italy), Nicos Tzavidis (University of Manchester, the UK).

The conference was sponsored by Fondazione Cassa di Risparmio di Lucca, Italian National Statistical Institute (ISTAT), Italian Statistical Society (SIS), University of Bergamo, University of Cassino, University of Florence, University of Perugia and the University of Trieste.

Prepared by: Tomasz Klimanek (University of Economics in Poznań), and Tomasz Żądło (University of Economics in Katowice).

ANNEX

The list of titles and authors at the SAE 2007 Conference:

Keynote Paper:

1. *J.N.K. Rao* Methodological developments in Small Area Estimation: overview and appraisal.

Invited Papers:

2. *Ray Chambers* Robust estimation of mean squared error for linear SAE.
3. *Daniela Cocchi* Bayesian least squares approximations in small area models.
4. *Wayne Fuller* Small Area Estimation subject to a restriction.
5. *Jiming Jiang* Fence methods for Small Area Estimation.
6. *Domingo Morales* Small Area Estimation of totals of unemployed and employed people in the Spanish labour force survey.
7. *Danny Pfeffermann* Bayesian Small Area Estimation of literacy under a two part random effects model.

Papers in specialized and contributed sessions:

Specialized session: Bayesian Methods

8. *N. Balgobin, J.W. Choi* A Bayesian Benchmarking for Small Areas.
9. *G.E. Montanari, G. Ranalli* Small Area Estimation when the Variable of Interest is Latent.
10. *M. Ghosh* Hierarchical Bayes Estimation For Bivariate Binary Data with Applications to Small Area Estimation.

Specialized session: Design and Weighting Issues

11. *R. Lehtonen, M. Myrskylä, C.E. Särndal, A. Veijanen* Estimation for Domains and Small Areas under Unequal Probability Sampling.
12. *P.D. Falorsi, P. Righi, C. Casciano* Multi-way Stratification Designs for Business Small Area Estimation.
13. *J. Kordos, J. Paradysz* New Development in Small Area Estimation Research in Poland.

Contributed session: SAE in Official Statistics

14. *L.U. Martinez, A.F. Militino, C. Prado, M. Ayestaràn, J.S. Vicente* Small area estimation in official statistics: The labour force survey of the Basque country.
15. *J. Wooton, D. Elazar, M. McEwin* Small Area Estimation: Issue Facing a National Statistical Office.
16. *P. Clarke, K. McGrath, H. Chandra, N. Tzavidis* Developments in Small Area Estimation in UK with focus on current research activities.
17. *M. Smeets* Model diagnostics and model selection in small area estimation - a case study in official statistics.

Contributed session: Semi-Parametric and Outlier Robust Methods

18. *W.R. Bell, E.T. Huang* Using the t-distribution to Deal with Outliers in Small Area Estimation.
19. *S. Casanova, Y. Aragon* Using conditional M-quantiles and quantiles to estimate a cumulative distribution function in a domain.
20. *J.L. Wywiał* Estimation of mean in domain when distribution of variable under study is skew.
21. *C.R. Sabater, F.G. Garcìa, J.A. Menèndez* Small Area Restricted Models.
22. *M.J. Lombardìa, W. Gonzàlez-Manteiga, I.Molina, D. Morales, L. Santamarìa* Small Area Estimation under Fay-Herriot Models with Nonparametric Estimation of Heteroscedasticity.
23. *D. Culliford* A Cross-Validatory Diagnostic Test under an Area-Level Model.

24. *T. Żądło* On prediction of domain total when population elements belong to domains at random.

Contributed session: Bayesian Approaches to SAE

25. *C. Pasquale* A New Bayesian Approach to Small Area Statistics and Resampling via Urns.
26. *B.L. Gunnels, T.J. Thompson, J.P. Boyle* Bayesian Hierarchical Methods for Small Area Estimation: A Case Study in Estimating Diabetes Prevalence by United States State and County.
27. *R.M. Baskin, J. Sommers* Comparison of direct state estimates and Bayesian state estimates for the medical expenditure panel survey.
28. *V.G. Rubio, N. Best, S. Richardson, P. Clarke* Bayesian models for Small Area Estimation and policy making.
29. *V.G. Rubio, N. Best, S. Richardson* A comparison of likelihood-based and Bayesian methods for Small Area Estimation.

Contributed session: Design Issues for SAE

30. *W. Niemiro, J. Wesołowski* Linear estimation and prediction under model - design approach with small area effects.
31. *K. Pruska* Empirical verification of usefulness of measures of small area similarity for synthetic estimation.
32. *P.A.V.B. Swamy, T.S. Zimmerman, J.S. Mehta* Two-Step versus Simultaneous Estimation of Survey- Non-Sampling Error and True Value Components of Small Area Sample Estimators.
33. *W. Gamrot* On Some Direct Estimator for the Domain Coefficient of Variation under Nonresponse.
34. *T. Jurkiewicz* Factors affecting efficiency of the modified synthetic estimation. A Monte Carlo analysis.

Contributed session: Labour Force Applications

35. *B. Meindl* Estimating unemployment-rates for small areas - A simulation-based approach.
36. *B. Buelens* The development of a tool for model-based small area estimation and its application to the Dutch Labour Force Survey.
37. *F. Hernàndez, M. Herrador, D. Morales, M.D. Esteban, A. Pérez* Sampling design variance estimation of small area estimators in the Spanish labour force survey.
38. *S. Haslett, A. Noble, F. Zabala* Small area estimation of unemployment using hierarchical Bayes methods.
39. *H.J. Boonstra* A comparison of several design-based and model-based estimators for municipal unemployment rates.

