STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 1—4

FROM THE EDITOR

Starting from 2007 the Editorial Board of *the Statistics in Transition* decided to extend its field of interest to broad area of application of statistical methods. Though the priority is given to analysis of the experience of the post-communist economies and other emerging markets, articles concerned with general application of statistical methods and teaching statistics are also welcome. To emphasize these changes a new title of the journal is introduced *"Statistics in Transition – new series"*. This is first issue published in the new journal – *Statistics in Transition – new series.*

This issue contains seven articles on sampling and estimation methods, four other articles on application different statistical methods, and an obituary.

There are following seven articles devoted to some problems of sampling and estimation methods:

1. ***Cross-sectional and longitudinal weighting in a rotational household panel applications to EU-SILC*** (by Vijay Verma from India, Gianni Betti and Giulio Ghellini from Italy). This paper provides a comprehensive description of an integrated system of cross-sectional and longitudinal weighting for rotational household panel surveys. It develops the weighting procedures with reference to the EU-SILC *integrated design*. EU-SILC (Statistics on Income and Living Conditions) covers data and data sources of various types: cross-sectional and longitudinal; household and personal; economic and social; and from registers and interview surveys. It describes a step-by-step procedure for construction of *initial weights* to be applied to each new sample as it is introduced into the survey. Starting from the initial weight of each individual in the original sample, the person's base weight is constructed for each subsequent year to compensate for panel attrition. The final objective is to construct *cross-sectional weights* and *longitudinal weights* for use in data analysis. Procedures are described for constructing these two types of weights from the same base weights. This makes the whole system of weights internally consistent and integrated.
2. ***Improved ratio-cum-product type estimators*** (by Pier Francesco Perri from Italy). In this paper an improved version of Singh's *ratio-cum-product* estimators is suggested in simple random sampling when two auxiliary variables are available. Using Taylor linearization method up to the first and second order of approximation, expressions for the mean square error of the proposed estimators are established that they are more efficient than the original ratio-cum-product estimators. Moreover, in

order to better evaluate the performance of the estimators, an application to real data is shown.

3. ***On PSPNR sampling scheme for mean estimation*** (by D. Shukla and Jayant Dubey from India). This paper presents a post-stratified partial non-response (PSPNR) sampling scheme useful to cope with the presence of partial non-response in surveys using post-stratification. An unbiased estimation strategy is proposed and its optimum properties are examined. Cost aspect, under PSPNR, is explored and observed having limitations about the solution of equations. An alternative cost strategy is proposed to deal with which provides cost-optimal selection of required sample-size and optimum allocation of sample fractions. All derived properties and results are numerically supported.
4. ***A Family of estimators for estimating population mean using known correlation coefficient in two phase sampling*** (by Rajesh Singh, Pankaj Chauhan and Nirmala Sawan from India). A family of estimators for estimating the population mean of the variable under study, using two auxiliary variables in two-phase sampling, and known correlation coefficient of the second auxiliary character is proposed. It has been shown that the proposed estimators are more efficient than usual unbiased estimator, usual two-phase ratio estimator and Chand (1975) estimator. An empirical study is carried out to illustrate the performance of the constructed estimator.
5. ***Effect of non-response on current occasion in search of good rotation patterns on successive occasions*** (by G. N. Singh and Kumari Priyanka from India). The present work is an attempt to study the effect of non-response at current occasion in search of good rotation patterns on successive occasions. Two difference type estimators have been proposed for estimating the population mean at current occasion in presence of non-response in two occasions successive (rotation) sampling. Detailed behavior of proposed estimators has been studied. Proposed estimators have been compared with the estimators for the same situations but in the absence of non-response. Empirical studies have been carried out to demonstrate the performance of the proposed estimators.
6. ***On synthetic and composite estimators for small area estimation under Lahiri – Midzuno sampling scheme*** (by G.C. Tikkiwal and K. K. Pandey from India**).** This paper studies performance of synthetic ratio estimator and composite estimator, which is a weighted sum of direct and synthetic ratio estimators, under Lahiri – Midzuno (L-M) sampling scheme. The synthetic estimator under L-M scheme is unbiased and consistent if the assumption of synthetic estimator is satisfied. Further, this paper compares performance of the synthetic and composite estimators empirically under L-M and SRSWOR schemes for estimating crop acreage for small domains. The study shows that both the estimators perform better under

L-M scheme as having comparatively smaller absolute relative biases and relative standard errors.

7. ***Simulation analysis of accuracy estimation of population mean on the basis strategy dependent of sampling design proportionate to the order statistics of an auxiliary variable*** (by Janusz Wywiał from Poland). The paper deals with an analysis of the accuracy of the strategies for estimating the mean as well as the total value of a variable under study in a fixed and finite population. Positive valued auxiliary variable is involved. Three strategies called quantile types are compared with a simple sample mean, with an order ratio estimator from the simple sample mean as well as with the order ratio estimator from a sample drawn according to the sampling design proportional to the sample mean of an auxiliary variable. The proposed ratio type strategy is dependent on the sampling design proportional to the value of the order statistic of the auxiliary variable, as proposed by Wywiał (2006). The comparison of the strategies' accuracy has been based on a computer simulation. In the case of a small size population, the mean square errors have been evaluated on the basis of all possible samples which can be selected. In the case of a larger population, the samples have been drawn according to the considered sampling schemes. Finally, appropriate conclusions have been formulated.

There are following four articles devoted to application of statistical methods in different fields:

8. ***Extreme value treatment for samples from skew income distributions*** (by Marek Balog from Slovakia and Daniel Thorburn from Sweden). Income distributions, as well as many other economic variables, are often quite skew with a few very large values. When a sample is taken one may get too many or too few extreme observations. Estimates of totals and means, as well as inequality measures like the Gini coefficient, can be strongly influenced when ordinary design-based methods are used, if the outliers are weighted according to their inclusion probabilities. In this paper the authors use a model for the tail of the distribution. The parameters of the tail distribution are estimated and the estimated tail distribution is used for estimation of means and inequality measures. In this way the possibility of extreme values are taken into account even when there is no extreme value in the sample and the impact of the extreme values are decreased if there are too many of them in the tail. The methods are applied to the full income distribution in Sweden. It is shown that this method gives better results than both classical design-based and classical outlier methods.

9. ***Spatial division of labour: an empirical evidence from Poland*** (by Henryk Gurgul and Paweł Majdosz from Poland). The paper is examining spatial division of labour in Poland. Defining a region's ratio of actual employment to a predicted one and using a non-survey method to estimate the latter, the

authors provide evidence that some regions tend to specialize in jobs requiring low educated workers whereas other are more likely to specialize in the high-skilled jobs associated with a high educational level of employees. Furthermore, it turned out that there exists a statistically significant relationship between the rural/urban-type of a region and the direction of its specialization; while more rural regions specialize in low-skilled jobs, urban regions specialize in high-skilled jobs.

10. ***The gravity model – the study of Poland's trade integration with the European Union: trade creation and trade diversion effect*** (by Maria Majewska and Jolanta Grala-Michalak from Poland). The authors use a version of the gravity model to analyze the effect of both trade creation and trade diversion on Polish external trade with European Union (EU) for the years: 1990, 1995, 1997, 1999, 2001 and 2003. The results show that both trade creation and trade diversion for Poland bilateral trade with EU have positive and statistically significant coefficients, which means that Poland have traded with outer-region countries as well as with intraregion countries above the hypothetical level and Poland trade integration with EU has not caused a negative trade diversion effect. It is also observed that the trade creation effects in the Poland exporting and importing activities are proved to be generally stronger in the case of the old EU (15 countries) than in the case of bilateral trade with the new EU.
11. ***The Quality Policy in the Management of the Polish Labour Costs Survey*** **(**by Jolanta Szutkowska from Poland). This article presents the quality policy in the labour costs survey conducted every 4-years strictly according to Eurostat recommendations. This survey is well advanced in the quality policy because earnings and labour costs statistics was one of the first domains covered by quality requirements of European Statistical System. The quality policy in labour costs survey is illustrated on the basis of different quality approaches, namely: (i) the labour costs survey as the coherent and comparable part of the integrated system on earnings and labour costs statistics, (ii) quality reporting in labour costs survey, (iii) the application of DESAP as the self-assessment quality tool for the survey manager, (iv) the quality audit in labour costs survey.

The editor announces with the deepest sorrow that ***Professor Marek Balicki, the Head of the Department of Statistics, University of Gdansk, died at the age of 63.*** Professor Balicki was well known as a researcher and expert in the field of environmental statistics. He was an active member of the Polish Statistical Association and different scientific committees and councils of the Central Statistical office of Poland and other institutions.

Jan Kordos,
The Editor-in-Chief

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 5—50

CROSS-SECTIONAL AND LONGITUDINAL WEIGHTING IN A ROTATIONAL HOUSEHOLD PANEL: APPLICATIONS TO EU-SILC

Vijay Verma[1], Gianni Betti[1], Giulio Ghellini[1]

ABSTRACT

This paper provides a comprehensive description of an integrated system of cross-sectional and longitudinal weighting for rotational household panel surveys. To be concrete and detailed, it develops the weighting procedures with reference to the EU-SILC *integrated design*. EU-SILC (Statistics on Income and Living Conditions) covers data and data sources of various types: cross-sectional and longitudinal; household and personal; economic and social; and from registers and interview surveys. The standard integrated design involves a rotational panel in which a new panel is introduced each year to replace one quarter of the existing sample; persons enumerated in each new panel are followed-up in the survey for four years (Verma and Betti, 2006). A common rotational sample of this type yields each year a cross-sectional sample as well as longitudinal samples of various durations. These sample data have to be weighted to make them more representative of the target populations they represent. The paper begins with a summary of the main features of EU-SILC and an overview of the integrated weighting system for the different types of data coming out of the rotational panel annually. It describes a step-by-step procedure for construction of *initial weights* to be applied to each new sample as it is introduced into the survey. An innovative feature of the weighting procedure is the concept of *base weights* (Verma and Clemenceau, 1996). Starting from the initial weight of each individual in the original sample, the person's base weight is constructed for each subsequent year to compensate for panel attrition. The final objective is of course to construct *cross-sectional weights* and *longitudinal weights* for use in data analysis. Procedures are described for constructing these two types of weights from the same base weights. This makes the whole system of weights internally consistent and integrated.

Key words: weighting, household panel, rotational design, cross-sectional weights, longitudinal weights, EU-SILC.

[1] V. Verma, e-mail: verma@unisi.it (corresponding author); G. Betti e-mail: betti2@unisi.it, G. Ghellini, e-mail: ghellini@unisi.it. Department of Quantitative Methods, University of Siena, P.za S. Francesco, 7, 53100, Siena, Italy.

1. The EU-SILC framework

1.1. Introduction: context and content of the paper

The present paper provides a comprehensive description of an integrated system of cross-sectional and longitudinal weighting for rotational household panel surveys, specifically with reference to the EU-SILC integrated design.

As is well-known by this time, EU-SILC is the major new source of comparative statistics on income and living conditions in Member States of the European Union and some neighbouring countries. It has been developed as a flexible yet comparable instrument for the follow-up and monitoring of poverty and social exclusion at the EU and national levels. It covers data and data sources of various types: cross-sectional and longitudinal; household-level and person-level; economic and social; from registers and interview surveys; from new and existing national sources.

In previous papers and reports (Verma, 2001; Verma and Betti, 2006) we have elucidated the structure and main characteristics of EU-SILC surveys, and the various technical considerations involved in the design and implementation of samples for EU-SILC. Despite the diversity of arrangements permitted under EU-SILC, the standard integrated design recommended by Eurostat has been adopted by a big majority of the participating countries. This integrated design involves a rotational panel in which a new sample of households and persons is introduced each year to replace a part (normally one quarter) of the existing sample. Persons enumerated in each new sample are followed-up in the survey for four (or more) years. A common rotational sample of this type yields each year a cross-sectional sample as well as longitudinal samples of various durations. In most situations, these sample data have to be weighted to make them more representative of the target population of the survey. The complex structure of the sample means that the corresponding weighting procedures can also be quite complex. The weighting procedure described here is in fact based on detailed recommendations originally developed by one of the present authors (Verma, 2006). These recommendations have been adopted by Eurostat and are being implemented in EU-SILC national surveys. The present paper aims to provide a more systematic and clearer description of the weighting procedures, also introducing some refinements so as to enhance the consistency and completeness of original recommendations. The rest of Section 1 summarises the main features of EU-SILC, such as its scope and content, data structure, longitudinal follow-up (tracing) rules, and in particular the integrated rotational panel design adopted by most countries. The objective is to provide the necessary background for describing the recommended weighting procedures. Further details may be found in various EU-SILC regulations published in the Official Journal of the European

Community, the many technical reports produced by Eurostat, and also in the references already cited above.

Section 2 provides an overview of the integrated weighting procedure. It clarifies the different types of weights required for the different types of data coming out of EU-SILC annually. With a 4-year rotational design, once established, one cross-sectional dataset and three longitudinal datasets, respectively of 2, 3 and 4 years duration are generated each year. We explain the concepts of 'initial' and 'base' weights, on the basis of which the required cross-sectional and longitudinal weights can be constructed as an integrated whole.

Section 3 deals with the construction of initial weights, by which we mean weights to be applied to each new sample as it is introduced into the survey. A step-by-step procedure is described, starting from design weights, followed by adjustments for non-response and calibration to external controls, and finally trimming and scaling as required to obtain the initial weights. Of course, these procedures are applicable to any survey and are not specific to a rotational panel. However, in a panel survey it is particularly important to construct good initial weights for each new sample as it is introduced, since these determine the quality of the longitudinal weights, as well as the cross-sectional weights, to be applied in subsequent years the sample remains in the survey.

An innovative feature of the weighting procedure is the concept of base weights, developed originally in the context of European Community Household Panel (Verma and Clemenceau, 1996). Starting from the initial weight of each individual in the original sample, the person's base weight is constructed - for each subsequent year the person remains in the survey - with the objective of compensating for panel attrition. The procedure for constructing base weights is described in Section 4.

Sections 5 and 6 describe the procedures for constructing, respectively, longitudinal weights and cross-sectional weights. These two types of weights can in fact both be obtained in a straightforward way from the same base weights. This makes the whole procedure internally consistent and integrated.

Finally, in Section 7 we comment on some important practical aspects in the implementation of the weighting procedures.

1.2.Scope and content of EU-SILC

As noted, EU-SILC aims to be a flexible yet comparable instrument covering data and data sources of various types.

In terms of the substantive content, four types of data are involved: (i) variables measured at the household level; (ii) information on household size and composition and basic characteristics of household members; (iii) income and other more complex variables measured at the personal level, but aggregated to construct household-level variables (which may then be ascribe to each member for analysis); and (iv) more complex non-income or 'social' variables collected and analysed at the personal level.

For set (i)-(iii) variables, a sample of households including all household members is required. Among these, sets (i) and (ii) are normally collected from a single, appropriately designated respondent in each sample household. Alternatively, some or all of these data may be compiled from registers or other administrative sources.

Set (iii) variables - concerning mainly, but not exclusively, the detailed collection of household and personal income - must be collected directly at the personal level, covering all persons in each sample household. In many countries, these income and related variables are collected through personal interviews with all adults aged 16+ in each sample household. This collection is normally combined with that for set (iv) variables, since the latter must also be collected directly at the personal level. These are the so-called survey countries. By contrast, in some countries, set (iii) variables are compiled from registers and other administrative sources, thus avoiding the need to interview all members (adults aged 16+) in each sample household. These are the so-called register countries.

Set (iv) variables are normally collected through direct personal interview in all countries. These are too complex or personal in nature to be collected by proxy, nor are they available from registers or other administrative sources. For the survey countries, this collection is normally combined with that for set (iii) variables as noted above, both being based on a sample of complete households, i.e. covering all persons in each sample household. However, from the substantive requirements of EU-SILC, it is not essential — in contrast to set (iii) variables — that set (iv) variables be collected for all persons in each sample household, since these variables need not to be aggregated to the household level. It is therefore sufficient to do this collection on a representative sample of persons. This option is normally followed in register countries, since for these countries interviewing all household members for set (iii) is not involved. In countries which choose to do so, the sampling process involves the selection of persons (usually one adult member aged 16+ per household) either directly or, optionally, through a sample of households.

1.3.Data structure

Hence different types of units of analysis are involved in EU-SILC for which sample weights have to be defined: (i) private households; (ii) all persons residing in sample households; (iii) all household members aged 16+; and optionally (iv) one selected adult per sample household. One may also be interested in special groups, such as children.[1]

Another dimension is that both cross-sectional and longitudinal data are required. The cross-sectional component covers information pertaining to the current and a recent period such as the preceding calendar year. It aims at providing estimates of cross-sectional levels and of net changes from one period (year) to another. The longitudinal component covers information compiled or collected through repeated enumeration of individual units, and then linked over time at the micro-level. It aims at measuring gross (micro-level) change and elucidating the dynamic processes of social exclusion and poverty. Both cross-sectional and longitudinal micro-data sets are updated on an annual basis. In EU-SILC a period of four years is taken as the minimum duration for longitudinal follow-up at micro level.

Combining the various types of units and the time dimension, the new data sets disseminated each year consist of the following: cross-sectional data pertaining to the most recent reference year for households and persons; data pertaining to three different longitudinal periods, covering 2, 3 and 4 years preceding the survey, only for persons.

1.4.The integrated design

Verma and Betti (2006) illustrate a typology of possible data source structures in EU-SILC. A single *integrated* source covering all components – cross-sectional and longitudinal, income and social - is by far the most common one adopted by countries up to now. Throughout, we will describe the weighting procedure with reference to this integrated design.

The basic idea is as follows. At any one time, the sample is made up of, say, 4 short-term subsamples or panels. Each year one new panel is added to stay in the survey for 4 years, and then dropped to be replaced by another new panel. Movers from the original sample are followed-up to their new location for up to the time their panel remains in the survey.[2] Each panel provides a longitudinal sample of

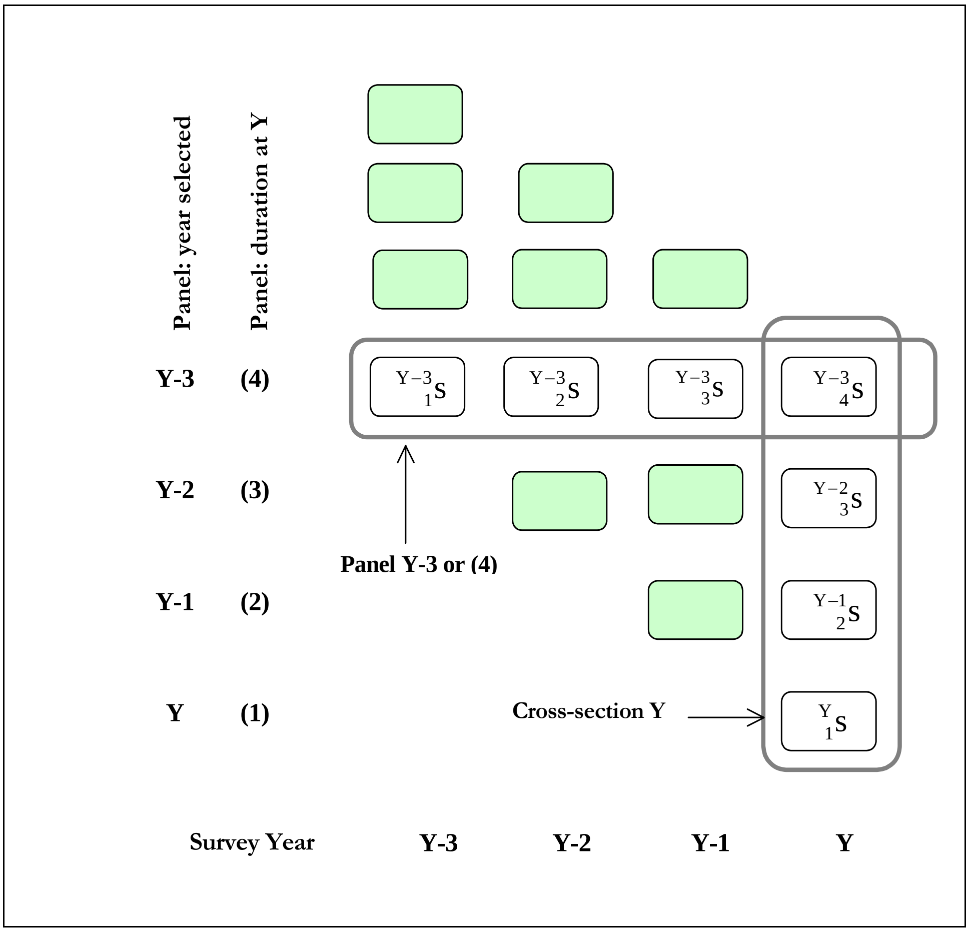
[1] Following the EU-SILC terminology, we will refer to these different types of units and to the associated data files as follows: H (‘household’), R (‘register’ covering all members), P (‘personal’ covering all adults aged 16+), S (‘selected respondent’), and Q (children or other special groups).

[2] According to *Commission Regulation (EC) No 1982/2003 as regards the Sampling and Tracing Rules* rotational design refers to “the sample selection based on a number of subsamples or replications, each of them similar in size and design and representative of the whole population.

the chosen duration. The units present at a given time from all the panels are appropriately put together to constitute the cross-sectional sample. Clearly, an important advantage of this scheme is that both cross-sectional and longitudinal data are obtained from the same common set of units. This overlap is highly economical, and also maximises internal consistency between longitudinal and cross-sectional statistics produced from the survey.

The rotational scheme is illustrated in Figure 1. It also specifies the notation we will use.

Figure 1. The panel and cross-sectional samples



From one year to the next, some replications are retained, while others are dropped and replaced by new replications".