Contributed session: Business and Household Surveys Applications

40. *E. Gołata* An Attempt to Estimate Household Composition in Poland (based on Household Budget Survey data).
41. *T. Klimanek* The Indirect Estimation of Some Farm´s Characteristics - an Attempt.
42. *D. Hedlin, M. Jansson* Use of auxiliary data to decompose enterprise-level data into local unit-level data.

Contributed session: Demographic Applications

43. *L.N. Pereira* Small Area Estimation of the Mean Price of the Habitation Transaction in Portugal: Methodological and Practical Issues.
44. *T. Waldhoer, H. Heinzl, M. Wald,* Associations between the Spatial Distributions of Infant Mortality by Subgroups in Austria 1984-2005.
45. *H. Heinzl, T. Waldhoer* Analysing Austrian infant mortality data with survival time methods.
46. *D. Swanson, J. McKibben, R. Prevost* New Directions in the Development of Population Estimates and Projections.

Contributed session: SAE for Poverty and Expenditure

47. *G. Jones, S. Haslett, J. Enright* Problems in small-domain estimation of expenditure patterns.
48. *S. Haslett, M. Isidro, G. Jones* Comparison of Regression Techniques for Small Area Estimation of Poverty.
49. *D.C. Broadstock, A. Druckman, L.C. Hunt, T. Jackson* A comparison of small area estimation techniques to estimate household expenditure at a local area level in the UK.
50. *C. Quintano, R. Castellano, G. Punzo* Estimating Supplementary Poverty in the Italian Provinces: A Multidimensional and Fuzzy Analysis through Small Area Models.

Contributed session: Estimation of Uncertainty

51. *Molina* Uncertainty under a multivariate nested-error regression model with logarithmic transformation.
52. *M. Myrskylä* Variance estimation for the logistic generalized regression estimator under equal and unequal probability sampling.
53. *S. Chatterjee, P. Lahiri* High order accurate mean squared prediction error estimation in general small area models.

Specialized session: Disease Mapping

54. *L.U. Martinez, T. Goicoa, A.F. Militino* Alternative MSE Estimators in Disease Mapping.

55. *Saei, M. Murry, A. Charlett, A. Pearson* Small Area Estimation of Mandatory Surveillance for Methicillin Resistant Staphylococcus aureus (MRSA).
56. *Biggeri, D. Catelan, E. Dreassi, G. Cringoli* Multivariate Spatially Structured Variability of Ovine Parasitic Infections.

Specialized session: Small Area Models 1

57. *H. Chandra, R. Chambers* Small Area Estimation for Skewed Data.
58. *E. Fabrizi, C. Trivisano* Assessing Robustness of Normal Random Effects Models in Small Area Estimation.
59. *M. D'Alo', S. Falorsi, F. Solari* Linear Mixed Models for Generalised Random Effects Structures for Small Area Estimation.

Specialized session: MSE Estimation

60. *M. Peralta, N. Salvati, M. Pratesi* Bootstrap Mean Squared Error of the Spatial EBLUP.
61. *L.C. Zhang* Conditional Mean Squared Error of Prediction for Small Area Composition Estimates Subjected of Informative Missing Data.
62. *S. Chen, P. Lahiri, J.N.K. Rao* Robust Estimation of the Mean Squared Prediction Error of an EBLUP of a Small Area Mean.

Specialized session: Small Area Models 2

63. *R. Chambers, M. Pratesi, N. Salvati, N. Tzavidis* M-quantile Geographically Weighted Models with Application to Small Area Estimation.
64. *J. Opsomer* Nonparametric Small Area Estimation Using Penalized Spline Regression.
65. *M. Trevisani, N. Torelli* Small area models for count data: Alternative hierarchical Bayesian specifications.
66. *M.R. Ferrante* Multivariate Small Area Models for Count Data.

Specialized session: SAE for Poverty and Labour Force Applications

67. *R. Benedetti, C. Rinaldelli* Local Stationarity in EBLUP Estimation of Poverty Parameters.
68. *G. Betti, L. Neri* Comparing M-quantile and ELL methods for poverty mapping.
69. *G. Demombynes, C. Elbers, P. Lanjouw, J.O. Lanjouw* How good a map? Putting small area estimation to the test.
70. *S. Rubin-Bleuer, S. Godbout, Y. Morin* Evaluation of Small Domain Estimators for the Survey of Employment, Payrolls and Hours in Canada.

Specialized session: Applications in Agricultural Surveys

71. *Bernetti, L. Casini, N. Marinelli* Spatially Explicity Small Scale Modelling of Land Use Change: an Application for the Analysis of CAP Reform Scenarios.
72. *C. Bocci, A. Giommi, A. Petrucci, E. Rocco* Small Area Estimation in Presence of Non Random Non Response. An Application to the Italian Farm Structure Survey.
73. *R. Salvatore, C. Russo* Clustered Data in Small Area Estimation: an Analysis of Rural Local Economies.

Abstracts of the SAE2007 Conference papers and the invited presentations are available at www.dipstat.ec.unipi.it/SAE2007 . The Editorial Board of *Survey Methodology* has agreed to dedicate a substantial part of one of their 2008 issues for publication of selected papers presented at the conference.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 611—616

Report from the National Scientific Conference "Statistics Yesterday, Today and Tomorrow"

The National Scientific Conference "Statistics Yesterday, Today and Tomorrow" took place in October 10—12, 2007 in Wrocław.

The aim of the conference was to ***celebrate 95th anniversary of the Polish Statistical Association (PTS).*** The conference was organized by Wrocław branch of PTS with the cooperation of the Department of Statistics of Wrocław University of Economics, Statistical Office in Wrocław and Main Council of PTS. This meeting was also called the First National Meeting of Statisticians, since it gathered all the groups of Polish statisticians on the first scientific meeting after the Second World War.

Following the presentation of the history of the Polish Statistical Association (PTS) given by Walenty Ostasiewicz *Over-long Foreword*, we gave a brief summary of the main turning points in the development and work of PTS.

For the first time the organization called PTS was founded in 1912 in Cracow. The founding initiator was K. Kumaniecki, who became the secretary of the Society. The first president of PTS was Juliusz Leo, a professor at Jagiellonian University. One of the first activities of the Association was the publishing of the statistical yearbook "*Statystyka Polska*" in 1915..

One of important events, which stimulated the development of Polish statistics, was the foundation of the Central Statistical Office (GUS) in 1918. Thanks to this institution, the most important statistical publications were published. In this context it is worth mentioning "*Kwartalnik Statystyczny*" (1924—1934) and "*Wiadomości Statystyczne*".

In 1917 the Society of the Polish Economists and Statisticians (TEiSP) was founded in Warsaw. In 1933 TEiSP created the Section of Statistics, which in 1937, mainly thanks to Z. Limanowski and E. Szturm de Sztrem, turned to an independent organization called PTS. Professor Edward Szturm de Sztrem (1885—1962) became the president of the Society. One of the first decisions taken by the committee was the foundation of the journal "*Przegląd Statystyczny*".

Four sections worked within the Association:

- Section of Mathematical Statistics,
- Section of Population Statistics,
- Section of Administration and Social Statistics,
- Section of Business Statistics.

There were two meetings of PTS in 1938 (April 10, 1938 in Warsaw and April 18, 1938 in Cracow). One of the activities of the Society was the organization of lectures related to different aspects of statistics and its applications. In particular, on April 11, 1938 J. Neyman gave a lecture on statistical estimation and on May 4, 1938 M. Presburger gave a lecture on applications of statistical demography in actuarial problems. The last meeting before the Second World War was on 2 April 1939.

In 1947 S. Szulc reactivated the Association. The first meeting after the war was in June 1947. The Association decided to choose Szulc for the president and to restore the publication of "*Przegląd Statystyczny*".

The next meeting, run by J. Czekanowski, took place in May 1949. Among others M. Fisz, S. Szulc, H. Steinhaus, E. Szturm de Sztrem participated in this meeting. The participants criticized the former work of the Society. M. Fisz proposed a program of the intensification of the activity of PTS. Unfortunately, in 1951 the Association virtually stopped its work, and in 1955 it was formally closed down.

After this break, the next founding meeting of PTS, initiated by Polish statisticians, took place in 1981. ***Prof. Mikołaj Latuch*** became the President of the Association from 1981—1985. Next President of the Polish Statistical Association for two terms: 1985—1989 and 1989— 1994 was ***Prof. Jan Kordos***. In 1993 PTS started publishing journal in English ***"Statistics in Transition"*** with J. Kordos as an Editor-in-Chief. In 1990 "*Wiadomości Statystyczne*" (earlier published by GUS) became a journal published jointly with GUS and PTS. ***Prof. Czesław Domański*** was President of PTS for two terms: 1994—2000 and 2000—2004. In 2004 ***Dr Kazimierz Kruszka*** was elected a President of PTS.

The more detailed description of the history of PTS can be found in the paper "*Activities of the Polish Statistical Association*" ["Statistics in Transition", Vol. 1, No 1, June 1993] written by Prof. Jan Kordos.

The conference was organized under the honorary auspices of *Professor Józef Oleński*, the President of the Central Statistical Office of Poland. Many persons engaged in the development of statistical science in Poland took part in the honour committee: *Prof. Zbigniew Czerwiński, Prof. Zdzisław Hellwig, Prof. Jan Kordos, Prof. Wiesław Sadowski, Prof. Władysław Welfe, Prof. Kazimierz Zając and Prof. Ryszard Zieliński,* who gave their strong support for the idea of joining statisticians from different disciplines and institutions.

The conference drew approximately 110 scientists, representing official statistics (statistical offices) and academic statistics (mainly economic universities and economic faculties).

The conference was opened by *Prof. Bogusław Fiedor*, Rector of Wrocław University of Economics. Then the chairman of the Organizing Committee *Prof. Walenty Ostasiewicz* introduced the keynote participants of the conference. The next speaker was *Dr. Kazimierz Kruszka*, the President of PTS, who presented the current activity of the Association and its plans for the future. The former

president of PTS, *Prof. Czesław Domański*, talked briefly on the tumultuous history of the Association and people involved in the foundation of PTS. Additionally, the conference was accompanied with the exhibition of unique documents presenting the history and achievements of the Association, prepared by Jan Berger and Kazimierz Latuch from Central Statistical Office of Poland.

Almost 40 scientists gave their presentations at the conference. Below we summarize all the plenary talks and the contributions that got the highest response.