In the above diagram, ${}^{Y-3}_{t}s$ is a typical panel sample introduced in year (Y-3); this sample (more strictly, the samples derived from it according to certain follow-up rules) is enumerated over four years (Y-3) to Y. Longitudinal samples of 2 to 4 years duration are constructed by putting together different panels of this type, as will be discussed later (Section 5). The full cross-sectional sample at year Y is composed of four panels ${}^{Y-t+1}_{t}s$, t = 1 to 4, as shown in Figure 1.

1.5.Sample follow-up (tracing) rules

Follow-up rules are required in a household panel to preserve its representative nature over time. In the case of EU-SILC, the relevant Commission Regulation (European Community, 2003) specifies these formally as follows.

“ (a) Initial sample: refers to the sample of households or persons at the time it is selected for inclusion in EU-SILC.

(b) Sample persons: means all or a subset of the members of the households in the initial sample who are over a certain age.

(c) Age limit used to define sample persons: … In countries with a four-year panel using a sample of addresses or households, all household members aged 14 and over in the initial sample shall be sample persons. In countries with a four-year panel using a sample of persons, this shall mean the selection of at least one such person per household. ...

(e) Sample household: means a household containing at least one sample person. A sample household shall be included in EU-SILC for the collection or compilation of detailed information where it contains at least one sample person aged 16 or more. ...

(g) Co-residents (non-sample persons): all current residents of a sample household other than those defined above as sample persons. ...

(i) Age: refers to the age at the end of the income reference period.”

Ideally, all persons, irrespective of age, in initial households should be followed-up over the panel duration. The age limit of 14 has been introduced in EU-SILC regulations merely for practical reasons. The impact of this limitation is generally not important because we can expect very few children under 14 to move to a new address ‘alone’, without being accompanied by one or more adults in their original household at the time of selection. Hence for the present discussion, we will consider as sample persons *all individuals in the initial households, irrespective of age, and assume that all these are supposed to be followed-up to their new address if they move*. The few cases where the slightly restrictive rules adopted in EU-SILC do not require a follow-up can simply be treated as non-respondents.

The above applies to the normal situation in survey countries where all adults in the household are supposed to be followed-up. Additional problems can arise in ‘register countries’ where only one adult per household is selected for follow-up. We comment on this in the concluding Section 7.3.

2. Overview of the sample weighting procedure

2.1.The importance of weighting sample data

As noted by Kish (1992), “why, when and how” to weight are among the fundamental and most common questions in estimation and analysis using sample survey data. The answer to these questions may depend on the context, the data source and the type of data and analysis involved. Sometimes there are sharp arguments as to whether or not it is appropriate or necessary to weight at all. In the case of an intensive and complex survey of limited size, such as a household panel like EU-SILC, we believe that the answer to the question is quite clear: in most situations it is both necessary and useful to weight the sample data to compensate for imperfection of the achieved sample.

Firstly, weighting is introduced to compensate for differences in sampling probabilities. Such weighting is essential if those differences are large. Another important reason making weighting necessary is the high rates of non-response and attrition over time in panel surveys. Non-response and attrition is often not only large but also selective, such as being higher for households at the extreme ends of the income distribution.

Further improvement in the representativeness of the sample can be made through its calibration against more reliable external information such as age-sex composition and geographical distribution of the population, and the distribution of households by type and size. Such calibration can reduce biases in the sample due to non-response, non-coverage and other distortions, and also reduce variances. In the EU context at least, a great deal of external information is available in most countries for calibration of the sample.

In household panel surveys, two additional factors make the weighting procedures complex: one, the survey deals with units of different types (households, all members, adults); and two, both cross-sectional and longitudinal components are involved, each with its own weighting requirements.

In all situations, the legitimate objective of weighting the sample data is to reduce the resulting mean-square-error of the estimates from the survey. In special situations (such as optimal allocation and calibration) weights are introduced to reduce variance. However, in most situations the introduction of unequal weights tends to inflate variances, and the purpose of weighting is to reduce biases in the estimation. Compromise is often required to balance these effects and control the mean-square-error. Even a small number of extreme values, in particular at the upper end of the weight distribution, can inflate the variance significantly. In recommending the rather complex weighting procedures for households’ panels in this paper, we must emphasise the importance of checking for extreme values or large variations being introduced in the weights. Alternative procedures have been suggested to limit the range of weight values encountered, such as ‘shrinkage’ of weights (Spencer and Cohen, 1991). In our experience, the trimming of extreme values is often the simplest.

2.2. Initial and follow-up weights in a household panel

There are two broad aspects of the weighting procedure in a household panel: initial weighting at the time of introduction of the panel units into the survey; and modification of the sample and of the initial weights in the following period.

Initial weights

Irrespective of the details of the design, any panel introduced into the survey begins with a sample of households, and hence also of persons who are members of those households. We assume that by design, these samples are probability samples. We term weights assigned to the achieved sample as it first appears in the survey as the *initial weights*. These will be discussed in more detail in Section 3.

The procedures for constructing the initial weights are in fact the same as those used for all types of surveys, including purely cross-sectional surveys. The development of initial weights is generally performed in stages: three or four stages may be involved, possibly some of them involving multiple steps. Some useful references in the literature include Kish (1992), Kalton and Kasprzyk (1986), Little (1986), and, for detailed application to ECHP, Verma and Clemenceau (1996).

Follow-up sample and weighting

The initial sample provides the set of individuals, called *sample persons*, who are followed-up over duration of the panel. The sample needs to be modified to reflect changes in the target population over time. At least three types of adjustments are involved. The first concerns changes in the target population in private households: individuals leaving the population due to death or out-migration, and new persons joining the population through births or in-migration. Unless supplementary samples of new entrants into the population are added, the panel cannot reflect in-migration fully. The second arises from non-response and other losses in the sample. The weights of units remaining in the sample need adjustment in order to reduce the impact of non-response on sample representativeness. The third source of change in the sample is the entrance of non-sample persons into the households included in the sample. Many household panel surveys, including EU-SILC, collect a great deal of information on these non-sample members or *cohabitants*, while they are living with some sample person, in order to measure the circumstances (such as income) of the whole household. It is desirable (efficient) to exploit the information at the individual level collected for these cohabitants. Appropriate weighting schemes are required for their incorporation into the sample, which of course must continue to reflect the corresponding target population. The basis of these procedures is provided by the *weight share* method expounded by Ernst (1989). Some other useful references in the literature include Lavallée (1995), Kalton and Brick (1995), Lavallée and Caron (2001), Deville and Lavallée (2006), and, in the specific context of ECHP, Verma and Clemenceau (1996).

2.3.Structure of the weighting procedure; notation

In this section, we provide an overview of the proposed weighting procedure. Another important objective is to introduce the notation we use throughout.

Consider the rotational structure represented in Figure 2. At a given time, the total sample is made up of a number of panels. The following notation will be used for panel-specific weights (w):

${}_{t}^{Y}w_{j}^{(U,X)}$ where the identifiers of the weight are as follows

Y refers to the year the panel was selected; alternatively, it may be a more convenient to identify the panel by its current age (T).

t $1 \le t \le T$, refers to the panel's age in general.

U identifies different types of units: H (households); R (all household members); P (all members aged 16+); S (where relevant, the selected respondent); and Q (children or other population subsamples of interest).

X identifies the type of weights at wave 1 (D, N, F, I, see below). Thereafter we deal with base weights X = B.

j is a particular unit (household or person). It is understood that (j) belongs to sample s identified in the same way as w: $j \in {}_{t}^{Y}s^{(U)}$.

When certain identifiers are nor needed, they are dropped for simplicity. For instance, most commonly we deal with base weight (X = B) of individual persons (U = R), for any unspecified unit (j), so that the weight can be identified as ${}_{t}^{Y}w$ and the corresponding sample as ${}_{t}^{Y}s$.

When the only needed reference is to t, we will employ the simplified and commonly used notation w_t and s_t, respectively for the weight and the sample.

A longitudinal sample, ${}_{t}^{Y}s^{(L)}$, is defined as the set of individuals (j) who have been enumerated in the survey throughout the period 1 to t inclusive:

$$ {}_{t}^{Y}s^{(L)} = {}_{1}^{Y}s \cap {}_{2}^{Y}s \cap \ldots \cap {}_{t}^{Y}s \,; \qquad (1) $$

when convenient, we use simplified notation corresponding to s_t, namely $s_{t}^{(L)}$.

Initial and base weights

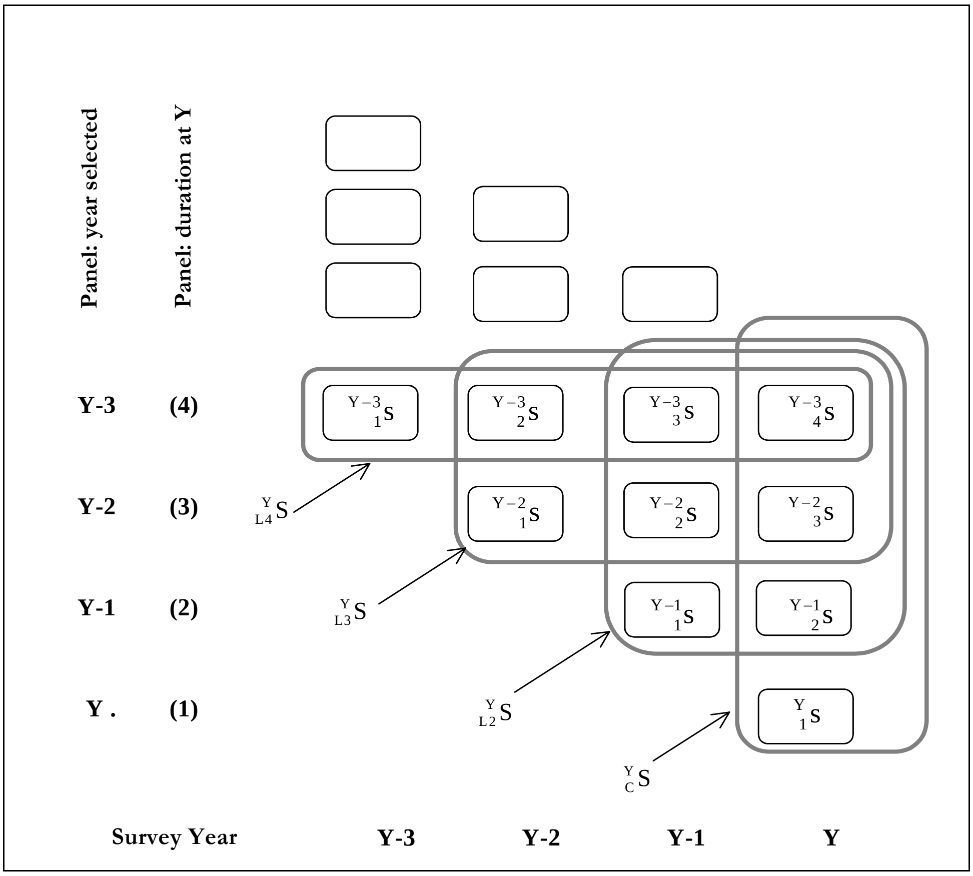
As described in Section 3, the final wave 1 weights are constructed involving a step-by-step procedure defining:

- D, design weights, in inverse proportion to the unit's selection probability;
- N, the above weights modified to reflect non-response at wave 1;
- F, the resulting weights calibrated to 'fit' certain specified control distributions;
- I, trimming and scaling of the above, giving the initial wave 1 weights.

For individuals remaining in the sample during subsequent periods t >1, we define a sequence of base weights (B), which take into account attrition of the panel. To start with, the base weight is defined to be identical to the wave 1 initial weight:

$${}^{Y}_{1}w_{j}^{(U,B)} \equiv {}^{Y}_{1}w_{j}^{(U,I)} . \qquad (2)$$

Figure 2. The annual cross-sectional and three longitudinal samples



Cross-sectional and longitudinal weights for the full sample

The panel-specific weights (w) are 'intermediate' quantities. These are put together to construct weights (v) for the full sample comprised of different panels.

Weights v are of different types (C, L2, L3, L4; see Figure 2) and represent the actual weight variables used in analysis. We indicate these as:

$${}^{Y}_{X}v^{(U)}_{j} \; .$$

Firstly, for cross-sectional weights (X = C), for reference year Y, we have:

$${}^{Y}_{C}v^{(U)}_{j} \; ,$$

defined for the full cross-sectional sample, i.e. all units in the survey at year Y:

$${}^{Y}_{C}S = {}^{Y}_{1}s \cup {}^{Y-1}_{2}s \cup {}^{Y-2}_{3}s \cup {}^{Y-3}_{4}s \, .$$

It can be seen that each year, results for three new longitudinal samples of different durations – 2, 3 and 4 years – become available. We identify the corresponding weights, respectively, X = L2, L3 and L4. The constituent samples, S, to which these apply are:

$$\begin{aligned}{}^{Y}_{L2}S &= ({}^{Y-3}_{3}s \cap {}^{Y-3}_{4}s) \cup ({}^{Y-2}_{2}s \cap {}^{Y-2}_{3}s) \cup ({}^{Y-1}_{1}s \cap {}^{Y-1}_{2}s) \\ &= ({}^{Y-3}_{3}s \cap {}^{Y-3}_{4}s) \cup ({}^{Y-2}_{2}s \cap {}^{Y-2}_{3}s) \cup {}^{Y-1}_{2}s^{(L)} \; ,\end{aligned} \tag{3}$$

$$\begin{aligned}{}^{Y}_{L3}S &= ({}^{Y-3}_{2}s \cap {}^{Y-3}_{3}s \cap {}^{Y-3}_{4}s) \cup ({}^{Y-2}_{1}s \cap {}^{Y-2}_{2}s \cap {}^{Y-2}_{3}s) \\ &= ({}^{Y-3}_{2}s \cap {}^{Y-3}_{3}s \cap {}^{Y-3}_{4}s) \cup {}^{Y-2}_{3}s^{(L)} \; ,\end{aligned} \tag{4}$$

$${}^{Y}_{L4}S = ({}^{Y-3}_{1}s \cap {}^{Y-3}_{2}s \cap {}^{Y-3}_{3}s \cap {}^{Y-3}_{4}s) = {}^{Y-3}_{4}s^{(L)} \; . \tag{5}$$

The quantity ${}^{Y}_{t}s^{(L)}$ has been defined in (1).

Note that generally there is a large overlap in terms of units and data among the longitudinal samples and, in an integrated rotational design, also with the cross-sectional sample.

Nevertheless, each sample $S = ({}^{Y}_{C}S, \, {}^{Y}_{L2}S, \, {}^{Y}_{L3}S, \, {}^{Y}_{L4}S)$, with corresponding X = (C, L2, L3, L4), has its own distinct set of weights. Furthermore, in general, the set of weights differ by the type of unit U = (H, R, P, S, Q).

3. Initial weights

3.1. A step-by-step procedure

We begin from the construction of initial weights given to households and persons as they first enter the survey. Throughout, we consider one panel at a time. This weighting procedure is of course the same as that in any comparable cross-sectional survey, but we aim at bringing out some important theoretical and practical points and make some recommendations.

In developing the weights, the best possible use has to be made of the information available, both internal to the sample and from external sources. This includes information on: (i) coverage and selection probabilities; (ii) characteristics and circumstances of non-responding units; (iii) the structure and characteristics of the target population.

Such information can be used in a systematic manner to apply weights in a step-by-step procedure. We may identify the following four steps:

- Design weights. Each household and person in the sample is weighted in inverse proportion to its probability of selection.
- Non-response weights. These are introduced to reduce the impact of non-response on the structure and characteristics of the achieved sample. These weights are derived on the basis of items of information which are available for both responding and non-responding units.
- Calibration weights. This refers to the adjustment of the distributions of the weighted sample according to various characteristics to agree with more reliable information on those distributions available from sources external to the survey. In a household survey, distributions both in terms of households and persons may be involved.
- Trimming and scaling. This refers to adjustments made to avoid extreme values in the distribution of weights. The objective is to avoid large increases in variance which result from the presence of extreme weights, even though such adjustment may introduce some bias. For many purposes, only the relative values of weights are relevant, and their overall scaling is immaterial. However, an appropriate scaling of these quantities is usually necessary for clarity and convenience.

Some basic features of the weighting procedure described below may be noted:

- The weights are applied in a sequence. That is, at any step after the first, the weights are computed from sample values already weighted according to the results of all preceding steps. Thus weighting for non-response is determined from the sample results weighted by the design weights. This simplifies the computations, and shows more clearly the contribution of each step in adjusting the weights.
- In a household panel survey, typically the data are used simultaneously for analysis at the household and the personal levels. It is desirable, therefore, to use a weighting procedure which ensures consistency between analyses involving the two types of units. The recommended procedure is 'integrative weighting' at wave 1, which ensures a uniform weight for all members of a household, the same as the weight of the household.
- The specific variables involved at each step and the sources of the data used may vary from one survey to another. Nevertheless, certain variables can be expected to be important in most circumstances, such as

geographic location of the household, household size and composition, and distribution of the population by age, sex and other basic characteristics.

- It is possible in principle to combine more than one step into a single procedure. However, it is desirable that each step is separately implemented, where possible.
- The final weight of a unit is the product of the weighting factors determined at each step. The resulting weights at each step may be appropriately scaled, such as to average 1.0 per sample unit.

3.2. Design weights

Design weight of a unit is inversely proportional to its probability of selection into the sample, though allowances can sometimes be made for known exclusions or under-coverage of parts of the study population.

In a panel survey, the design weights are defined only at wave 1, when the unit is selected into the sample. The unit, or derivations from it, at subsequent waves may be considered subject to 'indirect sampling' (Deville and Lavallée, 2006), but in the present case the concept of base weights described in the next section provides a clearer and simpler alternative.

There are essentially two ways of selecting the sample for a household panel. The first is to select households - or addresses, dwelling or other structures which contain households - directly from appropriate lists. The procedure directly gives the probability of selection of the household (Ph). In the notation introduced in Section 2.3, the household design weight is:

$${}_1w_h^{(H,D)} = k/P_h \,,$$

where k is an arbitrary scaling factor to be chosen.

In so far as all persons in the household are automatically taken into the sample once the household is selected, the members' selection probabilities and hence weights are identical to those of their household. The above equality applies also to any subpopulation in the household, such as persons aged 16+ (P), or children (Q):

$${}_1w_{j\in h}^{(P,D)} = {}_1w_{j\in h}^{(Q,D)} = {}_1w_{j\in h}^{(R,D)} = {}_1w_h^{(H,D)} \tag{6}$$

In situations where one adult in the household is selected as the respondent for a personal interview, we have the selection probability and weight of selected respondent (S) as:

$$P_{s\in h} = P_h / n_h \; ; \; {}_1w_{s\in h}^{(S,D)} = {}_1w_h^{(H,D)}.n_h \,, \tag{7}$$

where n_h is the number of household members eligible to be a selected respondent.

The other way of obtaining a sample of household is to first select a sample of individuals meeting certain specified criteria, and then construct a household around each such selected individual. A household is selected through its association with one or more such individuals. Normally, the latter are selected from a list of adults. In so far as each eligible household contains at least one such person in the list, the household receives a non-zero probability of selection. The population of households not containing such a persons will not be covered in the survey.

The probability of selection of a household (Ph) relates to that of the individuals (Ps) through which it could have been selected as:

$$P_h = \sum_{s \in h} P_s \, .$$

Concerning weights we have:

$$_1w_s^{(S,D)} = k/P_s \, ,$$

$$_1w_h^{(H,D)} = k \Big/ \sum_{s \in h} P_s \, .$$

The person weights (R) of household members are identical to this household weight, as in equation (6).

Usually, persons eligible for selection within a household all have the same probability of selection, so that:

$$\sum_{s \in h} P_s = n_h P_s \, ,$$

giving:

$$_1w^{(H,D)} = \left(\frac{1}{n_h} \right) \cdot {}_1w_s^{(S,D)} \, . \qquad (8)$$

3.3.Non-response adjustment

In a panel, the largest loss of the sample due to non-response generally occurs at the first wave when the household is introduced into the survey. Good, efficient procedures to re-weight the responding cases is, therefore, a critical requirement at wave 1. However, the possibilities are often constrained by lack of information: non-response adjustment has to be based on characteristics which are known for both responding and non-responding households.

Here the concern is with non-response at the stage of the *household interview*, which obtains information at the household level and enumerates the population in the household. The procedure involves estimating response rates or propensities to response as functions of characteristics available for both responding and non-responding households, including characteristics of the area where the household is located. This is also true when a sample of selected respondents has been used, except that personal characteristics of the selected individuals are also useful as determinants.

There are two commonly used procedures for non-response weighting.

Weighting within classes

The first is to modify the design weights by a factor inversely proportional to the response rate within each weighting class:

$$ {}_1w_{j\in K}^{(H,N)} = {}_1w_{j\in K}^{(H,D)} . \frac{\overline{R}}{R_K}, \qquad (9)$$

where R_K is the response rate in weighting class K, and $\overline{R}$ their average (this is just a constant introduced for convenience). The response rates should be computed with data weighted by the design weights:

$$R_K = \frac{\sum_{j\in Kr} {}_1w_j^{(H,D)}}{\sum_{j\in Ks} {}_1w_j^{(H,D)}},$$

where the set Kr represents the responding units in the class, and Ks the selected units in the same class.

By ‘class’ we mean an appropriately determined grouping of units. It is common to use sampling strata or other geographical partitions as weighting classes. Numerous, very small weighting classes can result in large variations in R_K values, and should be avoided. On the other and, if only a few broad classes are used, little variation in the response rates across the sample may be captured – making the whole re-weighting process ineffective. On practical ground, classes of average size 100-300 units may be recommended.

Weighting according to response propensities

An alternative is to use a regression-based approach. When many auxiliary variables are available, this approach is preferable to the previous one. Using an appropriate model such as logit regression, response propensities can be estimated as a function of auxiliary variables (X) which are available for both responding and non-responding cases:

$$R_h = \Pr(r_h = 1 | X_h),$$

where rh is a (0,1) response indicator, equal to 1 if household h has been successfully enumerated; Xh is an appropriate set of auxiliary variables related to the households’ response propensity, and Rh is that propensity predicted using regression.

The weighting of each responding unit is adjusted by the inverse of the estimated response propensity, in the same way as by the inverse of response rates in the previous method (Little, 1986).

A very important point when using the regression approach is to ensure that weights assigned are confined to be within reasonable limits. The regression can predict zero or even negative values, which of course must be rejected. The problem is more general than that: extreme values should not be permitted. It is for this reason that it is very important to check the distribution of the resulting weight adjustments, and apply trimming or similar procedures to remove extreme values. This is discussed further in Section 3.5 below.

Non-response affects the enumeration of households and of all household members in exactly the same way. Hence the non-response adjustment to weights retains the basic equality in equation (6).

Some practical aspects

Generally it is useful to apply the adjustment in two steps:

(i) for non-contact (of households and/or of selected individuals); and

(ii) for non-response, once a contact with the households or the persons concerned has been made.

For both steps, especially for (i), area-level characteristics provide a main part of the auxiliary variables explaining non-response. This is because they are more easily available for both responding and non-responding units.

In dealing with the effect of non-response, it is of crucial importance to correctly distinguish between non-eligible and non-responding units. In fact, selected units which turn out to be non-eligible or non-existent must be excluded and not counted as non-respondents. Sometimes, imputation has to be made for units with unknown eligibility-response status, i.e. when it is not clear whether they are non-eligible or non-respondents.

For this purpose it is useful to distinguish the following categories:

1. Households that are out-of-scope: all individuals in these households are, of course, also out-of-scope.
2. Households that are successfully contacted and enumerated. These contain persons that are:
   - 2.a in scope;
   - 2.b out-of-scope;
   - 2.c persons for whom we do not know whether or not they are in-scope.
3. Households that are known to be in-scope, but are non-respondents. These households may contain persons who are actually out-of-scope, but this information is not available.
4. Households that are not contacted, and the eligibility status of the whole household is not known.

In order to impute a definite status for an individual in (2.c) or (3), we may assume that the propensity to be in-scope for these persons can be determined on the basis of the combined group (2.a+2.b), controlling for appropriate auxiliary variables. A logit regression model with appropriate control variables may be used. However, a simpler approach – for instance assigning individuals in (2.c) and (3) an eligibility status with probability equal to the proportion in that status in group (2.a+2.b), would suffice if the incidence of missing information is small.

Persons in group (4) are more likely to be out-of-scope than group (2) or (3) alone. It seems reasonable to assume that the status for group (4) can be imputed on the basis of groups (1)-(3) combined. A regression or a simple approach on the same lines as above may be used.[1]

3.4. Calibration to external data sources

Calibration is carried out in an attempt to ensure that weighted sample sums of specified control variables or categories equal to the known population totals for those variables.

In so far as the first wave of a panel like EU-SILC is subject to high rates of non-response and possibly also to random and systematic distortions of the sample, we consider calibration to be particularly important in the first wave. Once the initial sample has been so adjusted, recalibration at subsequent waves may be done more selectively.

A robust and convenient method of adjusting the weighted sample distribution to a number of external controls simultaneously is the classical iterative proportional fitting or raking (Deming, 1943). Using special algorithms such as those implemented in the INSEE program CALMAR, upper and lower bounds can be imposed on the weight adjustment during calibration. (The limits, however, cannot be made too narrow as the iterative procedure involved becomes slow in converging, and may fail to converge altogether beyond a certain limit.) Deville and Särndal (1992) develop a family of calibration estimators of which the standard general regression estimator (GREG) is a first approximation. A problem with GREG and similar procedures is that they can yield negative weights, which of course does not make sense. In our experience the classical raking procedure, with simple trimming of any weight values outside desirable limits, continues to provide a practical and acceptable approach in many situations. (Concerning trimming, see Section 3.5 below).

Both household and person-level control variables are useful in calibration. Useful variables tend to be similar to those used for non-response adjustment, assuming availability: geographical location, tenure status, household type and size, age-sex composition of the population, etc. In some situations, additional variables may be available.

However, the crucial requirement in calibration is to ensure that the external control variables are strictly comparable to the corresponding survey variables, the distribution of which is being adjusted.

It is desirable that at wave 1, all persons in the same household receive the same weight, so that the weight given to each member is the same as the weight of

[1] Some surveys allow substitution of non-responding units with new units (despite the general undesirability of this practice). In such cases, non-responding original units for which successful substitutions have been made are to be considered as ‘responding units’ in the computation of response rates for the purpose of determining non-response weights.

the household, as in equation (6). Such uniform within-household weighting can be retained, even when the external controls in the calibration are at both the household and the personal level, by special technique known as ‘integrative calibration’.[1]

3.5.Trimming and scaling

Trimming

It is important to ensure that no step in the weighting procedure results in extreme values of the weights. More precisely, what is required is to ensure that large variation in the weight values is not introduced as a result of the adjustments. This is because that variation inflates variance of the estimates from the survey (Kish, 1992). In fact, control of extreme values and large variations in weights is desirable at each stage of adjusting the weights - after non-response adjustment, and then again after calibration.

A common approach, as indicated earlier, is to trim extreme values.2 However, there is no rigorous procedure for general use for determining the limits for trimming. While more sophisticated approaches are possible, it is desirable to have a simple and practical approach. Such an approach may be quite adequate for the purpose, at least in situations where the main problem is caused by a limited number of extreme values assigned during the adjustment process.

After calculation of non-response weights, we have recommended and used the following simple procedure: any computed non-response weights outside the following limits are recoded to the boundary of these limits:

$$\frac{1}{L} \le \left\{ \frac{{}_1w_h^{(H,N)}}{{}_1\overline{w}^{(H,N)}} \bigg/ \frac{{}_1w_h^{(H,D)}}{{}_1\overline{w}^{(H,D)}} \right\} \le L\,, \qquad (10)$$

where ${}_1\overline{w}^{(H,D)}, {}_1\overline{w}^{(H,N)}$ are respectively the mean values of household design and non-response weights, and L is some appropriate upper bound for the adjustment in weights. L = 3 could be a reasonable value for this parameter.[3]

After calibration, we can follow the same form of check and correction for extreme values:

$$\frac{1}{L} \le \left\{ \frac{{}_1w_h^{(H,F)}}{{}_1\overline{w}^{(H,F)}} \bigg/ \frac{{}_1w_h^{(H,D)}}{{}_1\overline{w}^{(H,D)}} \right\} \le L\,, \qquad (11)$$

[1] The procedure is described in, for instance, CALMAR documentation from INSEE (document no. F9310, November 1993). See also a brief description in Verma and Clemenceau (1996), Section 5.

[2] For a more technical discussion of the issue, see Potter (1988, 1990).

[3] Since trimming alters the mean value of the weights, the above adjustment may be applied iteratively, with the mean re-determined after each cycle. A very small number of cycles should suffice normally.

where each quantity is divided by its mean, so as to appropriately scale the weight values being compared.

Note that in both (10) and (11), the limits imposed are in terms of the ratio of the adjusted to the original design weights, i.e. the factor by which the design weights are being modified.

Any trimming or similar adjustment can be applied identically to household and person level weights, retaining the equality (6) for the trimmed weights, for instance:

$$ {}_1w_{j\in h}^{(R,F)} = {}_1w_h^{(H,F)} \; . $$

Scaling

As will be noted in Sections 5 and 6, the final longitudinal and cross-sectional weights actually used in data analysis may be scaled such that their sum is proportional to the size of the target population. Such scaling allows data from different countries and surveys to be put together for combined analysis without further adjustment to the weights (Verma, 1999).

However, at the present stage of the weighting procedure, we are considering one out of a number of panels which constitute the total sample at any given time. These panels are to be subsequently put together to constitute the full sample. Unlike the previous case, this is not an aggregation over different populations, but of samples representing the same population. In such aggregation, it is appropriate that each sample contributes inversely proportional to its variance, i.e., approximately in direct proportion to its sample size. Hence the weights should be scaled such that their sum is proportional to the sample size of the panel concerned – which is obtained most simply by scaling the weights to average 1.0 per unit. Such scaling allows data from different panels to be put together to construct the total sample without further adjustment to the weights.

Finally, we note that it is desirable to retain equality of the type in equation (6) also for the final rescaled weights for households and persons. To be precise, we need to choose whether to determine the required scaling factor with reference to the household or the person level sample. The latter is more appropriate in so far as the individual person is the more commonly used unit of analysis, as is the case for instance in the analysis of income distribution and poverty. Hence we define the required re-scaling factor as:

$$ F_R = n_R \Big/ \sum_{j=1}^{n_R} {}_1w_j^{(R,F)} \; , $$

where nR is the number of individuals enumerated in households of the panel sample at wave 1. The rescaled initial weights are:

$$ {}_1w_j^{(R,I)} = {}_1w_j^{(R,F)}.F_R \, , \quad {}_1w_h^{(H,I)} = {}_1w_h^{(H,F)}.F_R \; , $$

so that the following equality is maintained:

$$ {}_1w_{j\in h}^{(R,I)} = {}_1w_h^{(H,I)} . \tag{13}$$

It is not necessary to define these quantities for other subpopulations (P, Q, S) at this stage, but these can be taken simply as:

$$ {}_1w_j^{(P,I)} = {}_1w_j^{(Q,I)} = {}_1w_j^{(R,I)} , \tag{14}$$

$$ {}_1w_j^{(S,I)} = \left\{ \frac{{}_1w_j^{(R,I)}}{{}_1w_j^{(R,D)}} \right\} \cdot {}_1w_j^{(S,D)} , \tag{15}$$

that is, for all the different type of units, the same modification (the first factor in the last equation) is applied to the original design weights.

4. Base weights

4.1.Longitudinal population and longitudinal sample

In the following we consider a panel selected fresh at time t = 1 from population P1, and then enumerated for a total of T (say 4), waves, t = 1 to T. In this section we are primarily concerned with base weights (B) of the total population of persons (R) and for convenience will use the following simplified notation unless required otherwise (see Section 2.3):

$$ w_t = {}_tw_j^{(R,B)} ,\ j \in s_t^{(L)} ,$$

where $s_t^{(L)} = s_{t-1}^{(L)} \cap s_t$, the longitudinal sample of persons from time 1 to t. We define the base weight at wave 1 as being identical to the initial weight defined in the previous section: $w_1 = {}_1w_j^{(R,B)} = {}_1w_j^{(R,I)}$.

From Section 3, these weights have been determined at wave 1 such that sample s1 with weights w_1 represents population P1, which we write as $(s_1, w_1) \longrightarrow P_1$.

Base weights are defined for persons but not for households. According to the following procedure, members of the same household may have different weights after wave 1.

By longitudinal population in the interval t = t1 and t2 is meant all persons who have remained in the target population throughout the period t1 to t2, inclusive. Let $P_t^{(L)}$ be the longitudinal population in the interval 1 to t. It comprises persons who were in the target population at wave 1, and have remained in the population up to and including time t. $P_2^{(L)}$ differs from P_1 by persons (say OUT2) who have left the population between years 1 and 2: $P_2^{(L)} = P_1 - OUT_2$. This may for instance be due to death, migration out of the country or movement to an institution. Similarly, $P_3^{(L)}$ differs from P_1 by persons who have left between times 1 and 2 (OUT_2) or between times 2 and 3 (OUT_3):

$$P_3^{(L)} = P_2^{(L)} - OUT_3 = P_1 - (OUT_2 + OUT_3).$$

In general, $P_T^{(L)} = P_1 - \sum_{t=2}^{T} OUT_t$. The longitudinal population $P_t^{(L)}$ differs from the actual (cross-sectional) population P_t at t, as the former does not include persons who are born or have migrated into the target population since time 1, and have remained in that population since that time.

Longitudinal sample $s_t^{(L)}$ is defined as individuals who have been members of an enumerated household throughout the period 1 to t inclusive. (The person is not necessarily a member of the same household throughout this period). $s_2^{(L)}$ differs from s_1 by persons in the original sample who left the population between years 1 and 2 (out_2), and by persons still in the population (x2) whose household was enumerated at t = 1 but not at t = 2: $s_2^{(L)} = s_1 - (out_2 + x_2)$. We assume that (out_2, w_1) is a representative sample of OUT_2; consequently:

$$\left[(s_2^{(L)} + x_2), w_1\right] \longrightarrow (P_1 - OUT_2) = P_2^{(L)}. \qquad (16)$$

4.2. The evolution of base weights starting with wave 1

We assume that all persons enumerated at wave 1 are eligible for follow-up to the next wave. Those not successfully followed-up are considered non-respondents (unless re-classified as out-of-scope at the later time).1

The objective is to determine new base weights, w_t, at $t \geq 2,$ such that:

$$\left[\left(s_t^{(L)} + x_t\right), w_{t-1}\right] \longrightarrow \left(P_{t-1}^{(L)} - OUT_t\right) = P_t^{(L)},$$

is transformed with the new weight wt to:

$$\left(s_t^{(L)}, w_t\right) \longrightarrow \left(P_{t-1}^{(L)} - OUT_t\right) = P_t^{(L)}. \qquad (17)$$

In order to determine base weight w_t from known w_{t-1}, we use the following procedure. Consider the set $\left(s_{t-1}^{(L)} - out_t\right)$ of persons enumerated at (t-1) who are still in-scope at t. For each person (j) in this set, we can define a binary variable:

rj=1 if the person is in $s_t^{(L)}$, i.e. is successfully enumerated at t,

rj=0 otherwise, i.e. the person is not successfully enumerated at t.

Using a logit model, for instance, we can determine the response propensity Rj of each person in the above set as a function of a vector of auxiliary variables Xj: $R_j = \Pr(r_j = 1 | X_j)$. For any person (j) in $s_t^{(L)}$ (i.e., with $r_j = 1$) the required weight is $w_t = \frac{w_{t-1}}{R_j}$.

In distinction from non-response adjustment at wave 1 (Section 3.3), the set of auxiliary variables (X) can be rich in content because of the information available on the non-enumerated persons from preceding waves.

In so far as most non-response occurs at the household (rather than the personal) level, a majority of the relevant auxiliary variables (X) will be geographical and household level variables (region, household size and type, tenure), including constructed variables such as household income and household work status.

[1] The following applies concerning EU-SILC. There are certain (small) categories of households and individuals, which, according to EU-SILC rules, are not followed-up. Examples are households not enumerated at wave 1 or at two consecutive waves thereafter, or even not enumerated at a single wave for some specified reasons. Also, persons below a certain age (under 14, or under 16 in some countries) are not followed up if they move ‘alone’, i.e. without being accompanied by an adult sample person. For the present purpose, all these categories are treated as non-respondents – even if these have not been recorded as such in the survey because of the particular tracing rules.

Many person-level variables are also likely to be useful (gender, age, employment status) – the sort of variables correlated with persons moving to a new address, setting up a new household, remaining traceable, etc.

$R_j \leq 1$ by definition. It is necessary to ensure that no negative, zero, or indeed very small values of Rj are allowed. As before, the following practical trimming limit is suggested:

$$\frac{1}{L} \leq \frac{\left\{{}_t w_j^{(R,B)} \Big/ {}_t \overline{w}^{(R,B)}\right\}}{\left\{{}_1 w_j^{(R,D)} \Big/ {}_1 \overline{w}^{(R,D)}\right\}} \leq L$$

, with L = 3, for instance.

Here each quantity is divided by its mean, and (R,D) refers to the individual's original design weight.

4.3.Determining eligibility and response status of units

Application of the above procedure requires that for each person enumerated at (t-1), the person's eligibility and response status at t is known. This means that it is possible to classify each person in $s_{t-1}^{(L)}$ into one of the following categories uniquely: the person

(1) is enumerated at t (i.e., is in set $s_t^{(L)}$);

(2) remains in the population, but is not enumerated at t (is in set x_t);

(3) has moved out of the population (is in set out_t).

In practice, for a proportion of non-enumerated persons, information is not available to determine whether they belong to category (2) or to category (3). Each such person has to be assigned to one or the other of these two groups on the basis of some appropriate model. The procedure is similar to that in Section 3.3 for wave 1, but generally simpler because fewer ambiguous categories are involved.

4.4.Base weights for other categories of persons

In the above, base weights have been defined only for the longitudinal sample starting from wave t = 1, i.e., for all individuals enumerated in the survey throughout the interval 1 to t:

$$j \in s_t^{(L)} = s_1 \cap s_2 \cap \ldots \cap s_t = s_{t-1}^{(L)} \cap s_t \text{, with } s_1^{(L)} \equiv s_1 . \qquad (18)$$

There are two small additional categories of persons who can be assigned non-zero base weights on the basis of the base weights of these longitudinal individuals in the same household, without affecting the weights of the latter. These are:

(i) Children born to sample women. They receive the weight of the mother.

(ii) Persons moving into sample households from outside the survey population. They receive the average of base weights of existing sample members in the household, including (i).[1]

At any time, the sample households include additional categories of members who have so far been assigned zero weights. The main group among these are the co-residents, defined as persons moving into a sample household from another non-sample household in the population. These are given zero base weight. The same applies to children born to non-sample women.

In order to construct longitudinal and cross-sectional weights described in the following sections (which are the actual target variables of interest), procedures are required to assign non-zero weights to some or all of the zero-weighted persons, so that they can be included in analysis of the survey data.

5. Longitudinal weights

5.1.Set of samples requiring longitudinal weights

Consider the sample data becoming available after survey reference year Y in Figure 2. The total sample is made up four panels, selected in years Y (most recent), (Y-1), (Y-2) and (Y-3). For convenience, we will also identify these by reference to their current duration in years, respectively as (1), (2), (3) and (4).

With the exception of the most recently introduced panel (1), the other three panels contribute to newly available longitudinal data sets. As described in Section 2.3, by putting together these panels we obtain three longitudinal data sets $({}_{L2}^{Y}S, {}_{L3}^{Y}S, {}_{L4}^{Y}S),$ respectively of duration 2, 3 and 4 years.[2]

In terms of panels (2)-(4), the composition of these samples is as follows. Different types of panel segments are involved:

- Panels starting from their time of selection (t = 1):

A2: a 2-year longitudinal sample of panel (2), covering years (Y-1) to Y;

A3: a 3-year longitudinal sample of panel (3), covering years (Y-2) to Y;

A4: a 4-year longitudinal sample of panel (4), covering years (Y-3) to Y.

- Panels which are included from a later time (t >1):

[1] In determining the response propensities from one wave to the next as described in Section 4.2, these additional categories can be included in the sample base being followed-up.

[2] There are also other sequences of longitudinal data embedded in the data set shown in the diagram: a 3-year longitudinal sample from (Y-3) to (Y-1) in panel (4); and three 2-year samples, consisting of (Y-3)+(Y-2) and (Y-2)+(Y-1) in panel (4), and (Y-2)+(Y-1) in panels (3). These panels correspond to data which have been produced in previous years.

B2: a 2-year longitudinal sample from panel (3), covering years (Y-1) to Y;
B3: a 3-year longitudinal sample from panel (4), covering years (Y-2) to Y;
C2: a 2-year longitudinal sample from panel (4), covering years (Y-1) to Y.

The concerned longitudinal data sets becoming available at Y are (see equations (3)-(5) in Section 2.3):

$${}^{Y}_{L2}S = A2 + B2 + C2;$$

$${}^{Y}_{L3}S = A3 + B3;$$

$${}^{Y}_{L4}S = A4;$$

with $A2={}^{Y-1}_{1}s \cap {}^{Y-1}_{2}s = {}^{Y-1}_{2}s^{(L)}$; similarly $A3={}^{Y-2}_{3}s^{(L)}$, $A4={}^{Y-3}_{4}s^{(L)}$;

$B2={}^{Y-2}_{2}s \cap {}^{Y-2}_{3}s$; $B3={}^{Y-3}_{2}s \cap {}^{Y-3}_{3}s \cap {}^{Y-3}_{4}s$;

$C2={}^{Y-3}_{3}s \cap {}^{Y-3}_{4}s$.

5.2. Longitudinal weights as distinct from base weights

The longitudinal weights discussed in this section are the actual weights used in longitudinal analysis based on the total available longitudinal sample. For a period t0 to t, $1 \le t_0 < t$, these are defined for all units (from all the available panels) who have been enumerated in the survey throughout this period.

As noted earlier, base weights are for the longitudinal sample of each panel, starting from time t = 1 when the panel is first introduced into the survey. In the special case of t0 = 1, there is essentially no difference between these base and the required longitudinal weights, except that the former are defined for each panel while different panels need to be put together for the latter.

With t0 >1, the longitudinal sample over the period t0 to t includes certain additional categories of individuals not present in the longitudinal sample of original sample persons over the full period 1 to t, to which non-zero weights have to be assigned. The differences from cases (i)-(ii) of Section 4.4 is that assigning non-zero weights to these additional categories will require adjusting the weights of other members in the sample. The additional categories include:

(iii) longitudinal co-residents, entering the household at or before t0 and remaining as household members over the period (t0 to t);

(iv) returnees, defined as sample persons who left the household temporarily but later returned to the household at or before t0 and remained a household member over the period (t0 to t). [1]

Adjustment has also to be made for persons entering the target population before time t0 but subsequent to the time of selection of the panel (t = 1). Of course, these persons cannot be represented in the panel concerned, but they are a part of the longitudinal population over the period (t0 to t1) and therefore should be represented in the sample.

As will be described below, this can be done on the basis of other (later introduced) panels in the rotational design being put together to construct the total longitudinal sample (Ardilly and Lavallée, 2003).

Among those entering the target population, we will need to distinguish between those joining existing households (containing persons from the existing population), and those forming new households. Finally, we may mention that the longitudinal weights also need to take into account the 'non-standard' structure of the total sample in the first years of the survey (the first 3 years in a 4-year panel, as shown in Figure 2).