Prof. Danuta Strahl (Wrocław University of Economics) presented the problem of the measurements of innovation of European Union in her paper *Reference Border System in Innovation Measurement of European Regional Space*. This talk was the first one representing the economic statistics, in particular multidimensional comparison analysis.

Prof. Mirosław Krzyśko (Adam Mickiewicz University in Poznań) presented the paper *Polish Statistics — Present Status and Possibilities of Development* on the achievements of Polish statisticians with respect to a number of publications in statistical journals from the so called "*Philadelphian List*". Additionally he pointed the current state of the main statistical journals in Poland, mentioning also "*Statistics in Transition – new series*", journal of the Polish Statistical Association..

Prof. Tomasz Panek (Warsaw School of Economics**)** presented the paper *Poverty and Social Exclusion in Poland by Region (Voivodship)* on the poverty and social exclusions with respect to regions.

The paper *New Measures of Economic Polarization* by *Prof. Stanisław Maciej Kot* (Gdańsk University of Technology) dealt with new measures of economic polarization. It was mainly theoretically oriented, although this field of research has plenty of applications, for example in the research of the difference in the wealth of the population.

Prof. Tadeusz Bednarski (University of Wroclaw) — representing mathematical statistics — presented the paper *Von Mises Methodology and Statistical Inference for Time Series.* The paper discussed some problems concerning asymptotic marginal distribution of some statistics for time series and focused especially on the problem of unitary root.

The problems related to the national population and housing census obtained a great response, in particular talks by Prof. Jan Paradysz (Systematic Errors in Population Censuses) and Prof. Elżbieta Gołata (Population Census — Possibilities and Limitations), both authors from Poznań University of Economics.

Also the paper *Harmonization of Statistical Standards and Development of Statistical Surveys*, which was written by Dr Teresa Śmiłowska (Statistical Office in Łodź) on new statistical classifications to be introduced next year, met with great interest.

Prof. Mirosław Szreder (University of Gdańsk) presented the paper *On Some Aspects of Statistical Education in the Polish Society* and introduced the plenary discussion on the statistical education of Polish society. He talked about the problems of surveys, in particular about the support of voters for political parties before the voting. Next in the discussion many persons took part who gave their suggestions.

It is particularly worth talking over the first plenary presentation *of Prof. Józef Oleński*, the President of GUS, who encouraged all the statisticians to support official statistics and promote statistical culture in the whole society. He proposed to organize annual Statistical Weeks (as it is organized by German Statistical Society). Additionally, PTS should write an ethics code for statisticians and found an independent organization evaluating the quality of statistical information published by the media. All the cases of manipulations should be made widely known through the publication in the Internet. Besides, PTS should promote statistical education at higher education level, especially with respect to the designing and organization of statistical experiments (as it is already realized at sociology departments) and knowledge about ethics in statistics. Finally, he postulated to increase the basis of PTS, in order to incorporate not only scientists from economic faculties of universities and statistical offices, but also statisticians working on health and environment protection. He also proposed to found new sections, as e.g. biostatistics or statistics in technique.

Besides the above mentioned talks, the following papers were presented:

- Iwona Bąk, Katarzyna Wawrzyniak (University of Agriculture in Szczecin), *The Multidimensional Scaling as an Instrument of Segmentation of the Tourist Market in Subregions of Poland*;
- Jan Berger (Central Statistical Office), *Foundation of Municipal Statistical Office in Wrocław in 1873*;
- Marek Cierpiał-Wolan (Statistical Office in Rzeszów), *Economic Aspects of External Migration in Poland*;
- Izabela Cichocka (University of Information Technology and Management in Rzeszów), *Application of Methods of Statistical Data Analysis of Dynamics to Describing Changes in Passenger Cars Market in Poland*;
- Grażyna Dehnel (Poznań University of Economics), *GREG Estimator and Winsor-type Estimation in Micro-enterprises Survey*;
- Joanna Dębicka (Wrocław University of Economics), *An Approach to the Study of Future Cash Flows Arising from Multistate Insurance Contracts*;
- Magdalena Dyda (Wrocław University of Economics), *Importance Sampling*;
- Ewa Frątczak (Warsaw School of Economics), *Event History Analysis and Multilevel Analysis Unit — Teaching and Research*;

- Grzegorz Kończak (University of Economics in Katowice), *Interactive Tools for Teaching Statistics — Internet Resources and Own Development*;
- Jan Kordos (Formerly Central Statistical Office / Warsaw School of Economics), *Interplay Between Sample Survey Theory and Practice in Poland*;
- Cyprian Kozyra (Wrocław University of Economics), *Evaluation of Health Care Quality*;
- Kazimierz Latuch (Central Statistical Office), *The 25th Anniversary of the Reactivation of Polish Statistical Association (1981—1982)*;
- Bożena Łazowska (Central Statistical Library), *Centralna Biblioteka Statystyczna (CBS) — The Central Statistical Library Yesterday, Today and Tomorrow*;
- Markowicz Iwona, Stolorz Beata (the University of Szczecin), *Use of Survival Analysis Methods for Study of Socio-Economic Phenomena*;
- Małgorzata Markowska (Wrocław University of Economics), *Regional Innovation Measurement Problems in EU Statistics. Changes in Approaches*;
- Edyta Mazurek (Wrocław University of Economics), *Assessment of Tax System with Tax Allowance for Children in 2007 on the Basis of Dagum Decomposition of Gini Coefficient*;
- Aldona Migała-Warchoł (University of Information Technology and Management in Rzeszów), *Quality of Life Geographical Diversification in Podkarpackie Voivodeship*;
- Andrzej Młodak (Statistical Office in Poznań), *Methods of Complex Assessment of Properties of Composite Observations*;
- Anna Nikodem (Wrocław University of Economics), *Estimation of Ruin Probability*;
- Agnieszka Ordon (University of Information Technology and Management in Rzeszów) *"Demographic" Future of Universities — Forecast of the Population Undertaking Studies, 2007—2030*;
- Barbara Podolec, Paweł Ulman, Agnieszka Wałęga (Cracow University of Economics), *Economic Activity and Financial Situation of Households. Report from the Survey*;
- Bogdan Reiter (Statistical Office in Poznań), *Monitoring the Quality of Life in Cities and Possibilities of Information Coverage in Polish Public Statistics*;
- Agnieszka Sompolska-Rzechuła (University of Agriculture in Szczecin), *Quality of the Classification of Poland's Voivodships by the Natural Environment Pollution State*;