5.3.Longitudinal weights by age of panel

Our main objective is to determine longitudinal weights for units (of the total population R) in the six sets S = (A2, A3, A4, B2, B3, C2) for a certain reference year Y, as defined above.

Sets A2, A3, A4

In all cases of type A, the longitudinal population and the longitudinal sample are, for the panels concerned, exactly the same as those considered in the construction of the base weights in Section 4. For example, for A2, using the terminology of Section 4:

Longitudinal population P1 – OUT2
Longitudinal sample s1 – (out2+ x2).

Hence for any (j) in sets A2-A4, the required longitudinal weights involved are identical to the base weights defined earlier. Using simplified notation on the left for convenience we may write this, respectively for the three sets, as:

$$ {}^{Y}v_j = {}^{Y-1}_{\;2}w_j^{(R,B)},\ j \in A2 = {}^{Y-1}_{\;2}s^{(L)}, $$

$$ {}^{Y}v_j = {}^{Y-2}_{\;3}w_j^{(R,B)},\ j \in A3 = {}^{Y-2}_{\;3}s^{(L)}, $$

$$ {}^{Y}v_j = {}^{Y-3}_{\;4}w_j^{(R,B)},\ j \in A4 = {}^{Y-3}_{\;4}s^{(L)}. $$

[1] Strictly speaking, to be considered a returnee, the person leaving the household for a period must still have remained within the target population. Otherwise, on departure the person is treated in the weighting process as an out-migrant and on return as an in-migrant into the population.

Set B2

Now consider set B2, i.e. a longitudinal sample covering the second (t = 2) and third (t = 3) year of panel (3). The population at t = 2 may be written as:

$P_2 = P_1 - OUT_2 + BORN_2 + IN_2^{(old)} + IN_2^{(new)}$, where, between time t = 1 and t = 2, OUT_2 are persons who have left the target population, $BORN_2$ are children born to women in the population, $IN_2^{(old)}$ are persons who have moved into the population (from abroad or the non-household sector), but into existing households; $IN_2^{(new)}$ are persons who have moved into the population to set-up new households consisting only of such in-migrants. The sample at t = 2 may be written as $s_2 = s_1 - (out_2 + x_2) + born_2 + in_2^{(old)} + co_2$.

Quantities (out_2, $born_2$, $in_2^{(old)}$) are defined in the same way as the corresponding population quantities. There is no term in the sample corresponding to the population term $IN_2^{(new)}$, since entrants to the population after the sample selection who set-up entirely new households are not represented in the sample. On the other hand, there is no representation in the population for the sample term co_2. It represents co-residents, defined as persons who were already members of the population, and have simply moved from a non-sample private household to a sample household (i.e. to a household containing at least one sample person). Re-arranging, we can write:

$$\left(P_2 - IN_2^{(new)}\right) = P_2^{(L)} + \left(BORN_2 + IN_2^{(old)}\right),$$

$$s_2 = s_2^{(L)} + \left(born_2 + in_2^{(old)}\right) + co_2.$$

Here, the first population term on the right, $P_2^{(L)} = \left(P_1 - OUT_2\right)$, is the longitudinal population considered earlier in the determination of base weights. The first sample term on the right, $s_2^{(L)} = \left(s_1 - out_2 - x_2\right)$, is the longitudinal sample considered earlier; it estimates the longitudinal population with the base weights.

Extending the concept of base weights to cover additional categories of persons as noted earlier:

(i) All children, born2, born to sample mothers during the preceding year, are assigned the base weight of the mother; with these weights, they represent BORN2.

(ii) The mean of the base weights of persons $\left(s_2^{(L)} + born_2\right)$ in the household is assigned to each person in the group $in_2^{(old)}$ in that household; with these weights they represent $IN_2^{(old)}$.

(iii) Co-residents, co2, continue to be assigned a zero base weight.

This means that with these weights s2, the total sample at t = 2, represents the population P2 at t = 2, excepts for entrants $IN_2^{(new)}$ who have formed entirely new households without including any member from the existing population. We represent this as:

$$\left(s_2, {}_2^{(3)}w\right) \longrightarrow \left(P_2 - IN_2^{(new)}\right)$$

where ${}_2^{(3)}w = {}_2^{Y-2}w_j^{(R,B)}$, $j \in {}_2^{Y-2}s$, meaning the base weight at age t = 2 for panel (3), which was introduced into the survey in year (Y-2). Now, using exactly the same logic as used earlier in the construction of the base weights, it follows that:

$$\left((s_2 - out_3 - x_3), {}_3^{(3)}w\right) \longrightarrow \left((P_2 - OUT_3) - IN_2^{new}\right).$$

Since all co-residents are zero weighted, we can re-write the above as:

$$\left[\left\{s_2 - (out_3 + x_3 + co_2) + co_{23}\right\}, {}_3^{(3)}w\right] \longrightarrow \left((P_2 - OUT_3) - IN_2^{new}\right), \qquad (19)$$

where co2 are the co-residents at t = 2, and among those co23 are also present in the household at t = 3. The left side is the longitudinal sample between times t = 2 and t = 3, i.e. individuals (irrespective of whether or not they are from the original, t = 1, sample) who are present at both waves. The first term on the right, $(P_2 - OUT_3)$, is the corresponding longitudinal population, i.e. individuals who are members of the population at both times. With the already computed base weight ${}_3^{(3)}w$, the sample quantities on the left estimate the population values, except for persons $IN_2^{(new)}$ who entered and formed entirely new households between the sample selection and t = 2, and are not represented in the sample.

Set B3

For B3 – a longitudinal sample covering three years (t = 2 to t = 4) of a panel, the procedure is essentially the same, and we obtain the expression:

$$\left[\left\{s_2 - (out_3 + out_4 + x_3 + x_4 + co_2) + co_{234}\right\}, {}_4^{(4)}w\right]$$
$$\longrightarrow \left[P_2 - (OUT_3 + OUT_4)\right] - IN_2^{new}$$
(20)

where co_{234} are co-residents present in the household at t = 2 to 4. Further adjustment is still required to B2 and B3 weights in order to incorporate longitudinal co-residents (co_{23} and co_{234} respectively) with non-zero weights.[1]

[1] Note that in the common simplified notation used in equations (19) and (20), the various sample quantities on the left refer to different panels – to panel (3) in the former, and to panel (4) in the latter.

5.4. Adjustment for returnees and longitudinal co-residents

Set C2; returnees

Now we consider C2, i.e. a longitudinal sample covering the third and fourth years (t = 3, 4) of panel (4). The additional complication in this case is the following.

Sample s3 at t = 3 may contain returnees to the sample: sample persons at t = 1 who were non-respondents at t = 2, but were again enumerated at t = 3. If they also continue to be present at t = 4, they need to be included in the longitudinal sample being considered (since, as defined, they belong to the corresponding longitudinal population).

On return at t = 3 after the absence, these persons receive a zero base weight. These units need to be given a positive weight if they are to be included in the analysis. With the procedure described below, we can assign a non-zero weight to returnees not present in year (2) but again present in year (3). This requires, in compensation, a (slight) deflation of the base weights of the units in the longitudinal sample, from ${}^{(4)}_{3}\mathrm{w}$ to say ${}^{(4)}_{3}\tilde{\mathrm{w}}$, using the procedure indicated below.

The longitudinal sample of interest is not affected by any returnees between t = 3 and 4, so that the above can be used to estimate the revised weights:

$$ {}^{(4)}_{4}\tilde{\mathrm{w}} = {}^{(4)}_{4}\mathrm{w}.\left\{ \frac{{}^{(4)}_{3}\tilde{\mathrm{w}}}{{}^{(4)}_{3}\mathrm{w}} \right\}. \tag{21} $$

With this modification we can write the required expression similar to (19). The procedure is a little more involved algebraically, but the result is in the same form:

$$ \left[\left\{ s_3 - (\mathrm{out}_4 + x_4 + \mathrm{co}_3) + \mathrm{co}_{34} \right\}, {}^{(4)}_{4}\tilde{\mathrm{w}} \right] \longrightarrow \left(\mathrm{P}_3 - \mathrm{OUT}_4 \right) - \left[\mathrm{IN}_2^{\mathrm{new}} + \mathrm{IN}_3^{\mathrm{new}} \right]. $$

(22)

The left side is the longitudinal sample between times t = 3 and t = 4, i.e. individuals who are present at both waves, irrespective of whether they are from the original sample (t = 1) or were present at t = 2.

The first term on the right is the longitudinal population between t = 3 and t = 4. It is represented by the corresponding longitudinal sample with weights ${}^{(4)}_{4}\tilde{\mathrm{w}}$, except for entrants $\mathrm{IN}_2^{\mathrm{new}}$ and $\mathrm{IN}_3^{\mathrm{new}}$ forming entirely new households during the two years preceding t = 3.

Procedure for weight adjustment to incorporate returnees

Consider a longitudinal population N and the corresponding longitudinal sample n over a three-year period (1)-(3). For simplicity, let us only consider persons who remain members of N throughout the period. Let n1, n2 and n3 = n+r be the achieved samples in years 1-3, where r are the returnees, i.e. persons in n1 who were not enumerated in n2 at year (2), but were enumerated in n3 at year 3. Starting with weights w1 at t = 1, base weights for the longitudinal sample can be constructed as described in Section 4.2. We may write the base weights w2 at t = 2 as:

$$w_2 = w_1 / P(n_2 \mid n_1),$$

where the denominator stands for the propensity of a sample unit at t = 1 to remain in the sample at t = 2. Similarly:

$$w_3 = w_2 / P(n \mid n_2).$$

Using a similar procedure, we can estimate the propensities of units in n3 to be present in n. The adjustment (deflation) of the base weights of units in n allows for the incorporation of the returnees r = (n3 - n) into the sample with non-zero weights. The adjusted weights are:

$$\tilde{w}_3 = w_3 P(n \mid n_3) = \frac{w_1 P(n \mid n_3)}{P(n_2 \mid n_1) P(n \mid n_2)}, \text{ for units in longitudinal sample n,} \quad (23)$$

$$\tilde{w}_3 = \frac{w_1}{P(n_3 \mid n_1)}, \text{ for returnees r = (n3 - n).} \quad (24)$$

Incorporating longitudinal co-residents

As noted, it is desirable to incorporate longitudinal co-residents into longitudinal analysis. By these we mean co-residents present throughout the longitudinal interval of interest. Since, henceforth, such persons are zero weighted, they need to be assigned a non-zero longitudinal weight. The weight-share procedure described in Section 6.3 can be used for this purpose.

No modification need to be made to the weights of members in a household which does not contain any longitudinal co-residents. Otherwise, in a household containing one or more longitudinal co-residents, we compute the mean of longitudinal weights as the sum of base weights of longitudinal sample persons divided by the number of all longitudinal persons in the household, including longitudinal co-residents. This uniform weight is then assigned to all longitudinal household members, including longitudinal co-residents.

5.5. Construction of the target variables

We return to the three longitudinal data sets ${}^{Y}_{L2}S$, ${}^{Y}_{L3}S$ and ${}^{Y}_{L4}S$ for which the sample weights are required. The construction of those weights simply involves putting together the constituent panels. The components and the weight variables for the data sets are shown in Table 1. Generally, new entrants are represented in some but not all the constituent panels. To compensate for the missing ones, the weight of the included new elements can be appropriately inflated as follows. Let $W_{(A2)}$ be the sum of weights $\sum_{j \in A2} {}_2^{(2)}w_j$ excluding entrants $IN_{Y-2}^{(new)}$ and $IN_{Y-1}^{(new)}$. We define similar quantities for other components, e.g. $W_{(C2)}$ as the total of ${}_4^{(4)}\tilde{w}_j$, $j \in C2$.

Since population $IN_{Y-2}^{(new)}$ is represented only in panels (2) and (3), we inflate the weight of each element in category $in_{Y-2}^{(new)}$ by the factor:

$$\frac{W_{(A2)} + W_{(B2)} + W_{(C2)}}{W_{(A2)} + W_{(B2)}} \tag{25}$$

to compensate for its absence in component C2. With equal weighted panel sizes, this factor becomes 3/2. In accordance with Section 3.5, the weights in each panel are assumed scaled to average 1.0, so that the sum of weights is equal to the (unweighted) sample size.

Table 1. Coverage of new entrants in the panels comprising the longitudinal samples

Panel: year introduced	Panel duration at Y	Units	Base weights *	Whether entrants represented $\mathrm{IN}_{Y-2}^{(new)}$	$\mathrm{IN}_{Y-1}^{(new)}$	$\mathrm{IN}_{Y}^{(new)}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Longitudinal data-set of 2 year duration becoming available after year Y: ${}_{L2}^{Y}S$ = A2 + B2 + C2						
Y-1	(2)	$j \in A2$	${}_{2}^{(2)}w$	√	√	-
Y-2	(3)	$j \in B2$	${}_{3}^{(3)}w$	√	x	-
Y-3	(4)	$j \in C2$	${}_{4}^{(4)}\tilde{w}$	x	x	-
Longitudinal data-set of 3 year duration becoming available after year Y: ${}_{L3}^{Y}S$ = A3 + B3						
Y-2	(3)	$j \in A3$	${}_{3}^{(3)}w$	√	-	-
Y-3	(4)	$j \in B3$	${}_{4}^{(4)}w$	x	-	-
Longitudinal data-set of 4 year duration becoming available after year Y: ${}_{L4}^{Y}S$ = A4						
Y-3	(4)	$j \in A4$	${}_{4}^{(4)}w$	-	-	-
√ = YES, x = NO, - not relevant						

* for panel (4), modified base weights are defined in the text.

Similarly, the weight of each element in category $\mathrm{in}_{Y-1}^{(new)}$ is multiplied by the factor:

$$\frac{W_{(A2)} + W_{(B2)} + W_{(C2)}}{W_{(A2)}}.$$

With equal weighted panel sizes, this factor equals 3.

With these modifications for entrants, weights in col. (4) are the required longitudinal weight variables, expect for the following final adjustments.

(1) Scaling the weights to average 1.0 for each panel separately automatically ensures that the contribution of each panel is proportional to its sample size. Also, the average weight remains 1.0 or close to it when the panels are merged to form the full sample. However, sometimes weights are scaled to sum to the populations' size (or to equal inverse of the selection probabilities). In this case the weights in the individual panels have to be proportionally reduced to obtain the full sample weights. Ideally, before putting the panels together, these weights should be re-scaled such that for each panel the sum of weights is proportional to the panel sample size.

(2) For multi-country analysis using combined data, it is desirable to scale the weights such that their sum is proportional to the longitudinal population size of the country concerned. When (as is often the case), size of the

longitudinal population is not available, the sum of weights may be scaled to be proportional to the population at a certain time such as Y.

(3) In accordance with Section 3.4, each panel is assumed calibrated against external controls at the time of its entry into the sample. This is the most important application of calibration. Additional or repeated calibration at subsequent times is often not worthwhile. In any case, in many situations information on the longitudinal population is not available for the purpose.

(4) For units in any subpopulation such as children (Q), we can take the longitudinal weight to be identical to the units' sample weight in the total population (R): $v_j^{(Q)} = v_j^{(R)}$.

(5) After successful enumeration of the population (R) within households, additional non-response may be involved in the subsequent personal interview survey with adults (P). Adjustment to the weights is required to compensate for this within-household non-response. A simple approach is to calibrate the achieved P-sample to match the corresponding R-sample over a set of important characteristics. The raking procedure described in Section 3.3 can, for instance, be used for the purpose. We begin with the above determined R-weights applied to the achieved P-sample. Then the weights are adjusted (to give the required P-weights) such that P-sample, with these adjusted weights, agrees with R-sample (with R-weights) on the specified control characteristics. The above calibration corrects for overall distortions in the P-sample due to within-household non-response. Further correction needs to be applied to the data obtained for the particular households affected by within-household non-response, as described in Section 7.4.

(6) In countries where P-data are obtained from registers there is generally no within-household non-response of the above type. In such cases, we can take the longitudinal P-weight of a unit (in the P-sample) to be identical to the units' weight in the R-sample: $v_j^{(P,L)} = v_j^{(R,L)}$. In these countries personal interview is conducted over a sample of selected respondents (S), normally one adult per sample household. Here the S-weights firstly need to take into account the differences in the design weights between the two samples:

$$v_j^{(S)} = \left[\frac{{}_1w_j^{(S,D)}}{{}_1w_j^{(R,D)}} \right] .v_j^{(R)} . \tag{26}$$

Next, to control for the effect of sampling within households, we may calibrate the S-sample to match the corresponding R-sample over a set of characteristics, using for instance the raking procedure of Section 3.3. For this purpose, we can begin with the above defined S-weights applied to the S-sample, and adjust these weights to achieve agreement with the full R-sample (weighted by R-weights) on specified control characteristics.

(7) As to whether such calibration is also required to compensate for S-sample attrition depends on the follow-up rules and the field procedures followed (see Section 7.3 concerning the former). Two types of procedures can be envisaged:

(a) One option is to determine the follow-up of the R-sample entirely on the basis of the achieved S-sample – i.e. follow-up household and the associated household members only for successfully interviewed selected respondents (S-sample). In this case, there is no relative non-response between the two samples, and weights derived in (6) require no further adjustment.

(b) In our view, the above procedure is unnecessarily damaging to the quality of the income and other data collected on the R-sample from registers and other sources not themselves subject to non-response. This arises from imposing on the R-sample the interview non-response to which the S-sample is subject. The alternative is to follow-up the full R-sample independently of the outcome of the S-sample. Under this scenario, one should construct the whole set of base, longitudinal and cross-sectional weights described in Sections 4-6 for the S-sample in its own right.

5.6. Start-up of the integrated design

Figure 2 also shows how the integrated design may be started up. For a 4 years panel design, we may begin with 4 new panels in year 1, drop one of these and introduce a new one in year 2. The whole longitudinal sample, consisting of three panels, is of type A2, and the required longitudinal weights are given directly by the base weights as described in Section 5.3.

In year 3, we drop another of the original panels and introduce a new one. The longitudinal sample of 2-year duration consists of one panel of type A2 and two panels of type B2, say B2(1) and B2(2). Part $\mathrm{IN}_{Y-2}^{(new)}$ is represented in all these panels and no reweighting is required. The reweighting for part $\mathrm{IN}_{Y-1}^{(new)}$ is similar to that in the 'normal' case discussed in Section 5.5. The weight of any unit in this part is multiplied by:

$$\frac{W_{(A2)} + W_{(B2(1))} + W_{(B2(2))}}{W_{(A2)}}. \qquad (27)$$

From year 4 onwards the sample structure becomes normal for the integrated design.

6. Cross-sectional weights

6.1. Special aspects of cross-sectional weighting

Some special features of cross-sectional weights include the following.

- In the case of EU-SILC at least, the first objective is to obtain good quality cross-sectional estimates. This enhances the importance of constructing good cross-sectional weights: by ‘good’ we mean weights chosen to reduce mean-square-error.
- For the sake of efficiency, cross-sectional analysis should use all available sample cases at the time concerned. This requires that all units enumerated at that time, irrespective of whether or not they are original sample persons, be given non-zero positive weights.
- It is more convenient and consistent to assign a uniform cross-sectional weight to all current members of a household, the same as the weight of the household, rather than to allow the weights to vary across persons in the same household.
- It is both more desirable and more feasible to calibrate the sample weights of the cross-sectional sample each wave on external control distributions on relevant population characteristics, compared to the longitudinal samples. Much more information is usually available for cross-sectional calibration than for longitudinal calibration.

In the rotational integrated design, cross-sectional weights can be constructed from the base weights (Section 4) in a straightforward way, on lines similar to the construction of longitudinal weights in Section 5. Many of the details of the procedure are similar for the two types of weighting, and this section can therefore be brief. The full cross-sectional sample is constructed by putting together results from different panels for the same year (Y), as shown in Table 2 and illustrated in Figure 2.

Table 2. Coverage of new entrants in the panels comprising the cross-sectional sample

Panel: year introduced	Panel duration at Y	Units	Base weights*	Whether entrants represented		
				$IN^{(new)}_{Y-2}$	$IN^{(new)}_{Y-1}$	$IN^{(new)}_{Y}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Y	(1)	$j \in {}^{Y}_{1}s$	${}^{(1)}_{1}w$	√	√	√
Y-1	(2)	$j \in {}^{Y-1}_{2}s$	${}^{(2)}_{2}w$	√	√	x
Y-2	(3)	$j \in {}^{Y-2}_{3}s$	${}^{(3)}_{3}\tilde{w}$	√	x	x
Y-3	(4)	$j \in {}^{Y-3}_{4}s$	${}^{(4)}_{4}\tilde{\tilde{w}}$	x	x	x
√ = YES, x = NO						

* for panels (3) and (4), modified base weights are defined in the text.

6.2. Adjustment of the base weights

Base weights form the basis of the required cross-sectional weights. With extension of the base weights to additional categories of persons (Sections 4.4 and 5.2), further adjustment is still required to the incorporation of returnees into the sample with non-zero weights. With the procedure described in Section 5.4, the weights can be further adjusted for returnees as follows.

We have already computed the adjusted weights ${}^{(4)}_{3}\tilde{w}$, referring to panel (4) in its 3rd year. Exactly in the same way, we can estimate ${}^{(3)}_{3}\tilde{w}$, which refers to the 3rd year of panel (3). For panel (4), it is necessary to adjust the weights for (i) returnees between its 2nd and 3rd years, and then further for (ii) returnees between its 3rd and 4th years. An estimate of (i) is as before (equation (21)):

$$ {}^{(4)}_{4}\tilde{w} = {}^{(4)}_{4}w.\left\{\frac{{}^{(4)}_{3}\tilde{w}}{{}^{(4)}_{3}w}\right\}. $$

For the further adjustment (ii), the procedure is the same as that in Section 5.4. We estimate the propensity P of individuals in ${}^{(4)}_{4}s$ at wave 4 to also have been enumerated at wave 3 as $P = \left({}^{(4)}_{3}s \cap {}^{(4)}_{4}s\right) / \left({}^{(4)}_{4}s\right)$, giving ${}^{(4)}_{4}\tilde{\tilde{w}} = P.\left({}^{(4)}_{4}\tilde{w}\right)$.

The above are the modified base weights for persons other than returnees. For the returnees in the 4th year who were present in year 2 but not in year 3, we may use the slightly approximate expression[1]:

[1] We have assumed that individuals missed in only one wave may return to the sample; those missed in two or more waves consecutively are not allowed to return to the sample. This in fact is in line with the follow-up rules adopted in EU-SILC. The procedure, of course, can be extended in a straightforward way to remove the above limitation.

$$P = \left({}^{(4)}_{2}s \cap {}^{(4)}_{4}s\right) / \left({}^{(4)}_{2}s\right), \text{ giving } {}^{(4)}_{4}\tilde{\tilde{w}} = {}^{(4)}_{2}w / P.$$

6.3. Construction of cross-sectional weights

Coverage of in-migrants

New entrants into the population during the three years preceding Y are represented only in some of the constituent panels. To compensate for the missing ones, base weights of new entrants in the panel(s) where they have been included are inflated as follows.

Let $W_{(S1)}$ be the sum of base weights $\sum\left\{{}^{(1)}_{1}w_j\right\}$, $j \in {}^{1}_{1}s$, similarly $W_{(S2)} = \sum\left\{{}^{(2)}_{2}w_j\right\}$, $W_{(S3)} = \sum\left\{{}^{(3)}_{3}\tilde{w}_j\right\}$, $W_{(S4)} = \sum\left\{{}^{(4)}_{4}\tilde{\tilde{w}}_j\right\}$, in all cases excluding new entrants $in^{(new)}$.

Since population $IN^{(new)}_{Y-2}$ is not represented in panel (4), we inflate the weight of each element in $in^{(new)}_{Y-2}$ category in every of the panel (1)-(3) by the factor:

$$\frac{W_{(S1)} + W_{(S2)} + W_{(S3)} + W_{(S4)}}{W_{(S1)} + W_{(S2)} + W_{(S3)}}. \tag{28}$$

When the weighted panel sizes are equal, this factor equals 4/3.[1]

Similarly, weights of present units in category $in^{(new)}_{Y-1}$ are multiplied by the factor:

$$\frac{W_{(S1)} + W_{(S2)} + W_{(S3)} + W_{(S4)}}{W_{(S1)} + W_{(S2)}}, \tag{29}$$

which equals 2 when the weighted panel sizes are equal.

For units in category $in^{(new)}_{Y}$, the unit weights are inflated by the factor:

$$\frac{W_{(S1)} + W_{(S2)} + W_{(S3)} + W_{(S4)}}{W_{(S1)}}, \tag{30}$$

which equals 4 when the weighted panel sizes are equal.

Weight sharing

Within a household, j∈h, each member has been assigned a weight w_j, except for co-residents for whom w_j = 0. The average of these weights over all household members (including co-residents) is assigned to each member (including co-residents). We recommend applying *this averaging process to all households, including households not containing a co-resident.* This procedure

[1] As was noted earlier, in putting different panels together, it is best to scale the weights in each panel separately to average 1.0. In this case the sum of weights is simply the panel sample size.

ensures that non-zero weights are assigned to all persons enumerated at the cross-section concerned, and that within each household, all individuals enumerated have the same weight. The result is to assign the same uniform weight to all current members of a household. The sum of base weights (after any adjustment as above) remains unchanged for the household: this sum is simply shared among all current members.

In brief outline the weight-sharing procedure is as follows (Ernst, 1989). If u_k is a random variable associated with each unit k in the population of size N, then $\hat{X} = \sum_{k=1}^{N} x_k . u_k$ provides an unbiased estimator of the population total $X = \sum_{k=1}^{N} x_k$, provided $E(u_k) = 1$. In the conventional Horvitz-Thompson estimator, we take u_k as inverse of the selection probability of the sample unit, and take $u_k = 0$ for all units not in the sample. The weight-share approach is based on the observation that an unbiased estimate is also produced by defining a new set of weights w_j for all units (j) in the population in the following form:

$$w_j = \sum_{k=1}^{N} a_{jk} . u_k \text{, with } \sum_{k=1}^{N} a_{jk} = 1 \text{,}$$

where the constants a_{jk} are independent of u_k, but are otherwise determined according to freely chosen rules. This is because with the above definitions we have $E(w_j) = \sum_{k=1}^{N} a_{jk} . E(u_k) = \sum_{k=1}^{N} a_{jk} = 1$, thus satisfying the above requirement for the estimator to be unbiased. This form involves redefining the weight of each unit (j) in the population as a (weighted) average of the weights u_k for units in a certain set determined according to some specified rules. For instance, for each individual (j) in household h in the population, the set may refer to all members in the individual's household. The factors a_{jk} corresponding to each of the individual's household member may be taken as equal to $1/n_h$ (where n_h is the household size), and as equal to 0 corresponding to all persons in the population who are not members of (j)'s household. With these definitions the redefined weights w_j simply become the average of the original weights of individuals in the set (household) of (j).

Calibration

As noted, there is a greater need and possibility of calibration on external control distributions on relevant population characteristics in the cross-sectional compared to a longitudinal context. As a rule, it is recommended to calibrate the cross-sectional sample weights each wave. Again, integrative calibration will ensure that uniformity of the person weights within each household is retained. Calibration should normally be applied to the whole, pooled sample; separate calibration of individual panels is generally not necessary and often not even desirable.

Trimming

Following any operation like calibration, the resulting weights should always be checked for extreme values, followed by trimming or other appropriate treatment as necessary.

Scaling

The basic principles noted in Section 5.5 apply. Scaling the weights to average 1.0 for each panel separately, automatically ensures that the contribution of each panel is proportional to its sample size. For pooling across countries or other domains, it is desirable to scale the weights such that their sum is proportional to the population size.

Weights for other types of units

With weight sharing (and integrative calibration, uniform scaling etc., as applicable), the equality of individual (R) and household (H) weights can be retained. For units in any subpopulation such as children (Q), we can also take the cross-sectional weight to be identical to the units' cross-sectional weight in the total population (R): $v_{j\in h}^{(Q)} = v_{j\in h}^{(R)} = v_{h}^{(H)}$.[1]

For data from the personal interview survey with adults (P), adjustment is required to compensate for any additional non-response at that stage. The simple approach is to calibrate the achieved P-sample to match the corresponding R-sample over a set of important characteristics, described in Section 5.5. Further correction needs to be applied to the data obtained for the particular households affected by within-household non-response, as noted in Section 7.4.

In countries where the personal interview is conducted over a sample of selected respondents (S), the S-weights firstly need to take into account the differences in the design weights. Next, to control for the effect of sampling within households, we may calibrate the S-sample to match the corresponding R-sample over a set of characteristics (Section 3.3; see also Section 7.3).

6.4. Start-up of the integrated design

We refer again to the start-up of the integrated design in Figure 2. All panels in year 1 are freshly introduced. The required cross-sectional weights are given directly by the initial or base weights for year 1. Panels can be put together to construct the full cross-sectional sample, without further rescaling if the weights of each panel have been scaled to average 1.0.

[1] It has been suggested that, for special purposes, the weights for certain subpopulations (Q) may be subject to further, more precise calibration. See, for instance, Eurostat (2004) in relation to child-care data.

In year 2, category $\mathrm{IN}_2^{(\mathrm{new})}$ is represented in only one of the four panels (say with sum $W_{(S1)}$), and base weights of the cases belonging to it have to be inflated as in equation (30). No other adjustment is required.

In year 3, category $\mathrm{IN}_3^{(\mathrm{new})}$ is represented in only one of the four panels, and $\mathrm{IN}_2^{(\mathrm{new})}$ in two of the four. The weights of sample units in these categories have to be inflated, respectively, by expressions similar to equations (30) and (29). As in the longitudinal case, from year 4 onwards the cross-sectional sample structure becomes normal for the integrated design.

7. Some variations and practical aspects

In this concluding section, we note some further details and practical issues.

7.1.Information on migration prior to entering the survey

A precise application of longitudinal and cross-sectional weighting procedures (Sections 5 and 6) requires the identification of in-migrants into the target population of private households in the country (from abroad or from the institutional sector).

Information is also required as to whether an in-migrant formed a ‘new’ household, or moved into an already existing (‘old’) household containing a person from the existing population. Information on migration status is also required for co-residents who move into sample households at subsequent waves. The above requires the following items to be specified.

(1) For *all persons* enumerated in the household at the first wave: whether they entered the target population within the past 3 years, and if so when.

(2) At each subsequent wave, for each new co-resident (a non-sample person moving into the household that wave): whether they have moved from another household in the target population, or from outside that population (another country or an institution).[1]

Let $\mathrm{in}_{Y-t+1}^{(\mathrm{new})} = 1$ indicate an in-immigrant during the interval (Y-t) and (Y-t+1) into a ‘new’ household as defined above (this indicator is equal to 0 for all other persons). From Section 6, it can be seen that for the panel introduced at Y, this indicator is required for each person for the preceding three years (Y-t+1), t = 1, 2, 3. This can constructed from question (1) above as follows.

[1] In principle, the same set of information is required for returnees for their proper statistical treatment – see footnote (11) in section 5.2.

- If any current member of the household entering the survey has been a member of the target population for $t' > 3$ years, then the indicator $in_{Y-t+1}^{(new)} = 0$ for all members, all t = 1, 2, 3.
- Otherwise, let $t' \leq 3$ identify the interval for the set of current household members who are the earliest to have moved into the target population. Then $in_{Y-t'+1}^{(new)} = 1$ for each person in the above-mentioned set. For other $t \neq t'$ for this set, and for all t for any remaining member, this indicator equals 0.

Question (2) identifies co-residents entering the sample household from outside the target population at any time from wave 2 onwards. For persons who have moved into the household from outside the target population, index $in^{(old)} = 1$ for the waves they remain in the sample. As noted earlier, these persons are assigned a non-zero base weight (Section 5.3). By contrast, co-residents moving in from within the target population are treated differently. They originally have zero base weight. The longitudinal co-residents among them receive non-zero longitudinal weights, and all of them receive non-zero cross sectional weights, on the basis of weight sharing (Sections 5.4, 6.3).

7.2. Variations from the integrated design

A vast proportion of the countries conducting EU-SILC have used the integrated design described in Section 1.4. The main departures arise from the need to combine EU-SILC with an existing survey or sample (Finland, Germany, Norway). France uses the same structure as the integrated design, but a panel duration of 9 (in place of 4) years. (The panel duration is also longer – 8 years – in Norway). Luxembourg uses a panel of indefinite duration similar to the pure panel ECHP design, but complements this with annual samples to compensate for panel attrition (Clemenceau et al., 2006).

When the design departs from the standard, the weighted procedures described in this paper of course require adaptation. However, the basic approach, and many details as well, remain valid and useful for developing alternative procedures. When panels of long or indefinite duration are used, special care is needed to retain cross-sectional representativity of the sample. More careful weighting for panel attrition, and possibly also periodic supplementation of the sample, become important.

7.3. On follow-up rules in register countries

The analysis of income distribution, inequality and poverty – which is the primary aim of a survey like EU-SILC – is normally carried out with individual person as unit of analysis.

Longitudinal analysis, such as of persistent poverty, therefore requires that all persons in the original sample, irrespective of age or other characteristics, be followed-up in subsequent waves of the panel. As noted earlier (Section 1.5), restricting the follow-up to persons over a certain age, such as 14, normally does not have a major effect since only children under the specified age who move alone (that is, without being accompanied by an adult in the original household) are not followed-up. Usually such cases are rare and can be treated as non-response.

In so-called register countries, however, a more serious issue needs to be considered. Here we are dealing with two populations:

(i) all members of each sample household to be covered in the measurement of income and related characteristics, which can be obtained from registers or other sources not requiring detailed personal interviews; and

(ii) selected respondents, typically one per household, who are interviewed in detail for non-income or 'social' variables.

Which of these two populations – all persons or selected respondents – constitute the 'sample persons' to be followed-up in subsequent waves?

It is important to note that while for longitudinal analysis of variables covered under (ii) it is sufficient to follow-up selected respondents, that is not adequate for the longitudinal analysis of income and poverty.

Following-up only the selected respondents would mean that a representative sample of the whole population, especially of children, is not maintained. The follow-up rules for (i) have to be more inclusive.

7.4. Within-household non-response

By this is meant the failure to obtain the detailed interview with eligible individuals even when their household has been successfully enumerated. As a rule, this problem arises only in survey countries.

In Sections 5.5 and 6.4 we noted that in practice, the incidence of such non-response tends to be quite low – once the household interview has been successfully completed, the personal interview also tend to be obtained. The modest effect of within-household non-response on the composition of the sample can be controlled by calibrating the interviewed over the eligible sample for the personal interview in terms of age-sex and other characteristics.

However, the impact of such non-response on the particular households affected is generally much more important, and is not ameliorated by a calibration such as the above at the aggregate level.

Correction is required at the micro-level to the data of the households affected. This is because income and other characteristics at the household level have to be constructed by taking into account the contribution of all members of the household.

When the incidence of within-household non-response is low, the missing personal interviews may be imputed altogether. However, usually it is more appropriate and convenient to introduce a special weighting factor for the household concerned as a compensation for the information missing as a result of within-household non-response. We have not considered the construction of such special weighting factors in this paper, but for an example see Eurostat (2001).

Acknowledgements

The weighting procedures described here are based on the authors' detailed recommendations developed under a project with Eurostat, and we wish to thank Eurostat for providing the opportunity for their development. In fact, those recommendations are being implemented in EU-SILC national surveys. The present paper aims at providing a more systematic and clearer description of the weighting procedures, also introducing refinements to our original recommendations so as to enhance their consistency and completeness.

REFERENCES

ARDILLY, P., LAVALLÉE, P. (2003), The weight share method and the European Survey on Income and Living Conditions. *Proceedings Symposium 2003, Challenges in survey taking for the next decade*. Ottawa: Statistics Canada.

CLEMENCEAU, A., MUSEAUX, J-M., BAUER, M. (2006), EU-SILC: issues and challenges. Presented at conference *Comparative EU Statistics on Income and Living Conditions: Issues and Challenges*, Helsinki.

DEMING, W. E. (1943), *Statistical Adjustment of Data,* Chapter VII, Wiley (Paperback Dover 1964).

DEVILLE, J-C., LAVALLÉE, P. (2006), Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, **32**(2), pp. 165—176.

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

ERNST, L. (1989), Weighting issues for longitudinal household and family estimates. In *Panel Surveys* (eds. Kasprzyk D. *et al*.), pp. 135—159. Wiley.

EUROPEAN COMMUNITY (2003), Regulation (EC) No 1177/2003 of the European Parliament and of the Council. *Official Journal of the European Union*, pp. L165/1—9.

EUROSTAT (2001), ECHP UDB Construction of Variables, doc.PAN 167/01.

EUROSTAT (2004), First ideas on weighting for child care data, EU-SILC doc 135/04.

KALTON, G., BRICK, J. M. (1995), Weighting schemes for household panel surveys. *Survey Methodology*, **21**(1), pp. 33—44.

KALTON, G., KASPRZYK, D. (1986), The treatment of missing survey data. *Survey Methodology*, **12**, pp. 1—16.

KISH, L. (1992), Weighting for unequal P_i. *Journal of Official Statistics*, **8**, pp. 183—200.

LAVALLÉE, P. (1995), Cross sectional weighting for longitudinal survey of individuals and household using the weight share method. *Survey Methodology*, **21**(1), pp. 25—32.

LAVALLÉE, P., CARON, P. (2001), Estimation using the generalized weight share method: the case of record linkage. *Survey Methodology*, **27**(2), pp. 155—168.

LITTLE, R. J. A. (1986), Survey non-response adjustment for estimates of means. *International Statistical Review*, **54**, pp. 139—157.

POTTER, F. J. (1988), Survey of procedures to contrast extreme sampling weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 453—458.

POTTER, F. J. (1990), A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 225—230.

SPENCER, B., COHEN, T. (1991), Shrinkage weights for unequal probability samples. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 625—630.

VERMA, V. (1999), Combining national surveys for the European Union. *Bulletin of the International Statistical Institute*, **58**.

VERMA, V. (2001), *EU-SILC Sampling Guidelines*. EU-SILC doc 51/01. Luxembourg: Eurostat.

VERMA, V. (2006), *EU-SILC Weighting Procedures: an outline*. Report to Eurostat on Advanced Estimation Methods project.

VERMA, V., BETTI, G. (2006), EU Statistics on Income and Living Conditions (EU-SILC): Choosing the survey structure and sample design. *Statistics in Transition*, **7**(5), pp. 935—970.

VERMA, V., CLEMENCEAU, A. (1996), Methodology of the European Community Household Panel. *Statistics in Transition*, **2**(7), pp. 1023—1062.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 51—69

IMPROVED RATIO-CUM-PRODUCT TYPE ESTIMATORS

Pier Francesco Perri[1]

ABSTRACT

In this paper an improved version of Singh's *ratio-cum-product* estimators is suggested in simple random sampling when two auxiliary variables are available. Using Taylor linearization method we obtain, to the first and second order of approximation, the expressions for the mean square error of the proposed estimators and we establish that they are more efficient than the original ratio-cum-product estimators. Moreover, in order to better evaluate the performance of the estimators, an application to real data is shown.

Key words: Auxiliary Variable, Efficiency, Ratio Estimator, Product Estimator, Second Order Approximation.

1. Introduction

In sampling practice it is usual to make use of information on auxiliary variables to obtain improved designs and more efficient estimators. This information may be used at the planning stage of the survey, in the estimation procedure, or at both phases. The literature on survey sampling describes a great variety of techniques for using auxiliary information (see, e.g., Cochran, 1977; Murthy, 1977). In particular, it is well known that when the auxiliary information is used at the estimation stage, the *ratio*, *product* and *regression* estimators are widely utilized in many situations. Theoretically, it has been established that the regression estimator is more efficient than the ratio and product estimators except when the regression line of the character under study on the auxiliary variable passes by the origin. In this case, the efficiency of the estimators is almost equal. However, due to the stronger intuitive appeal, statisticians are more inclined towards the use of ratio and product estimators. Perhaps, that is why an extensive

[1] Department of Economics and Statistics, University of Calabria, Via Pietro Bucci, 87036 Arcavacata di Rende (CS), Italy. E-mail: pierfrancesco.perri@unical.it.

work has been conducted in the direction of improving the performance of these estimators.

In sampling literature, when a single auxiliary variable is available, many estimators have been proposed which, under some realistic conditions, are more efficient than the sample mean, the ratio and product estimators and are as efficient as the regression estimator in the optimum case. This paper is concerned with the problem of estimating the population mean of a survey variable (Y) using two auxiliary variables (X_1 and X_2).

When two or more auxiliary variables are available, many estimators may be defined by linking together different estimators such as ratio, product or regression, each one exploiting a single variable. Olkin (1958) suggested the use of information on more than one supplementary characteristics, positively correlated with the variable under study, using a linear combination of ratio estimators based on each auxiliary variable separately. The coefficients of the linear combination were determined so as to minimize the variance of the estimator. Singh (1967a) developed a multivariate expression of product estimator similar to Olkin's *multivariate ratio estimator*, while Raj (1965) suggested a method of using multi-auxiliary variables considering a linear combination of single *difference estimators*. Singh (1965, 1967b) and Rao and Mudholkar (1967) proposed multivariate estimators for the unknown mean of a finite population combining ratio and product estimators.

Many other contributions are present in sampling literature and, recently, some new estimators appeared. Here we mention, among others, Naik and Gupta (1991), Ceccon and Diana (1996), Abu-Dayyeh et al. (2003), Kadilar and Cingi (2004, 2005). Motivated by these recent proposals, in this paper we develop, when two auxiliary variables are available, some new estimators obtained from the *ratio-cum-product* estimators defined by Singh (1965, 1967b).

The paper is organized as follows: Section 2 introduces notation and ratio-cum-product estimators. In Section 3, we present four new estimators and obtain, up to the first degree of approximation, the approximate expressions for their bias and mean square error. Section 4 is devoted to the analysis of the efficiency of the proposed estimators. In Section 5, the expressions of second order mean square error are given. An empirical analysis of the second order mean square error is presented in Section 6, together with some concluding remarks.

2. Ratio-cum-Product Estimators

Let $P = \{1, 2, ..., N\}$ be a finite population, Y the study variable and X_1 and X_2 two auxiliary variables assuming, respectively, values Y_j, X_{1j} and X_{2j} for the j-th population unit, j=1,2,...,N. Let $\overline{Y} = N^{-1}\sum_{j=1}^{N} Y_j$ be the unknown mean of the study variable and suppose that the population means of the auxiliary variables, $\bar{X}_1 = N^{-1}\sum_{j=1}^{N} X_{1j}$ and $\bar{X}_2 = N^{-1}\sum_{j=1}^{N} X_{2j}$, are known. Let a sample of size n be drawn from the population according to simple random sampling without replacement (srswor) and let $\bar{y}$, $\bar{x}_1$ and $\bar{x}_2$ be the sample means of the variables Y , X_1 and X_2 .