- Elżbieta Stańczyk (Statistical Office in Wrocław), *Competitiveness of the Dolnośląskie Voivodship against Other Voivodships between 1995 and 2005*;
- Piotr Tarka (Poznań University of Economics), *Cluster Analysis in the Segmentation of Market Behaviour of the Recording Companies*;
- Paulina Ucieklak-Jeż (Jan Długosz University of Częstochowa), *Application of Sullivan's Method to the Comparison of Disability-Free Life Expectancy for Men and Women*;
- Janusz Wątroba (Statsoft Polska), *Aid for Statistical Data Analysis in STATISTICA 8*;
- Jan Zawadzki (University of Agriculture in Szczecin), *On the Research into the Methods of Forecasting Missing Data in Economic Time Series*;
- Beata Zmyślona (Wrocław University of Economics), *Application of IRT Models to Responses Randomization*.
- Dorota Bartosińska, Anna Jankiewicz-Siwek (Maria Curie-Skłodowska University in Lublin), *Statistical Analysis of Economic Activity of the Population in the Lublin Region and Poland*

Additionally we want to note the active participation of *Statsoft Polska* — one of the main sponsors of the conference.

Prepared by Cyprian Kozyra and Joanna Dębicka,
Department of Statistics, Wrocław University of Economics.

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 617—625

XXVI Conference on Multivariate Statistical Analysis (MSA 2007) & VI Conference on Statistics in Social and Economic Practice (SwPSG 2007)

Łódź, Poland, 5—7 November 2007

The XXVI Conference on **Multivariate Statistical Analysis** and the VI Conference on **Statistics in Social and Economic Practice** were held simultaneously from 5th to 7th November 2007 in Łódź. Organization of the conferences was charged to the Chair of Statistical Methods, University of Łódź.

The conference on Multivariate Statistical Analysis (MSA) has been organized, in Łódź, every year, since 1981. The goal has been focused on the latest achievements in the field of the multivariate statistical analysis and its applications.

The conference on Statistics in Social and Economic Practice has been organized every year, since 2001, in one of the university centers in Poland. Its overall goal was to bring together people representing different scientific and research centers in the field of applying statistics in social and economic practice.

Altogether there were 112 participants from various academic and research centers in Poland. Concerning the papers , 67 papers were presented in 18 sessions.

The conferences were opened by the Chairman of the Organizing Committee: **Prof. Czesław Domański**. The opening speech was given also by the Rector of the University of Łódź **Prof. Wiesław Puś** and the Dean of the Faculty of Economics and Sociology of the University of Łódź **Prof. Paweł Starosta**.

The scientific programme of the conferences was very broad and covered most of the statistical problems, such as multivariate distributions, multivariate statistical tests, nonparametric inference, factor analysis, cluster analysis, Bayesian inference, multivariate Monte Carlo analysis, data mining, robust procedures, pattern recognition and applications of multivariate methods in marketing, finance, insurance, capital markets, risk management, medicine and health services.

The first plenary session (chair: *Prof. Bronisław Ceranka*) was devoted to three famous Polish statisticians:

- Hugo Steinhaus (1887—1972)
- Jan Czekanowski (1882—1985)
- Fryderyk Moszyński (1737—1817)

In 2007 we celebrated 120th birth anniversary of **Professor Hugo Steinhaus**, ***a famous Polish mathematician and statistician***. Some of his achievement were reminded and impact on contemporary statistics discussed.

In 1923 he published in *Fundamenta Mathematicae* the first rigorous account of the theory of tossing coins based on the measure theory. In 1925 he was the first to define and discuss the concept of strategy in game theory.. In 1929, together with Banach, he started a new journal *Studia Mathematica* and Steinhaus and Banach became the first editors.

Various authors stress in their works the precursory role played by Steinhaus. For example, Norbert Wiener states that he was inspired by Steinhaus' ideas when he was developing the theory of stochastic processes.

From his early years Steinhaus, engaged in the *popularization of mathematics*. He delivered popular lectures, wrote articles to various periodicals, and published "*What Mathematics is and What is Not"* (1923 in Polish). Later, he wrote the "*Kalejdoskop Matematyczny*", which appeared also in English under the title "*Mathematical Snapshots*" (1938). A considerable place in the latter publications is occupied by the problems of chess, draughts and other games. The earlier article on games (1929), defining the fundamental notion of a determined game, caused Steinhaus to be seen as a pioneer of the rapidly developing *theory of games.*

Steinhaus was always interested in various *practical problems of biology, medicine, geography* and sometimes *technology*. His works on *applications of mathematics* to those and also other fields began to appear in 1924. There are practical problems at stake as regards measuring areas and lengths and later of what is known as Groer's law of patergy concerning infant tuberculosis (1934). Steinhaus designed an instrument for localization of strange bodies in the body of a sick person by means of X-rays, based on a simple and elegant geometrical conception (1938). Steinhaus' bibliography contains 170 articles. He did important work on functional analysis and application of mathematical and statistical methods in different fields (but he himself described his greatest discovery in this area as Stefan Banach).

Prof. Jan Kordos (Formerly Warsaw School of Economics and Central Statistical Office), a student of Prof. H. Steinhaus in 1953—1955 at Wrocław University, attended Prof. Steinhaus lectures on *probability theory*, *statistical quality control* and *seminars.* He prepared his master thesis under Steinhaus' supervision. In his paper entitled *"**Over 50 years in Statistical Practice**"* he has presented some statistical problems he had met during his over 50 years of work in the country and abroad. Among others, Steinhaus' lectures on *statistical quality*

control had impact on his interest in *quality of statistical data*. He used *Steinhaus' Tables of Iron Numbers* in *implementing sample surveys.*

The next speaker **Prof. Mirosław Krzyśko** (Poznań) presented the lecture ***"Jan Czekanowski — the anthropologist and statistician"***. The lecture was concerned with the life and researches of a great, Polish statistician — Jan Czekanowski (1882—1965). He attended school in Warsaw but was transferred to Latvia, where he finished his education in 1901. Then he entered a university in Zurich in 1902; there, he studied anthropology, mathematics, anatomy and ethnography as a student of Swiss anthropologist Rudolph Martin. In 1907 Czekanowski defended his PhD. dissertation. For the dissertation's research he traveled to the Royal Museum in Berlin and to Middle Africa from 1906—1907. While in Africa, he led a team into the Congo to collect ethnographic materials. While working on studying the societies of Africa, he developed various statistical methods and contributed to the field of taxonomy. The research he made in Africa has since been published in five volumes and sent in 1910 to Saint Petersburg ethnography. He then became a professor at the University of Lviv and University of Poznan. While working he introduced an innovative take on mathematical statistics. He played numerous scientific roles at the UAM, including vice-chairman of the Polish Social Statistic Company. He created so called *Czekanowski diagram.*

Prof. Czesław Domański (Łódź) reported a paper titled ***"Fryderyk Moszyński –the first Polish statistician"***. The statistical activity of Fryderyk Moszyński (1737—1817) was initiated at the times of the so-called Great Sejm of 1788—1791 (the longest Parliamentary Session in the Polish history). On Moszyński's initiative the Sejm decided in its resolutions of June 22, 1789 to carry out the first-ever general census combined with smoke-counting. He also proposed a statistical method of measuring the army tax.

Moszyński was the author of constant record of natural population movement conducted in cooperation with the church authorities.Working on his own he collected a very valuable statistical material related to the Treasury incomes and expenses.

The statistical material collected by Moszyński was received very well both in Poland and abroad and described as "highly reliable" while the scholar himself was acclaimed as "an eminent statesman". His first attempt to estimate the population of Poland, which was partitioned at that time, with the use of "presumed calculation" (1 smoke—6 souls) is also worth mentioning.

Titles of the papers of the next sessions of the conferences MSA and SwPSG, with the names of the authors, are presented below:

5 November 2007:

Session II MSA:

Chair: ***Prof. Grażyna Trzpiot***

- *A comparison of nonparametric actual error rateestimators in classification problems* (Mirosław Krzyśko, Michał Skorzybut, Poznań),
- *On some estimator of finite population skewness under nonresponse* (Wojciech Gamrot, Katowice),
- *On predictor of domain total under some model which belongs to the class of general linear mixed models* (Tomasz Żądło, Katowice),
- *On an improvement of the model-based clustering method* (Ewa Witek, Katowice).

Session II SwPSG:

Chair: ***Prof. Janusz Wywiał***

- *The indirect estimation of equalized income for rural households-Households Budget Surveys 1999—2005* (Tomasz Klimanek, Poznań),
- *ARIMA modelling in research on the relationship between the biological condition of human populations and the level of economic stress* (Elżbieta Żądzińska, Czesław Domański, Iwona Rosset, Artur Mikulec, Łódź),
- *Statistical analysis of social and demographic factors forming expected wage in Poland in 2005 (*Marcin Salamaga, Kraków).

Session III MSA:

Chair: ***Prof. Wiesław Wagner***

- *Perspectives on development bayesian models in technical reserving used in insurance companies* (Marek Karwański, Piotr Jałowiecki, Arkadiusz Orłowski, Warszawa),
- *Robustness of depth based classification rules* (Daniel Kosiorowski, Kraków),
- *Application of clustering methods to classification of income distributions* (Piotr Łukasiewicz, Grzegorz Koszela, Arkadiusz Orłowski , Warszawa),
- *Unbiased recursive partitioning algorithm in regression trees* (Dorota Rozmus, Katowice).