To estimate the unknown $\bar{Y}$, Singh (1965, 1967b) introduced the estimators

$$\hat{\bar{Y}}_{R1} = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}\frac{\bar{x}_2}{\bar{X}_2},\quad \hat{\bar{Y}}_{R2} = \bar{y}\frac{\bar{X}_1}{\bar{x}_1}\frac{\bar{X}_2}{\bar{x}_2},\quad \hat{\bar{Y}}_{P1} = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}\frac{\bar{x}_2}{\bar{X}_2},\quad \hat{\bar{Y}}_{P2} = \bar{y}\frac{\bar{x}_1}{\bar{X}_1}\frac{\bar{X}_2}{\bar{x}_2} \qquad (1)$$

named ratio-cum-product since they are constructed combining ratio and product estimators based on each auxiliary variable. It is known that the estimators are biased, but the bias may be considered negligible for large samples. Moreover, under suitable conditions involving the correlation coefficients between the variables, it is easy to prove (Singh, 1965) that they may be more efficient than the ratio, $\hat{\bar{Y}}_r = \bar{y}\,\bar{X}_i/\bar{x}_i$, and product, $\hat{\bar{Y}}_p = \bar{y}\,\bar{x}_i/\bar{X}_i$, estimators which make use of a single auxiliary variable.

In spite of many efforts conducted in the direction of improving ratio and product estimators based on a single auxiliary variable, few attempts have been made to improve the efficiency of the ratio-cum-product estimators. Sahoo and Swain (1980) proposed a Hartley-Ross type unbiased ratio-cum-product estimator and derived the conditions under which it is asymptotically more efficient than the biased $\hat{\bar{Y}}_{R1}$. Sahoo (1991) suggested an unbiased ratio-cum-product estimator in two-stage simple random sampling which reduces to the estimator of Sahoo and Swain (1980) in the case of uni-stage sampling. Tracy et al. (1996) adapted to $\hat{\bar{Y}}_{R1}$ the transformation suggested by Srivenkataramana and Tracy (1981), obtaining an estimator with exact expression for the bias and the mean square error.

Perri (2004), using the transformation suggested by Srivenkataramana (1980) and Bandyopadhyay (1980), proposed, for a generic sampling design, the modified ratio-cum-product estimators:

$$\hat{\bar{Y}}^*_{R1} = \hat{\bar{Y}}\frac{\bar{X}_1}{\hat{\bar{X}}^*_1}\frac{\hat{\bar{X}}^*_2}{\bar{X}_2},\quad \hat{\bar{Y}}^*_{R2} = \hat{\bar{Y}}\frac{\bar{X}_1}{\hat{\bar{X}}^*_1}\frac{\bar{X}_2}{\hat{\bar{X}}^*_2},\quad \hat{\bar{Y}}^*_{P1} = \hat{\bar{Y}}\frac{\hat{\bar{X}}^*_1}{\bar{X}_1}\frac{\hat{\bar{X}}^*_2}{\bar{X}_2},\quad \hat{\bar{Y}}^*_{P2} = \hat{\bar{Y}}\frac{\hat{\bar{X}}^*_1}{\bar{X}_1}\frac{\bar{X}_2}{\hat{\bar{X}}^*_2} \qquad (2)$$

with

$$\hat{\bar{X}}^*_i = \frac{N\bar{X}_i - n\hat{\bar{X}}_i}{N-n},\quad i = 1,2$$

where $\hat{\bar{Y}}$ and $\hat{\bar{X}}_i$ are unbiased estimators of $\bar{Y}$ and $\bar{X}_i$, defined on the sample drawn, while $\hat{\bar{X}}^*_i$ is also an unbiased estimator of $\bar{X}_i$ but it is based on the population units not sampled. Under the realistic hypothesis that the sampling fraction, $f = n/N$, is less than 0.5, the conditions that make the modified estimators more efficient than the original ones are obtained. In particular, after some algebra, it is easy to prove that:

- $\hat{\bar{Y}}^*_{R1}$ is more efficient than $\hat{\bar{Y}}^*_{P2}$ if $\dfrac{1+m}{2} > \dfrac{C_{101} - C_{110}}{C_{020} + C_{002} - 2C_{011}}$;
- $\hat{\bar{Y}}^*_{R2}$ is more efficient than $\hat{\bar{Y}}^*_{P1}$ if $\dfrac{1+m}{2} > -\dfrac{C_{101} + C_{110}}{C_{020} + C_{002} + 2C_{011}}$;
- $\hat{\bar{Y}}^*_{P1}$ is more efficient than $\hat{\bar{Y}}^*_{R2}$ if $\dfrac{1+m}{2} > \dfrac{C_{101} + C_{110}}{C_{020} + C_{002} + 2C_{011}}$;
- $\hat{\bar{Y}}^*_{P2}$ is more efficient than $\hat{\bar{Y}}^*_{R1}$ if $\dfrac{1+m}{2} > -\dfrac{C_{101} - C_{110}}{C_{020} + C_{002} - 2C_{011}}$,

where

$$m = \frac{f}{1-f}, \quad C_{rst} = \frac{1}{N-1} \frac{\sum_{j=1}^{N} (Y_j - \bar{Y})^r (X_{1j} - \bar{X}_1)^s (X_{2j} - \bar{X}_2)^t}{\bar{Y}^r \bar{X}_1^s \bar{X}_2^t}. \tag{3}$$

The above conditions may be easily adapted to the srswor. In this case, we find that:

- $\hat{\bar{Y}}^*_{R1}$ is more efficient than $\hat{\bar{Y}}^*_{P2}$ if $\dfrac{1+m}{2} > C_0 \dfrac{C_2\rho_{02} - C_1\rho_{01}}{C_1^2 + C_2^2 - 2C_1C_2\rho_{12}}$;
- $\hat{\bar{Y}}^*_{R2}$ is more efficient than $\hat{\bar{Y}}^*_{P1}$ if $\dfrac{1+m}{2} > -C_0 \dfrac{C_2\rho_{02} + C_1\rho_{01}}{C_1^2 + C_2^2 + 2C_1C_2\rho_{12}}$;
- $\hat{\bar{Y}}^*_{P1}$ is more efficient than $\hat{\bar{Y}}^*_{R2}$ if $\dfrac{1+m}{2} > C_0 \dfrac{C_2\rho_{02} + C_1\rho_{01}}{C_1^2 + C_2^2 + 2C_1C_2\rho_{12}}$;
- $\hat{\bar{Y}}^*_{P2}$ is more efficient than $\hat{\bar{Y}}^*_{R1}$ if $\dfrac{1+m}{2} > -C_0 \dfrac{C_2\rho_{02} - C_1\rho_{01}}{C_1^2 + C_2^2 - 2C_1C_2\rho_{12}}$,

where C_0, C_1 and C_2 denote the coefficients of variation of the variables Y, X_1 and X_2 and $\rho_{..}$ is the correlation coefficient between the corresponding variables.

For an optimal choice of the sampling fraction, $f_{opt} = m_{opt}/(1-m_{opt})$, the mean square error of the modified estimators may be minimized. Particularly, $\hat{\bar{Y}}^*_{R1}$ and $\hat{\bar{Y}}^*_{R2}$ attain their minimum mean square error, respectively, for

$$m_{opt} = \frac{C_{101} - C_{110}}{C_{020} + C_{002} - 2C_{011}} \text{ and } m_{opt} = -\frac{C_{101} + C_{110}}{C_{020} + C_{002} + 2C_{011}},$$

while $\hat{\bar{Y}}^*_{P1}$ and $\hat{\bar{Y}}^*_{P2}$ are optimum, respectively, for

$$m_{opt} = \frac{C_{101} + C_{110}}{C_{020} + C_{002} + 2C_{011}} \text{ and } m_{opt} = -\frac{C_{101} + C_{110}}{C_{020} + C_{002} - 2C_{011}}.$$

Adopting the optimum values of m_{opt} the above mentioned conditions are always satisfied and the modified estimators turn out to be always more efficient than the original ones.

Now, we intend to modify the structure of the original ratio-cum-product estimators with the aim of further improving the performance of the modified ones defined in (2). To do this, we will substitute in (1) the sample means with other different estimators.

3. The Proposed Estimators

To estimate the unknown mean $\bar{Y}$ in srswor, we suggest using in (1) the estimator $\hat{t}_i = \bar{x}_i + \alpha_i(\bar{X}_i - \bar{x}_i)$ instead of $\bar{x}_i$ (i=1,2), being α_i a real constant to be suitably chosen. In this way, we obtain the following estimators:

$$\hat{\bar{Y}}^\alpha_{R1} = \bar{y}\frac{\bar{X}_1}{\hat{t}_1}\frac{\hat{t}_2}{\bar{X}_2},\quad \hat{\bar{Y}}^\alpha_{R2} = \bar{y}\frac{\bar{X}_1}{\hat{t}_1}\frac{\bar{X}_2}{\hat{t}_2},\quad \hat{\bar{Y}}^\alpha_{P1} = \bar{y}\frac{\hat{t}_1}{\bar{X}_1}\frac{\hat{t}_2}{\bar{X}_2},\quad \hat{\bar{Y}}^\alpha_{P2} = \bar{y}\frac{\hat{t}_1}{\bar{X}_1}\frac{\bar{X}_2}{\hat{t}_2} \tag{4}$$

It can easily be checked that the estimators are biased for $\bar{Y}$ even if, as we will see later, the bias may be considered negligible for a large sample size. Now, in order to evaluate the efficiency of the proposed estimators, we obtain the expression for their mean square error using Taylor's series approximation (Wolter, 1985).

Let $\mathbf{t} = (\bar{y}, \bar{x}_1, \bar{x}_2)$ and $h(\mathbf{t})$ be a function that satisfies the conditions:

1. whatever the sample drawn is, $\mathbf{t}$ assumes values in a convex and bounded set $D \subseteq R^3$, containing the point $\mathbf{T} = E(\mathbf{t}) = (\bar{Y}, \bar{X}_1, \bar{X}_2)$;
2. in D the function $h(\mathbf{t})$ is continuous and bounded;
3. the first and second partial derivatives of $h(\mathbf{t})$ exist and are continuous and bounded in D.

Assume $h(\mathbf{t}) = \hat{\bar{Y}}^\alpha_W$ (with $W = R1,\ R2,\ P1,\ P2$) and, for the sake of simplicity, put $Y = X_0$ (hence $\bar{Y} = \bar{X}_0$ and $\bar{y} = \bar{x}_0$). Let d_i and d_{ij} indicate, respectively, the first and second partial derivative of $h(\mathbf{t})$ evaluated in $\mathbf{t} = \mathbf{T}$:

$$d_i = \left.\frac{\partial h(\mathbf{t})}{\partial x_i}\right|_{\mathbf{t}=\mathbf{T}}, \quad d_{ij} = \left.\frac{\partial^2 h(\mathbf{t})}{\partial x_i \partial x_j}\right|_{\mathbf{t}=\mathbf{T}}, \quad i, j = 0,1,2.$$

Expanding $h(\mathbf{t}) = \hat{\bar{Y}}_W^\alpha$ at the point $\mathbf{T}$ in a second order Taylor's series we get:

$$\hat{\bar{Y}}_W^\alpha \cong \bar{Y} + \sum_{i=0}^{2} d_i(\bar{x}_i - \bar{X}_i) + \tfrac{1}{2}\sum_{i=0}^{2}\sum_{j=0}^{2} d_{ij}(\bar{x}_i - \bar{X}_i)(\bar{x}_j - \bar{X}_j). \tag{5}$$

Taking in the expanded expression the expectation term-by-term up to include terms of order n^{-1}, we obtain the first order mean square error (MSE_I) of $\hat{\bar{Y}}_W^\alpha$, given by:

$$MSE_I(\hat{\bar{Y}}_W^\alpha) = E(\hat{\bar{Y}}_W^\alpha - \bar{Y})^2 \cong E\left[\sum_{i=0}^{2} d_i(\bar{x}_i - \bar{X}_i)\right]^2 = \mathbf{d}\boldsymbol{\Sigma}\,\mathbf{d}' \tag{6}$$

where $\mathbf{d} = [d_0, d_1, d_2]$ is the gradient vector:

$$\mathbf{d} = \begin{cases} [1,\ -R_1(1-\alpha_1),\ R_2(1-\alpha_2)] & \text{for } \hat{\bar{Y}}_{R1}^\alpha \\ [1,\ -R_1(1-\alpha_1),\ -R_2(1-\alpha_2)] & \text{for } \hat{\bar{Y}}_{R2}^\alpha \\ [1,\ R_1(1-\alpha_1),\ R_2(1-\alpha_2)] & \text{for } \hat{\bar{Y}}_{P1}^\alpha \\ [1,\ R_1(1-\alpha_1),\ -R_2(1-\alpha_2)] & \text{for } \hat{\bar{Y}}_{P2}^\alpha \end{cases} \tag{7}$$

with $R_i = \bar{Y}/\bar{X}_i$ and $\boldsymbol{\Sigma}$ the symmetric variance-covariance matrix of the sample means $\bar{y}$, $\bar{x}_1$ and $\bar{x}_2$:

$$\boldsymbol{\Sigma} = \frac{1-f}{n}\begin{pmatrix} S_0^2 & S_{01} & S_{02} \\ S_{10} & S_1^2 & S_{12} \\ S_{20} & S_{21} & S_2^2 \end{pmatrix} \tag{8}$$

where:

$$S_0^2 = \frac{1}{N-1}\sum_{j=1}^{N}(Y_j - \bar{Y})^2, \qquad S_{0i} = S_{i0} = \frac{1}{N-1}\sum_{j=1}^{N}(X_{ij} - \bar{X}_i)(Y_j - \bar{Y})$$

$$i=1,2$$

$$S_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}(X_{ij} - \bar{X}_i)^2, \qquad S_{12} = S_{21} = \frac{1}{N-1}\sum_{j=1}^{N}(X_{1j} - \bar{X}_1)(X_{2j} - \bar{X}_2).$$

As mentioned in Wolter (1985), it is worth emphasizing that, to the first order of approximation, the mean square error defined by (6) is the same as variance of the estimators, $MSE_I(\hat{\bar{Y}}_W^\alpha) = Var(\hat{\bar{Y}}_W^\alpha)$.

Using (7) and (8) we get the following expressions:

$$MSE_I(\hat{\bar{Y}}_{R1}^\alpha) = \frac{1-f}{n}[S_0^2 + \delta_1^2 + \delta_2^2 - 2(\delta_{01} - \delta_{02} + \delta_{12})], \tag{9}$$

$$MSE_I(\hat{\bar{Y}}_{R2}^{\alpha}) = \frac{1-f}{n}[S_0^2 + \delta_1^2 + \delta_2^2 - 2(\delta_{01} + \delta_{02} - \delta_{12})], \qquad (10)$$

$$MSE_I(\hat{\bar{Y}}_{P1}^{\alpha}) = \frac{1-f}{n}[S_0^2 + \delta_1^2 + \delta_2^2 + 2(\delta_{01} + \delta_{02} + \delta_{12})], \qquad (11)$$

$$MSE_I(\hat{\bar{Y}}_{P2}^{\alpha}) = \frac{1-f}{n}[S_0^2 + \delta_1^2 + \delta_2^2 - 2(\delta_{01} - \delta_{02} - \delta_{12})], \qquad (12)$$

with

$$\delta_{12} = R_1R_2S_{12}(1-\alpha_1)(1-\alpha_2), \quad \delta_i = R_iS_i(1-\alpha_i), \quad \delta_{0i} = R_iS_{0i}(1-\alpha_i), \quad i = 1,2.$$

The optimum values of α_1 and α_2 which minimize the mean square error of the estimators, can be easily shown as:

$$\alpha_1^* = 1 - \frac{\beta_{01.2}}{R_1}, \quad \alpha_2^* = 1 + \frac{\beta_{02.1}}{R_2} \quad \text{for} \quad \hat{\bar{Y}}_{R1}^{\alpha},$$

$$\alpha_1^* = 1 - \frac{\beta_{01.2}}{R_1}, \quad \alpha_2^* = 1 - \frac{\beta_{02.1}}{R_2} \quad \text{for} \quad \hat{\bar{Y}}_{R2}^{\alpha},$$

$$\alpha_1^* = 1 + \frac{\beta_{01.2}}{R_1}, \quad \alpha_2^* = 1 + \frac{\beta_{02.1}}{R_2} \quad \text{for} \quad \hat{\bar{Y}}_{P1}^{\alpha},$$

$$\alpha_1^* = 1 + \frac{\beta_{01.2}}{R_1}, \quad \alpha_2^* = 1 - \frac{\beta_{02.1}}{R_2} \quad \text{for} \quad \hat{\bar{Y}}_{P2}^{\alpha},$$

where

$$\beta_{01.2} = \frac{S_0}{S_1}\frac{\rho_{01} - \rho_{02}\rho_{12}}{1-\rho_{12}^2}, \quad \beta_{02.1} = \frac{S_0}{S_2}\frac{\rho_{02} - \rho_{01}\rho_{12}}{1-\rho_{12}^2}$$

are, respectively, the partial regression coefficients of Y on X_1 and of Y on X_2. Replacing the optimum values α_1^* and α_2^* in (9)-(12) it may be proved, after performing some calculations, that the proposed estimators defined in (4) have the same mean square error given by:

$$MSE_I(\hat{\bar{Y}}_W^{\alpha}) = \frac{1-f}{n}S_0^2(1-R_{0.12}^2), \qquad (13)$$

where $R_{0.12}^2$ denotes the squared multiple correlation coefficient of Y on X_1 and X_2:

$$R_{0.12}^2 = \frac{\rho_{01}^2 + \rho_{02}^2 - 2\rho_{01}\rho_{02}\rho_{12}}{1-\rho_{12}^2}.$$

It is interesting to note that the minimum mean square error in (13) coincides with that of the traditional regression estimator based on two auxiliary variables. Therefore, in the optimum case, the proposed estimators perform as well as the regression estimator.

The bias (B) of the proposed estimators may be evaluated using the approximation provided by (5). Since the sample means $\bar{y}$, $\bar{x}_1$ and $\bar{x}_2$ are unbiased estimators, considering expectations up to terms of order n^{-1} it follows that:

$$B_I(\hat{\bar{Y}}_W^{\alpha}) = E(\hat{\bar{Y}}_W^{\alpha} - \bar{Y}) \cong \tfrac{1}{2}\sum_{i=0}^{2}\sum_{j=0}^{2} d_{ij} E\left[(\bar{x}_i - \bar{X}_i)(\bar{x}_j - \bar{X}_j)\right]. \tag{14}$$

Substituting the different expressions for d_{ij} we obtain:

$$B_I(\hat{\bar{Y}}_{R1}^{\alpha}) = \frac{1-f}{n}\bar{Y}\left[-(1-\alpha_1)\rho_{01}C_0C_1 + (1-\alpha_2)\rho_{02}C_0C_2 + \right.$$
$$\left. -(1-\alpha_1)(1-\alpha_2)\rho_{12}C_1C_2 + (1-\alpha_1)^2 C_1^2\right],$$

$$B_I(\hat{\bar{Y}}_{R2}^{\alpha}) = \frac{1-f}{n}\bar{Y}\left[-(1-\alpha_1)\rho_{01}C_0C_1 - (1-\alpha_2)\rho_{02}C_0C_2 + \right.$$
$$\left. +(1-\alpha_1)(1-\alpha_2)\rho_{12}C_1C_2 + (1-\alpha_1)^2 C_1^2 + (1-\alpha_2)^2 C_2^2\right],$$

$$B_I(\hat{\bar{Y}}_{P1}^{\alpha}) = \frac{1-f}{n}\bar{Y}\left[(1-\alpha_1)\rho_{01}C_0C_1 + (1-\alpha_2)\rho_{02}C_0C_2 + \right.$$
$$\left. +(1-\alpha_1)(1-\alpha_2)\rho_{12}C_1C_2\right],$$

$$B_I(\hat{\bar{Y}}_{P2}^{\alpha}) = \frac{1-f}{n}\bar{Y}\left[(1-\alpha_1)\rho_{01}C_0C_1 - (1-\alpha_2)\rho_{02}C_0C_2 + \right.$$
$$\left. -(1-\alpha_1)(1-\alpha_2)\rho_{12}C_1C_2\right].$$

Finally, taking into account the optimum values for α_1 and α_2, we get the expressions:

$$B_I(\hat{\bar{Y}}_{R1}^{\alpha^*}) = -\frac{1-f}{n}\bar{X}_2C_0C_2\rho_{02}\beta_{02.1},$$

$$B_I(\hat{\bar{Y}}_{R2}^{\alpha^*}) = -\frac{1-f}{n}\bar{Y}R_1^{-1}R_2^{-2}C_1C_2\rho_{12}\beta_{01.2}\beta_{02.1},$$

$$B_I(\hat{\bar{Y}}_{P1}^{\alpha^*}) = -\frac{1-f}{n}\frac{\bar{Y}}{1-\rho_{12}^2}C_0\left(R_1^{-1}C_1\rho_{01}\beta_{01.2} + R_2^{-1}C_2\rho_{02}\beta_{02.1} - C_0\rho_{01}\rho_{02}\rho_{12}\right),$$

$$B_I(\hat{\bar{Y}}_{P2}^{\alpha^*}) = -\frac{1-f}{n}\bar{X}_1C_0C_1\rho_{01}\beta_{01.2}.$$

As we can see, the bias of the estimators is of the order of n^{-1} and, hence, its contribution to the true mean square error will be of the order n^{-2}. Therefore, since the true mean square error satisfies the well known identity $MSE(\hat{\bar{Y}}_W^{\alpha}) = Var(\hat{\bar{Y}}_W^{\alpha}) + [B(\hat{\bar{Y}}_W^{\alpha})]^2$, the bias becomes negligible with respect to the variance as n becomes large.

To conclude, we observe that both the mean square error and the bias of the proposed estimators depend on different quantities, such as correlation or variation coefficients, which are generally unknown in practical situations. In this work, in order to derive theoretical results, we assume that such quantities are known or available on the basis of previous data, pilot survey, past experience or efficient estimates.

4. Efficiency Comparison

In this section, we analyse the performance of the proposed estimators in order to evaluate the efficiency gain one could get by their use. Firstly, we observe that, to the first order of approximation, the proposed estimators turn out to be more efficient than the sample mean, $\bar{y}$, and the ratio and product estimators which exploit a single auxiliary variable. The proof, based on some known results concerning the regression estimator, may be found in Sukhatme et al. (1984).