Session III SwPSG:

Chair: ***Prof. Elżbieta Gołata***

- *Decision trees in statistical analysis — a virtual supply chain example* (Grażyna Trzpiot, Alicja Ganczarek, Katowice),
- *Classification methods in data analysis of virtual supply chains* (Grażyna Trzpiot, Alicja Ganczarek, Katowice),
- *The measurement of respondents sensibility and finding the stressors with the largest strength* (Tomasz Żądło, Katowice).

Session IVa MSA:

Chair: ***Prof. Mirosław Krzyśko***

- *Effectiveness of symbolic classification trees vs. noisy variables* (Andrzej Dudek, Marcin Pełka, Wrocław),
- *Classification of patients with respect to some group of factors* (Ewa Nowakowska-Zajdel, Grażyna Trzpiot, Alicja Ganczarek, Małgorzata Muc-Wierzgoń, Bytom, Katowice),
- *Choice of covariates in probability models based on roc and cap curves* (Iwona Schab, Warszawa),
- *On the application of distance-based algorithms in medical research* (Małgorzata Misztal, Maciej Banach, Łódź).

Session IVb MSA:

Chair: ***Prof. Krzysztof Jajuga***

- *Almost stochastic dominance in multicriteria decision aiding* (Maciej Nowak, Katowice),
- *The power of some tests of multivariate normality based on the moments* (Czesław Domański, Izabela Wojek, Łódź),
- *On the modification of the empty cells test* (Grzegorz Kończak, Katowice),
- *Variance balanced block designs with repeated blocks* (Bronisław Ceranka, Małgorzata Graczyk, Poznań).

6 November 2007:

Plenary Session I:

Chair: ***Prof. Jan Paradysz***

- *Fertility changes in regional aspect, tempo and quantum effect* (Elżbieta Gołata, Poznań),
- *Some aspects of post-enumeration surveys in Poland* (Jan Kordos, Warszawa),
- *About the estimation of a mean of the random variable with asymmetric distribution* (Janusz Wywiał, Katowice).

Plenary Session II:

Chair: ***Prof. Jan Kordos***

- *Polish population census 2011 as a challenge for small area statistics* (Jan Paradysz, Poznań),
- *Extreme value distribution and robust estimation* (Grażyna Trzpiot, Katowice),
- *Proofs of the normalization the density function of the class one-dimensional normal distributions* (Wiesław Wagner, Dariusz Parys, Lechosław Stępień, Rzeszów, Łódź).

Session III MSA:

Chair: ***Prof. Walenty Ostasiewicz***

- *Advantages and disadvantages of application for conjoint analysis* (Marcin Hundert, Szczecin),
- *Determinants of tourist attractiveness of communities* (Rafał Czyżycki, Szczecin),
- *Graphical presentation of a multi-way contingency table in the R software* (Iwona Kasprzyk, Katowice)
- *Multivariate statistical analysis of tourist attraction of provinces in Poland* (Rafał Klóska, Szczecin).

Session III SwPSG:

Chair: ***Prof. Eugeniusz Gatnar***

- *On internet sample survey* (Janusz Wywiał, Katowice),
- *A new approach to small business research* (Grażyna Dehnel, Poznań),

- *About one more method of estimating the return on sale* (Aleksandra Witkowska, Marek Witkowski, Poznań),
- *The application of Sullivan's method in calculating Healthy Life Expectancy in Poland in 2004"* (Paulina Ucieklak-Jeż, Częstochowa).

Session IVa MSA:

Chair: ***Prof. Janusz Wywiał***

- *Interpretative aspects of the matrix of coefficients of correlation of lineal and median* (Andrzej Mantaj, Rzeszów),
- *Using the Var model in time-structure analysis of interest rates in Poland* (Jerzy Rembeza, Grzegorz Przekota, Koszalin),
- *The minority game and quantum game theory* (Katarzyna Bolonek, Łódź).

Session IVb MSA:

Chair: ***Prof. Czesław Domański***

- *Efficiency of the modified synthetic estimator — Monte Carlo analysis* (Tomasz Jurkiewicz, Gdańsk),
- *The survey of economic activity of people in rural areas — the analysis using the econometric hazard models* (Joanna Landmesser, Warszawa)
- *Self-organizing neural networks in cluster analysis* (Kamila Migdał-Najman, Krzysztof Najman, Gdańsk),
- *Modifications of agglomerative clustering methods* (Jerzy Korzeniewski, Łódź).

7 November 2007:

Session I MSA:

Chair: ***Prof. Wiesław Wagner***

- *Statistical analysis of values of units and rates of return of Open Pension Funds* (Jacek Białek, Artur Mikulec, Łódź),
- *Abbreviated social welfare function proposed by Sen and its application to the analysis of household budgets in Poland* (Alina Jędrzejczak, Łódź),
- *Trends in cigarettes consumption in Poland — a comparison of adaptative and ARIMA models* (Ewa Jałowiecka, Piotr Jałowiecki, Arkadiusz Orłowski, Warszawa),
- *Application of logistic regression for firms survival analysis* (Iwona Markowicz, Beata Stolorz, Szczecin).

Session I SwPSG:

Chair: ***Prof. Jerzy Rembeza***

- *An attempt to use calibration estimators in household budget surveys in Poland* (Marcin Szymkowiak, Poznań),
- *Application of location depth and regression depth in analysis of level of environment pollution in Poland* (Dorota Pruska, Łódź),
- *Regression depth based analysis of clustered data* (Daniel Kosiorowski, Kraków),
- *On the use of weighting adjustments to estimate the finite population covariance under nonresponse* (Wojciech Gamrot, Katowice).