Let us now consider the comparison between the proposed and the modified ratio-cum-product estimators defined, respectively, in (4) and (2). For these latter we consider the minimum mean square error assuming the optimum values of m given in Section 2. To the first order of approximation we find:

$$MSE_I(\hat{\bar{Y}}^*_{R1}) = \bar{Y}^2\left[C_{200} - \frac{(C_{101} - C_{110})^2}{C_{020} + C_{002} - 2C_{011}}\right], \tag{15}$$

$$MSE_I(\hat{\bar{Y}}^*_{R2}) = \bar{Y}^2\left[C_{200} - \frac{(C_{101} - C_{110})^2}{C_{020} + C_{002} + 2C_{011}}\right], \tag{16}$$

$$MSE_I(\hat{\bar{Y}}^*_{P1}) = \bar{Y}^2\left[C_{200} - \frac{(C_{101} + C_{110})^2}{C_{020} + C_{002} + 2C_{011}}\right], \tag{17}$$

$$MSE_I(\hat{\bar{Y}}^*_{P2}) = \bar{Y}^2\left[C_{200} - \frac{(C_{101} - C_{110})^2}{C_{020} + C_{002} - 2C_{011}}\right]. \tag{18}$$

Let $\Delta(\hat{\bar{Y}}^*_{RP}, \hat{\bar{Y}}^{\alpha^*}_W) = MSE_I(\hat{\bar{Y}}^*_{RP}) - MSE_I(\hat{\bar{Y}}^{\alpha^*}_W)$, where $\hat{\bar{Y}}^*_{RP} = \hat{\bar{Y}}^*_{R1}$, $\hat{\bar{Y}}^*_{R2}$, $\hat{\bar{Y}}^*_{P1}$, $\hat{\bar{Y}}^*_{P2}$. Neglecting the not influential term $(1-f)/n$, after some algebra we obtain:

$$\Delta(\hat{\bar{Y}}^*_{R1}, \hat{\bar{Y}}^{\alpha^*}_W) = [\rho_{12}(C_2\rho_{02} - C_1\rho_{01}) - (C_1\rho_{02} + C_2\rho_{01})]^2,$$

$$\Delta(\hat{\bar{Y}}^*_{R2}, \hat{\bar{Y}}^{\alpha^*}_W) = [\rho_{12}(C_2\rho_{02} - C_1\rho_{01}) + (C_1\rho_{02} + C_2\rho_{01})]^2,$$

$$\Delta(\hat{\bar{Y}}^*_{P1}, \hat{\bar{Y}}^{\alpha^*}_W) = [\rho_{12}(C_2\rho_{02} - C_1\rho_{01}) - (C_1\rho_{02} - C_2\rho_{01})]^2,$$

$$\Delta(\hat{\bar{Y}}^*_{P2}, \hat{\bar{Y}}^{\alpha^*}_W) = [\rho_{12}(C_2\rho_{02} + C_1\rho_{01}) + (C_1\rho_{02} + C_2\rho_{01})]^2,$$

which denote quantities that are always positive. Therefore, the proposed estimators are more efficient than the optimum modified ratio-cum-product estimators. Consequently, for previous efficiency considerations (see Section 2), they are also more efficient than Singh's original ratio-cum-product estimators. Obviously, the proposed estimators will be also more efficient than the modified ratio-cum-product estimators in the case of a not optimum sampling fraction being adopted.

5. Second Order Approximation

Up to the first order of approximation, the proposed ratio-cum-product type estimators are equivalent in terms of efficiency. Consequently, each one may be used more efficiently than the modified and original ratio-cum-product estimators if the constants α_1 and α_2 are chosen in an optimum way. Therefore, in order to better investigate the efficiency of the proposed estimators, the second order approximation for the mean square error needs to be determined. The same procedure has been adopted by Kothwala and Gupta (1988) and Hossain et al. (2003) to study the efficiency of different ratio type strategies based on a single auxiliary variable.

The second order expressions, including terms of order n^{-2}, are quite long and for their derivation we have used some results given in Sukhatme et al. (1984) and Nath (1968). Let $\beta_i = 1 - \alpha_i$ (i=1,2) and put

$$k_1 = \frac{(N-n)(N-2n)}{(N-1)(N-2)},$$

$$k_2 = \frac{N-n}{n^3}\frac{(N+1)N-6n(N-n)}{(N-1)(N-2)(N-3)},$$

$$k_3 = \frac{N-n}{n^3}\frac{N(N-n-1)(n-1)}{(N-1)(N-2)(N-3)}.$$

Then, to the second order of approximation, the mean square error of $\hat{\bar{Y}}_{R1}^{\alpha}$ is:

$$MSE_{II}(\hat{\bar{Y}}_{R1}^{\alpha}) = MSE_I(\hat{\bar{Y}}_{R1}^{\alpha}) + \hat{\bar{Y}}^2 A_{R1},$$

where

$$\begin{aligned} A_{R1} = &-2a_1\beta_1 + 2a_2\beta_2 - 2a_3\beta_1\beta_2 + a_4\beta_1^2 + a_5\beta_2^2 - 2a_6\beta_1^3 + \\ &-2a_7\beta_1\beta_2^2 + 2a_8\beta_1^2\beta_2 - 6a_9\beta_1^3\beta_2 + 3a_{10}\beta_1^2\beta_2^2 + 6a_{11}\beta_1^4 \end{aligned}$$

and

$$\begin{aligned} a_1 &= k_1C_{210} \\ a_2 &= k_1C_{201} \\ a_3 &= 2[k_2C_{211} + k_3(C_{200}C_{011} + 2C_{110}C_{101})] + 3k_1C_{110} \\ a_4 &= 4k_1C_{120} + 3[k_2C_{220} + k_3(C_{200}C_{020} + 2C_{110}^2)] \\ a_5 &= 2k_1C_{102} + k_2C_{202} + k_3(C_{200}C_{002} + 2C_{101}^2)] \\ a_6 &= k_1C_{030} + 3(k_2C_{130} + 3k_3C_{110}C_{020}) \\ a_7 &= k_1C_{012} + 2k_2C_{112} + 2k_3(C_{002}C_{110} + 2C_{101}C_{011})] \\ a_8 &= 2k_1C_{021} + 5k_2C_{121} + 5k_3(C_{020}C_{101} + 2C_{110}C_{011})] \\ a_9 &= k_2C_{031} + 3k_3C_{011}C_{020} \\ a_{10} &= k_2C_{022} + k_3(C_{020}C_{002} + 2C_{011}^2) \end{aligned}$$

$$a_{11} = k_2 C_{040} + 3k_3 C_{020}^2 .$$

For $\hat{\bar{Y}}_{R2}^{\alpha}$ we obtain:

$$MSE_{II}(\hat{\bar{Y}}_{R2}^{\alpha}) = MSE_{I}(\hat{\bar{Y}}_{R2}^{\alpha}) + \hat{\bar{Y}}^2 A_{R2} ,$$

where

$$\begin{aligned}A_{R2} = & -2b_1\beta_1 - 2b_2\beta_2 + 2b_3\beta_1\beta_2 + b_4\beta_1^2 + b_5\beta_2^2 - 2b_6\beta_1^3 - 2b_7\beta_2^3 - 2b_8\beta_1\beta_2^2 + \\ & -2b_9\beta_1^2\beta_2 + 7b_{10}\beta_1^2\beta_2^2 + 6b_{11}\beta_1^3\beta_2 + 6b_{12}\beta_1\beta_2^3 + 3b_{13}\beta_1^4 + 3b_{14}\beta_2^4\end{aligned}$$

and

$$\begin{aligned}
b_1 &= k_1 C_{210} \\
b_2 &= k_1 C_{201} \\
b_3 &= 2[k_2 C_{211} + k_3(C_{200}C_{011} + 2C_{110}C_{101})] + 3k_1 C_{111} \\
b_4 &= 4k_1 C_{120} + 3[k_2 C_{220} + k_3(C_{200}C_{020} + 2C_{110}^2)] \\
b_5 &= 4k_1 C_{102} + 3[k_2 C_{202} + k_3(C_{200}C_{002} + 2C_{101}^2)] \\
b_6 &= k_1 C_{030} + 3(k_2 C_{130} + 3k_3 C_{110}C_{020}) \\
b_7 &= k_1 C_{003} + 3(k_2 C_{103} + 3k_3 C_{101}C_{002}) \\
b_8 &= 2k_1 C_{012} + 5[k_2 C_{112} + k_3(C_{002}C_{110} + 2C_{101}C_{011})] \\
b_9 &= 2k_1 C_{021} + 5[k_2 C_{121} + k_3(C_{020}C_{101} + 2C_{110}C_{011})] \\
b_{10} &= k_2 C_{022} + k_3(C_{020}C_{002} + 2C_{011}^2) \\
b_{11} &= k_2 C_{031} + 3k_3 C_{011}C_{020} \\
b_{12} &= k_2 C_{013} + 3k_3 C_{011}C_{002} \\
b_{13} &= k_2 C_{040} + 3k_3 C_{020}^2 \\
b_{14} &= k_2 C_{004} + 3k_3 C_{002}^2 .
\end{aligned}$$

The second order mean square error of $\hat{\bar{Y}}_{P1}^{\alpha}$ is:

$$MSE_{II}(\hat{\bar{Y}}_{P1}^{\alpha}) = MSE_{I}(\hat{\bar{Y}}_{P1}^{\alpha}) + \hat{\bar{Y}}^2 A_{P1} ,$$

where

$$A_{P1} = 2c_1\beta_1 + 2c_2\beta_2 + c_3\beta_1^2 + c_4\beta_2^2 + 2c_5\beta_1\beta_2 + 2c_6\beta_1\beta_2^2 + 2c_7\beta_1^2\beta_2 + c_8\beta_1^2\beta_2^2$$

and

$$\begin{aligned}
c_1 &= k_1 C_{210} \\
c_2 &= k_1 C_{201} \\
c_3 &= 2k_1 C_{120} + k_2 C_{220} + k_3(C_{200}C_{020} + 2C_{110}^2) \\
c_4 &= 2k_1 C_{102} + k_2 C_{202} + k_3(C_{200}C_{020} + 2C_{101}^2) \\
c_5 &= 3k_1 C_{111} + 2[k_2 C_{211} + k_3(C_{200}C_{011} + 2C_{110}C_{101})] \\
c_6 &= k_1 C_{012} + 2[k_2 C_{112} + k_3(C_{002}C_{110} + 2C_{101}C_{011})] \\
c_7 &= k_1 C_{021} + 2[k_2 C_{121} + k_3(C_{020}C_{101} + 2C_{110}C_{011})]
\end{aligned}$$

$$c_8 = k_2 C_{022} + k_3 (C_{020} C_{002} + 2C_{011}^2)\,.$$

Finally, for $\hat{\bar{Y}}_{P2}^{\alpha}$ we find:

$$MSE_{II}(\hat{\bar{Y}}_{P2}^{\alpha}) = MSE_{I}(\hat{\bar{Y}}_{P2}^{\alpha}) + \hat{\bar{Y}}^2 A_{P2}\,,$$

where

$$\begin{aligned} A_{P2} = {} & 2d_1\beta_1 - 2d_2\beta_2 - 4d_3\beta_1\beta_2 + d_4\beta_1^2 + d_5\beta_2^2 - 2d_6\beta_2^3 + \\ & +2d_7\beta_1\beta_2^2 - 2d_8\beta_1^2\beta_2 - 6d_9\beta_1\beta_2^3 + 3d_{10}\beta_1^2\beta_2^2 + 3d_{11}\beta_2^4 \end{aligned}$$

and

$$\begin{aligned} d_1 &= k_1 C_{210} \\ d_2 &= k_1 C_{201} \\ d_3 &= k_2 C_{211} + k_3 (C_{200} C_{011} + 2C_{110} C_{101}) \\ d_4 &= k_1 C_{120} + k_2 C_{220} + k_3 (C_{200} C_{020} + 2C_{110}^2) \\ d_5 &= k_1 C_{102} + 3[k_2 C_{202} + k_3 (C_{200} C_{002} + 2C_{101}^2)] \\ d_6 &= k_1 C_{003} + 3(k_2 C_{103} + 3k_3 C_{101} C_{002}) \\ d_7 &= 2k_1 C_{012} + 5[k_2 C_{112} + k_3 (C_{002} C_{110} + 2C_{101} C_{011})] \\ d_8 &= k_1 C_{021} + 2[k_2 C_{121} + k_3 (C_{020} C_{101} + 2C_{110} C_{011})] \\ d_9 &= k_2 C_{013} + 3k_3 C_{011} C_{002} \\ d_{10} &= k_2 C_{022} + k_3 (C_{020} C_{002} + 2C_{011}^2) \\ d_{11} &= k_2 C_{004} + 3k_3 C_{002}^2). \end{aligned}$$

The optimum values of β_i and, hence of $\alpha_i = 1 - \beta_i$, which minimize the second order mean square error of $\hat{\bar{Y}}_{R1}^{\alpha}$, $\hat{\bar{Y}}_{R2}^{\alpha}$, $\hat{\bar{Y}}_{P1}^{\alpha}$ and $\hat{\bar{Y}}_{P2}^{\alpha}$ may be obtained using numerical optimization routines.

6. Empirical Analysis

In this section we analyse the performance of the suggested estimators by means of a numerical evaluation of the first and second order mean square error. To the first order of approximation and, for a fixed sample size, we consider the efficiency of the estimators with respect to: (i) the ratio and product estimators exploiting a single auxiliary variable; (ii) the original ratio-cum-product estimators defined in (1); (iii) the modified ratio-cum-product estimators defined in (2) with a not optimum sampling fraction $f = m/(1+m)$.

To the second order of approximation, we evaluate the performance of the suggested estimators for different sample sizes. The study is based on two real data sets.

Data Set 1. The data are taken from the Survey of Household Income and Wealth 2002 conducted by the Bank of Italy (Bank of Italy, 2002). The survey covers 8011 Italian households composed of 22148 individuals and 13536 income-earners. In the analysis, we assume the 8011 households as the target population on which three variables are considered: the household net disposable income (Y), the household consumption (X_1) and the number of household income-earners (X_2). The following values are obtained for the considered variables:

$$\bar{Y} = 28229.427 \quad \bar{X}_1 = 20418.618 \quad \bar{X}_2 = 1.6897$$
$$C_0 = 0.787 \quad C_1 = 0.668 \quad C_2 = 0.4596$$
$$\rho_{01} = 0.74 \quad \rho_{02} = 0.458 \quad \rho_{12} = 0.348.$$

Data Set 2. The data have been collected by a market research company. The population consists of 2376 points of sale for which three variables are surveyed: the sale area (Y) (in square meters), the number of employees (X_1) and the amount of soft drinks sales (X_2) (in euro × 1000) in a year. For this population, the following values are obtained:

$$\bar{Y} = 1701.946 \quad \bar{X}_1 = 40.617 \quad \bar{X}_2 = 615.637$$
$$C_0 = 1.285 \quad C_1 = 2.35 \quad C_2 = 1.651$$
$$\rho_{01} = 0.898 \quad \rho_{02} = 0.861 \quad \rho_{12} = 0.773.$$

The performance of the estimators is evaluated through the relative efficiency with respect to the sample mean $\bar{y}$ defined, for the generic estimator $\hat{\theta}$, as

$$eff(\hat{\theta}) = \frac{Var(\bar{y})}{MSE(\hat{\theta})}.$$

Table 1 shows, the first order mean square error and the relative efficiency for the considered estimators. According to the theoretical results, we notice that the proposed estimators ($\hat{\bar{Y}}_W^{\alpha^*}$) dominate all the others in the sense that they show the smallest mean square error and, hence, the highest efficiency. Moreover, for the data sets in the analysis, we observe the poor performance of both the original and modified ratio-cum-product estimators. We point out that, except for the estimators $\hat{\bar{Y}}_{R1}^*$, $\hat{\bar{Y}}_{R2}^*$, $\hat{\bar{Y}}_{P1}^*$ and $\hat{\bar{Y}}_{P2}^*$, the relative efficiency does not depend on the sample size.

As mentioned earlier, the proposed estimators are equally efficient to the first order of approximation. Therefore, if the purpose is to find the most efficient estimator among the proposed ones, the approximation up to the second order needs to be considered.

For both the data sets considered, Table 2 reproduces the values of the second order minimum mean square error and the relative efficiency of the proposed estimators, $\hat{\bar{Y}}_{R1}^{\alpha^*}$, $\hat{\bar{Y}}_{R2}^{\alpha^*}$, $\hat{\bar{Y}}_{P1}^{\alpha^*}$ and $\hat{\bar{Y}}_{P2}^{\alpha^*}$. To better explore the second order approximation, we give the results for different sample sizes. The optimum values of α_1 and α_2 which minimize the mean square errors are shown in Table 3. These values have been obtained numerically through the optimization routine `fminsearch` in Matlab.

The behaviour of the second order mean square error seems to be different in the two data sets. In the first one, the relative efficiency of the four estimators is slightly higher than that at the first order ($eff_I = 2.461$). Indeed, the difference between first and second order efficiency may be considered negligible, even for small sample size. However, for the different sample sizes, the most efficient estimator seems to be $\hat{\bar{Y}}_{P2}^{\alpha^*}$.

In the second data set, all the estimators, except $\hat{\bar{Y}}_{P2}^{\alpha^*}$, show a relative efficiency lower than that at the first order ($eff_I = 8.028$). For n=50, the estimator $\hat{\bar{Y}}_{P1}^{\alpha^*}$ shows the maximum loss in efficiency: the decrease in the second order efficiency as compared to the first one is 19.8%. On the contrary, for the same sample size, the estimator $\hat{\bar{Y}}_{P2}^{\alpha^*}$ improves its performance of 92.26%. For n=250, the increase in the second order efficiency of $\hat{\bar{Y}}_{P2}^{\alpha^*}$ is 7.17%.

Table 1. First order efficiency comparison among different estimators.

		Data Set 1, *n=750*		**Data Set 2, *n=250***	
$\hat{\theta}$	*Auxiliary variables*	$MSE_I(\hat{\theta})$	$eff_I(\hat{\theta})$	$MSE_I(\hat{\theta})$	$eff_I(\hat{\theta})$
$\bar{y}$	*none*	596932.400	1	13323.699	1.000
$\hat{\bar{Y}}_r$	X_1	276716.830	2.157	5346.911	2.492
	X_2	481390.424	1.240	4424.899	3.011
$\hat{\bar{Y}}_p$	X_1	1776976.857	0.336	70036.653	0.190
	X_2	1119496.427	0.533	56465.206	0.236
$\hat{\bar{Y}}_W^{\alpha^*}$	X_1, X_2	242524.632	**2.461**	1659.555	**8.028**
$\hat{\bar{Y}}_{R1}$	X_1, X_2	593459.953	1.006	16917.322	0.788
$\hat{\bar{Y}}_{R2}$	X_1, X_2	366995.759	1.627	28019.206	0.476
$\hat{\bar{Y}}_{P1}$	X_1, X_2	2505361.788	0.238	144749.255	0.092
$\hat{\bar{Y}}_{P2}$	X_1, X_2	1455613.977	0.410	29566.757	0.451
$\hat{\bar{Y}}_{R1}^*$	X_1, X_2	646021.191	0.924	14204.582	0.938
$\hat{\bar{Y}}_{R2}^*$	X_1, X_2	716324.021	0.833	21197.210	0.629
$\hat{\bar{Y}}_{P1}^*$	X_1, X_2	495448.863	1.205	7470.723	1.783
$\hat{\bar{Y}}_{P2}^*$	X_1, X_2	556967.959	1.072	12717.114	1.048

Going for second order approximation is certainly more useful in Data Set 2 than in Data Set 1. In fact, the study of the second order mean square error allows us to better understand the behaviour of the proposed estimators and to find the most efficient estimator when the sample size changes.

Probably, the different behaviour of the estimators in the two data sets is caused by the different types of populations we have considered. The population described in Data Set 2 shows, as compared to that in Data Set 1, a higher variability and higher correlation between the variables. Particularly, the high variability in the auxiliary variables may affect the first order mean square error making it inaccurate. Therefore, the second order approximation may be also useful to evaluate the accuracy of the first order approximation when non-homogeneous populations are investigated. For a fairly homogeneous population, like the one described in Data Set 1, it should not be strange to observe small differences between the first and second order mean square errors even for small sample sizes.

Table 2. Second order mean square error and efficiency (in bold) for the proposed estimators.

Data Set 1				
$\hat{\theta}$	*n=150*	*n=300*	*n=750*	*n=1000*
$\hat{\bar{Y}}_{R1}^{\alpha^*}$	1303089.93 **2.480**	641578.41 **2.470**	242234.98 **2.464**	175492.26 **2.463**
$\hat{\bar{Y}}_{R2}^{\alpha^*}$	1303272.34 **2.479**	641622.16 **2.470**	242241.27 **2.464**	175495.60 **2.463**
$\hat{\bar{Y}}_{P1}^{\alpha^*}$	1305264.69 **2.476**	642124.60 **2.468**	242320.04 **2.463**	175539.15 **2.463**
$\hat{\bar{Y}}_{P2}^{\alpha^*}$	1301549.21 **2.483**	641248.44 **2.472**	242203.16 **2.465**	175480.19 **2.463**

Data Set 2				
$\hat{\theta}$	*n=50*	*n=100*	*n=250*	*n=500*
$\hat{\bar{Y}}_{R1}^{\alpha^*}$	9684.318 **7.526**	4593.21 **7.764**	1682.46 **7.919**	737.23 **7.974**
$\hat{\bar{Y}}_{R2}^{\alpha^*}$	9296.012 **7.841**	4491.50 **7.939**	1667.50 **7.990**	734.28 **8.006**
$\hat{\bar{Y}}_{P1}^{\alpha^*}$	11318.73 **6.439**	4989.35 **7.147**	1737.81 **7.667**	747.8757 **7.860**
$\hat{\bar{Y}}_{P2}^{\alpha^*}$	4722.056 **15.435**	3499.44 **10.190**	1548.78 **8.603**	716.6847 **8.202**

Table 3. Optimum values of α_1 and α_2 which minimize the second order mean square error.