Session II MSA:

Chair: ***Prof. Tadeusz Gerstenkorn***

- *Models probabilistic for random variables about Bernoulli distribution* (Andrzej Mantaj, Wiesław Wagner, Rzeszów),
- *Multidimensional scaling for symbolic interval data* (Andrzej Dudek, Wrocław),
- *On some properties of support vector clustering* (Michał Trzęsiok, Katowice),
- *Optimum chemical balance weighing designs for p=v+1 objects based on balanced block designs* (Małgorzata Graczyk, Poznań).

Session II SwPSG:

Chair: ***Prof. Czesław Domański***

- *Polish labour market situation modeling and forecasting in the voivodship perspective in years 1999—2007* (Katarzyna Frodyma, Kraków),
- *Determinants of job awaiting time and their interaction* (Iwona Markowicz, Beata Stolorz, Szczecin),
- *Application of cluster analysis in regression estimation for small areas (*Krystyna Pruska, Łódź),
- *Analysis of relationships between prices of agriculture products and prices of alimentary products* (Jerzy Rembeza, Joanna Giłka-Zaporska, Koszalin).

Session III MSA and SwPSG:

Chair: ***Prof. Krystyna Pruska***

- *Introduction to the problem of truncated power series distribution* (Tadeusz Gerstenkorn, Łódź),
- *Bayesian method in interval estimation of Sharpe style weights in the OPF style analysis model* (Agnieszka Orwat, Katowice),
- *Multivariate data classification — comparison of ISODATA and approximation of points methods* (Arkadiusz Maciuk, Wrocław),
- *Statistical analysis of economic activity of the population in Poland in comparison to the European Union (* Dorota Bartosińska, Anna Jankiewicz-Siwek, Lublin),
- *Statistical analysis of logistic processes in enterprise* (Czesław Domański, Łódź).

The next Conference on Multivariate Statistical Analysis will be held in November 3—5, 2008 in Łódź. Scientists interested in attending the Conference are kindly requested to send their application to the Scientific Secretary of MSA'2008 to the following address:

Katarzyna Bolonek
27th Conference MSA'2008
Chair of Statistical Methods, University of Łódź
90-214 Łódź, Rewolucji 1905 r. nr 41, Poland
phone: (4842) 635 5335; fax: (4842) 635 53 07
e-mail: msa@uni.lodz.pl

Prepared by Aleksandra Baszczyńska and Jacek Białek , University of Łódź

STATISTICS IN TRANSITION-new series, December 2007
Vol. 8, No. 3, pp. 627—628

ACKNOWLEDGEMENTS

Referees of Volume 8

The Editorial Board wishes to thank the following referees who have generously given their time and skills to the ***Statistics in Transition- new series*** *during the period from January 2005 to December 2007.*

Czesław Bracha, Warsaw School of Economics, Poland.
Ray Chambers, University of Southampton, UK
Kari Djerf, Statistics Finland, Finland
Czesław Domański, University of Łódź, Poland
Ewa Frątczak, Warsaw School of Economics, Poland.
Wojciech Gamrot, Academy of Economics, Katowice, Poland
Elżbieta Gołata, University of Economics, Poznań, Poland.
Dan Hedlin, Statistics Sweden, Sweden
Johan Heldal, Statistics Norway, Oslo, Norway
Krzysztof Jajuga, Wrocław University of Economics, Poland
Graham Kalton, WESTAT, Inc. Washington, USA
Jan Kordos, Formerly Warsaw School of Economics, Poland.
Danutê Krapavickaitê, Institute of Mathematics and Informatics, Lithuania
Irena Kotowska, Warsaw School of Economics, Poland
Marek Kozak, Warsaw Agricultural University, Poland
Liliana Kursa, Central Statistical Office, Poland
Risto Lehtonen, University of Helsinki, Finland
Marek Męczarski, Warsaw School of Economics, Poland.
Mikko Myrskylä, Statistics Finland, Finland
Wojciech Niemiro, Warsaw University, Poland
Lucyna Nowak, Central Statistical Office, Poland.
Tomasz Panek, Warsaw School of Economics, Poland.
Jan Paradysz, University of Economics, Poznań, Poland
Dorota Pekasiewicz, University of Łódź, Poland
Aleksandras Plikusas, Institute of Mathematics and Informatics, Lithuania
Richard Platek, Formerly Statistics Canada, Canada
Waldemar Popiński, Central Statistical Office, Poland
Krystyna Pruska, University of Łódź, Poland
J.N. K. Rao, Carleton University, Canada
Teresa Słaby, Warsaw School of Economics, Poland.

Adam Szulc, Warsaw School of Economics, Poland
Mirosław Szreder, University of Gdańsk, Poland
Daniel Thorburn, Stockholm University, Sweden
Imbi Traat, University of Tartu, Estonia
Ari Veijanen, Statistics Finland, Finland
M.R. Verma, Umiam (Barapani), Meghalaya, India,
Vijay Verma, Consultant in Survey Methodology, India
Jacek Wesołowski, Warsaw University of Technology, Poland
Janusz Wywiał, Academy of Economics, Katowice, Poland
Jan Wretman, Stockholm University, Sweden
Li Chun Zhang, Statistics Norway, Oslo, Norway
Agnieszka Zgierska, Central Statistical Office, Poland
Tomasz Żądło, Academy of Economics, Katowice, Poland