Data Set 1				
$\hat{\theta}$	*n=150*	*n=300*	*n=750*	*n=1000*
$\hat{\bar{Y}}_{R1}^{\alpha^*}$	1.7790, 0.6099	1.7790, 0.6100	1.7790, 0.6101	1.7790, 0.6101
$\hat{\bar{Y}}_{R2}^{\alpha^*}$	1.7788, 1.3903	1.7790, 1.3901	1.7790, 1.3900	1.7790, 1.3899
$\hat{\bar{Y}}_{P1}^{\alpha^*}$	0.2314, 0.6065	0.2260, 0.6083	0.2228, 0.6095	0.2222, 0.6097
$\hat{\bar{Y}}_{P2}^{\alpha^*}$	0.2315, 1.3990	0.2261, 1.3943	0.2228, 1.3915	0.2223, 1.3910

Data Set 2				
$\hat{\theta}$	*n=50*	*n=100*	*n=250*	*n=500*
$\hat{\bar{Y}}_{R1}^{\alpha^*}$	1.4640, 0.6810	1.4451, 0.6578	1.4330, 0.6422	1.4289, 0.6366
$\hat{\bar{Y}}_{R2}^{\alpha^*}$	1.4347, 1.3662	1.4299, 1.3668	1.4272, 1.3672	1.4263, 1.3674
$\hat{\bar{Y}}_{P1}^{\alpha^*}$	0.6149, 0.6357	0.5955, 0.6320	0.5823, 0.6317	0.5776, 0.6320
$\hat{\bar{Y}}_{P2}^{\alpha^*}$	0.6132, 1.4803	0.6028, 1.4291	0.5878, 1.3921	0.5804, 1.3778

REFERENCES

ABU-DAYYEH, W.A., AHMED, M.S., AHMED, R.A. and MUTTLAK, H.A. (2003). Some estimators of finite population mean using auxiliary information, Applied Mathematics and Computation, 139, 287—298.

BANDYOPADHYAY, S. (1980). Improved ratio and product estimators, Sankhyā C, 42, 45—49.

BANK OF ITALY (2002). Survey of Household Income and Wealth 2002. http://www.bancaditalia.it/statistiche/.

CECCON, C. and DIANA, G. (1996). Classi di stimatori di tipo rapporto in presenza di più variabili ausiliarie, Proceedings of the XXXVIII Scientific Meeting of the Italian Statistical Society, Rimini, 9-13 April 1996, Vol. 2, 119—126.

COCHRAN, W.G. (1977). Sampling Techniques, John Wiley & Sons, New York.

HOSSAIN, M.I., RAHMAN, M.S. and AHMED, M.S. (2003). Second order properties of some estimators under double sampling, Statistics in Transition, 6, n. 4, 543—554.

KADILAR, C. and CINGI, H. (2004). Estimator of a population mean using two auxiliary variables in simple random sampling, International Mathematical Journal, 5, 357—360.

KADILAR, C. and CINGI, H. (2005). A new estimator using two auxiliary variables, Applied Mathematics and Computation, 162, 901—908.

KOTHWALA, N.H. and GUPTA, P.C. (1988). A study of second order approximation of some ratio type strategies, Biometrical Journal, 30, 369—377.

MURTHY, M.N. (1977). Sampling Theory and Methods, Statistical Publishing Society, Calcutta.

NAIK, V.D. and GUPTA, P.C. (1991). A general class of estimators for estimating population mean using auxiliary information, Metrika, 38, 11—17.

NATH, S.N. (1968). On the product moments from a finite universe, Journal of the American Statistical Association, 63, 535—541.

OLKIN, I. (1958). Multivariate ratio estimation for finite populations, Biometrika, 45, 154—165.

PERRI, P.F. (2004). Alcune considerazioni sull'efficienza degli stimatori rapporto-cum-prodotto, Statistica & Applicazioni, 2, n. 2, 59—75.

RAO, P.S.R.S. and MUDHOLKAR, G.S. (1967). Generalized multivariate estimator for the mean of finite populations, Journal of the American Statistical Association, 62, 1009—1012.

RAJ, D. (1965). On a method of using multi-auxiliary information in sample surveys, Journal of the American Statistical Association, 60, 154—165.

SAHOO, L.N. (1991). An unbiased ratio-cum-product estimator in two-stage sampling, Metron, Vol. XLIX, n. 1—4, 213—217.

SAHOO, L.N. and SWAIN, A.K.P.C. (1980). Unbiased ratio-cum-product estimator, Sankhyā C, 42, 56—62.

SINGH, M.P. (1965). On the estimation of ratio and product of the population parameters, Sankhyā B, 27, 321—328.

SINGH, M.P. (1967a). Multivariate product method of estimation for finite populations, Journal of the Indian Society of Agricultural Statistics, 31, 375—378.

SINGH, M.P. (1967b). Ratio cum product method of estimation, Metrika, 12, 34—42.

SRIVENKATARAMANA, T. (1980). A dual to ratio estimator in sample surveys, Biometrika, 67, 199—204.

SRIVENKATARAMANA, T. and TRACY, D.S. (1981). An alternative to ratio method in sample surveys, Annals of the Institute of Statistical Mathematics, 32 A, 111—120.

SUKHATME, P.V., SUKHATME, B.V., SUKHATME, S. and ASOK, C. (1984). Sampling Theory of Surveys with Applications, Iowa State University Press, Ames, Iowa, U.S.A

TRACY, D.S., SINGH, H.P. and SINGH, R. (1996). An alternative to the ratio-cum-product estimator in sample surveys, Journal of Statistical Planning and Inference, 53, 375—387.

WOLTER, K.M. (1985). Introduction to Variance Estimation, Spinger-Verlag, New York.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 71—87

ON PSPNR SAMPLING SCHEME FOR MEAN ESTIMATION

D.Shukla and Jayant Dubey[1]

ABSTRACT

This paper presents a "Post-stratified partial non-response (PSPNR)" sampling scheme useful to cope with the presence of partial non-response in surveys using post-stratification. An unbiased estimation strategy is proposed and its optimum properties are examined. The cost aspect, under PSPNR, is explored and observed having some limitations regarding solution of equations. An alternative cost strategy is proposed to deal with which provides cost-optimal selection of required sample-size and optimum allocation of sample fractions. The derived properties and results are numerically supported.

Key words : Simple random sampling without replacement (SRSWOR), Post-stratified partial non-response (PSPNR), Response (R), non-response (NR).

1. Introduction

The presence of non-response in a sample survey is mainly of two types (i) complete non-response (ii) partial non-response. Hansen and Hurwitz (1946) tackled the problem of non-response in SRSWOR set-up for mail surveys. Jagers et.al. (1985) advocated that post-stratification with respect to the relevant criteria may improve upon estimation substantially over the sample mean or ratio estimator. Shukla and Dubey (2001) proposed a PSNR sampling design for dealing the non-response in sample surveys. This paper extents the similar approach in the post- stratified set-up in a more efficient manner for estimating the population mean.

1.1 Let a finite population U consists of units labeled (U_1, U_2,................. U_N) in some order. With each unit U_d (d = 1, 2, 3, ...N) a variable X is associated, used for stratification into k strata, each of size N_i (i = 1, 2, 3, ..k) with

[1] Department of Mathematics and Statistics, Dr. H.S.Gour University Sagar, (M.P.),470003, INDIA
E-mail: diwakarshukla@rediffmail.com

$\sum_{i=1}^{k} W_i = \sum_{i=1}^{k}\left[\frac{N_i}{N}\right] = 1$. Every i^{th} strata contains a group who used to response ($N_i^{'}$) and a group who never response ($N_i^{''}$) during the survey ($N_i = N_i^{'} + N_i^{''}$). The sizes (or proportions) regarding N, N_i, $N_i^{'}$ and $N_i^{''}$ are assumed known while frames of each strata are unknown. The variable of main interest is Y having m^{th} value of j^{th} group of response (R) or non-response (NR) in the i^{th} strata denoted as Y_{ijm} {j =1 for R and 2 for NR; m = 1, 2, 3….($N_i^{'}, N_i^{''}$ based on j)}. Within the i^{th} strata, S_{1i}^{2} for R and S_{2i}^{2} for NR; between strata S_i^{2} and overall S^2 are population mean squares.

2. Post — Stratified Partial Non-Response (PSPNR) Scheme

Following Shukla and Dubey (2001), we propose a scheme PSPNR assuming a compulsive partial non-response as under :

STEP I : Draw a random sample of size n by SRSWOR from population N.

STEP II : Conduct a low cost pilot personal-enquiry for information on X about n.

STEP III : Post-stratify n units into random n_i units, $\sum n_i = n$, based on X, and assume n_i is a moderate number.

STEP IV : Using an economical data collection procedure like by post, e-mail, telephone, internet or any other (avoiding face-to-face personal interview) collect information on Y from n units until a prefixed deadline. If response is cent percent, $\bar{y} = \left(n^{-1}\right)\sum_{i=1}^{k} n_i \bar{y}_i$ is sample mean derived from $\bar{y}_i$ on n_i units along with a post-stratified estimate $\bar{y}_{ps} = \sum_{i=1}^{k} W_i \bar{y}_i$ for mean $\bar{Y}$. The $\bar{y}_i$ is mean based on i^{th} strata units n_i in the sample n and $\sum n_i = n$.

STEP V : When the deadline is over, let $n_i^{'}$ units responded and $n_i^{''}$ have no response in the i^{th} strata with the assumption $n_i^{'} > 0$, $n_i^{''} > 0$, ($n_i^{'} + n_i^{''} = n_i$) which compel the scheme to be a partial non-response (PNR). The mean of $n_i^{'}$ units is $\bar{y}_i^{'}$.

STEP VI : To cope with non-response, draw sub-sample of size $n_i^{'''} > 0$ from $n_i^{''}$ ($n_i^{'''} < n_i^{''}$) by SRSWOR, maintaining a prefixed ratio f_i = ($n_i^{''} / n_i^{'''}$) for all *i* over *k.*

STEP VII : Conduct a face-to-face personal enquiry with high cost and high accuracy, over $n_i^{'''}$ units for gathering information on *Y.* Assume all responded well, with mean $\bar{y}_i^{'''}$.

Remark 2.0: While $n_i^{''} = 0$, the PSPNR converts into usual post-stratified scheme discussed in sukhatme et.al. (1984).

Remark 2.1: Assumptions $n_i^{'}$ > 0, $n_i^{''}$ > 0 in step V are extra condition to the PSNR scheme of Shukla and Dubey (2001). These avoid many mathematical complications likely to generate due to earlier scheme. Moreover, in the proposed PSPNR the more efficient estimation is possible as discussed below.

3. Estimation

Define functions $\phi_1(a,b) = a/b, \phi_2 = (a,b) = a.b$ and a constant $\alpha \in [0,1]$. For any real *a, b* and design D as PSPNR, the proposed strategy $[D, \phi_2(\bar{y}_\alpha, 1)]$, for estimating $\bar{Y}$, is

$$\phi_2(\bar{y}_\alpha,1) = \sum_{i=1}^{k} W_{i\alpha}\left[\phi_1(n_i^{'}, n_i)\phi_2(\bar{y}_i^{'},1) + \phi_1(n_i^{''}, n_i)\ \phi_2(\bar{y}_i^{'''},1)\right]$$

where, $W_{i\alpha} = \phi_1(\alpha, p_i) + \phi_2(1-\alpha, W_i),\ p_i = \phi_1(n_i\ , n\ ),\ W_i = \phi_1(N_i, N)$

Remark 3.1: The term $W_{i\alpha}$ is a weight structure adopted by Agrawal and Panda (1993) depicting a gain in efficiency due to suitable choice of α. Some useful contributions for the post-stratification are by Holt and Smith (1979), Jagers (1986), Smith (1991), Casady and Valliant (1993), Agrawal and Panda (1995), Wywil (2001) etc. and for the non-response are by Hinde and Chambers (1991), Jackway and Boyce (1987), Jones (1983), Khare (1987), Kott (1994), Little (1982, 1986), Sarandal and Swensson (1987), Lessler and Kalsbeek (1992), Groves and Couper (1998) etc.

Note 3.1: While $n_i^{''} = 0$, the estimator $\phi_2(\bar{y}_\alpha, 1) = \bar{y}_{ps}$ where $\bar{y}_{ps}$ is a usual post-stratified estimator discussed in Sukhetme (1984).

Note 3.2: $E[(...)] = EEE\left[(...)/(n_i^{'}, n_i^{''})\right] = E\left[E\{E(...)/(n_i\ , n_i^{''})\}/n_i\right]$

THEOREM 3.1: Under $\left[D,\phi_2\left(\bar{y}_\alpha,1\right)\right]$ the $\phi_2\left(\bar{y}_\alpha,1\right)$ is unbiased for $\bar{Y}$.

Proof: $E\left[\phi_2\left(\bar{y}_\alpha,1\right)\right]=E\left[E\left[E\left\{\phi_2\left(\bar{y}_\alpha,1\right)\right\}/\left(n_i^{'},n_i^{''}\right)\right]\right]$

$$=E\left[E\left[\left\{\sum_{i=1}^{k}W_{i\alpha}\left[\phi_1\left(n_i^{'},n_i\right)\phi_2\left(\bar{y}_i^{'},1\right)+\phi_1\left(n_i^{''},n_i\right)E\left\{\phi_2\left(\bar{y}_i^{'''},1\right)\right\}\right]\right\}\Big/\left(n_i,n_i^{''}\right)\right]\right]$$

$$=E\left[E\left[\left\{\sum_{i=1}^{k}W_{i\alpha}\left[\phi_1\left(n_i^{'},n_i\right)\phi_2\left(\bar{y}_i^{'},1\right)+\phi_1\left(n_i^{''},n_i\right)\left\{\phi_2\left(\bar{y}_i^{''},1\right)\right\}\right]\right\}\Big/\left(n_i\right)\right]\right]$$

$$=E\left[\left\{\sum_{i=1}^{k}W_{i\alpha}\left[E\ \left\{\phi_2\left(\bar{y}_i\ ,1\right)\right\}\right]\right\}\Big/n_i\right]$$

$$=\sum_{i=1}^{k}\left\{\phi_2\left(\bar{Y}_i\ ,1\right).E(W_{i\alpha})\right\}\ =\sum_{i=1}^{k}W_i\bar{Y}_i$$

THEOREM 3.2: Under the set-up of strategy $\left[D,\phi_2\left(\bar{y}_\alpha,1\right)\right]$

$$=E\left[E\left[V\left[\left\{\phi_2\left(\bar{y}_\alpha,1\right)\right\}/\left(n_i,n_i^{''}\right)\right]\right]\right]=\sum_{i=1}^{k}\left[\left\{\left(\frac{\alpha^2}{n}\right)E\left(p_i\right)+\left(1-\alpha\right)^2W_i^{\ 2}E\left(\frac{1}{n_i}\right)+2\alpha\left(1-\alpha\right)\left(\frac{W_i}{n}\right)\right\}A_i\right]$$

where Ai = $N_i^{''}\left(N_i\right)^{-1}\left[\left(f_i-1\right)S_{2i}^2\right]$and $f_i=\left(n_i^{''}\right)\left(n_i^{'''}\right)^{-1}$

Proof: $=E\left[E\left[V\left\{\sum_{i=1}^{k}W_{i\alpha}\left\{\phi_1\left(n_i^{''},n_i^{'}\right)\phi_2\left(\bar{y}_i^{''},1\right)\right\}\right\}\Big/n_i\ ,n_i^{''}\right]\right]$

$$=E\left[E\left[\left\{\sum_{i=1}^{k}W_{i\alpha}^2n_i^{-2}\left(n_i^{''}\right)^2V\left\{\phi_2\left(\bar{y}_i^{'''},1\right)\right\}\right\}\Big/n_i\ ,n_i^{''}\right]\right]$$

$$=E\left[\left\{\sum_{i=1}^{k}\left(W_{i\alpha}^2n_i^{-1}\right)\left\{E\left(\frac{n_i^{''}}{n_i}\right)\right\}\left(f_i-1\right)S_{2i}^2\right\}\Big/n_i\right]$$

$$=\sum_{i=1}^{k}A_i\left[\frac{\alpha^2}{n}E\left(p_i\right)+\left(1-\alpha\right)^2W_i^{\ 2}.E\left(\frac{1}{n_i}\right)+2\alpha\left(1-\alpha\right)\left(\frac{W_i}{n}\right)\right]$$

Remark 3.2: The detailed form of Theorem 3.2 is

$$=E\left[E\left[V\left[\left\{\phi_2\left(\bar{y}_\alpha,1\right)\right\}/n_i,n_i^{''}\right]\right]\right]\ =\left(\frac{1}{n}\right)\sum_{i=1}^{k}W_iA_i+\left[\frac{\left(1-\alpha\right)^2\left(N-n\right)}{\left(N-1\right)n^2}\right]\sum_{i=1}^{k}\left(1-W_i\right)A_i$$

Remark 3.3: The conditional covariance, for $,i\neq\ j=1,2.........k$, is

$Cov\left[\left\{\phi_2\left(\bar{y}_i,1\right),\phi_2\left(\bar{y}_j,1\right)\right\}/n_i,n_j^{''}\right]\ =0$, since $\bar{y}_i,\bar{y}_j$ are independent for given n_i, n_j

THEOREM 3.3: Under strategy $\left[D,\phi_2\left(\bar{y}_\alpha,1\right)\right]$

$$E\left[V\left[E\left[\phi_2(\bar{y}_\alpha,1)\right]/n_i,n_i^{''}\right]\right]$$

$$=\sum_{i=1}^{k}S_i^2\left[\alpha^2\left\{\frac{1}{n}E(p_i)-\frac{1}{N_i}E\left(p_i^2\right)\right\}+(1-\alpha)^2\left\{W_i^2E\left(\frac{1}{n_i}\right)-\left(\frac{W_i}{N}\right)\right\}+2\alpha(1-\alpha)\left\{\left(\frac{W_i}{n}\right)-\left(\frac{1}{N}\right)E(p_i)\right\}\right]$$

Proof : $E\left[V\left[\sum_{i=1}^{k}W_{i\alpha}\left\{\phi_1\left(n_i^{'},n_i\right)\phi_2\left(\bar{y}_i^{'},1\right)+\phi_1\left(n_i^{''},n_i\right)E\left\{\phi_2\left(\bar{y}_i^{'''},1\right)\right\}\right\}/\,n_i,n_i^{''}\right]\right]$

$$=E\left[V\left[\sum_{i=1}^{k}W_{i\alpha}\left[\phi_1\left(n_i^{'},n_i\right)\phi_2\left(\bar{y}_i^{'},1\right)+\phi_1\left(n_i^{''},n_i\right)\left\{\phi_2\left(\bar{y}_i^{''},1\right)\right\}\right]/\,n_i\right]\right]$$

$$=\sum_{i=1}^{k}S_i^2\left[\alpha^2\left\{\frac{1}{n}E(p_i)-\frac{1}{N_i}E\left(p_i^2\right)\right\}+(1-\alpha)^2\left\{W_i^2E\left(\frac{1}{n_i}\right)-\left(\frac{W_i}{N}\right)\right\}+2\alpha(1-\alpha)\left\{\left(\frac{W_i}{n}\right)-\left(\frac{1}{N}\right)E(p_i)\right\}\right]$$

Remark 3.4: The detail form of theorem 3.3 is

$$E\left[V\left[E\left\{\phi_2(\bar{y}_\alpha,1)\right\}/n_i,n_i^{''}\right]\right]=\left(\frac{1}{n}-\frac{1}{N}\right)\sum_{i=1}^{k}W_iS_i^2-\frac{(N-n)}{n(N-1)}\left\{\frac{\alpha^2}{N}-\frac{\left(1-\alpha^2\right)}{n}\right\}\sum_{i=1}^{k}(1-W_i)S_i^2$$

THEOREM 3.4: Under strategy $\left[D,\phi_2(\bar{y}_\alpha,1)\right]$

$$V\left[E\left[E\left\{\phi_2(\bar{y}_\alpha,1)\right\}/n_i,n_i^{''}\right]\right]=\alpha^2\left[\sum_{i=1}^{k}V(p_i)\bar{Y}_i^2+\sum_{i\neq}^{k}\sum_{j=1}^{k}Cov\left(p_ip_j\right)\bar{Y}_i\bar{Y}_j\right]$$

Proof : $V\left[E\left[\left\{\sum_{i=1}^{k}W_{i\alpha}\left\{\phi_1\left(n_i^{'},n_i\right)\phi_2\left(\bar{y}_i^{'},1\right)+\phi_1\left(n_i^{''},n_i\right)E\left\{\phi_2\left(\bar{y}_i^{'''},1\right)\right\}\right\}\right\}\Big/n_i,n_i^{''}\right]\right]$

$$=V\left[\sum_{i=1}^{k}W_{i\alpha}\;E\left\{\phi_2\left(\bar{y}_i,1\right)\right\}/n_i\right]=\alpha^2\left[\sum_{i=1}^{k}V(p_i)\bar{Y}_i^2+\sum_{i\neq}^{k}\sum_{j=1}^{k}Cov\left(p_ip_j\right)\bar{Y}_i\bar{Y}_j\right]$$

Remark 3.5: $V\left[E\left[E\left\{\phi_2(\bar{y}_\alpha,1)/n_i,n_i^{''}\right\}\right]\right]=\frac{\alpha^2\;(N-n)}{n(N-1)}\left\{S^2-\sum_{i=1}^{k}W_iS_i^2\right\}$ after simplification.

THEOREM 3.5: Under $\left[D,\phi_2(\bar{y}_\alpha,1)\right]$, the variance of $\phi_2(\bar{y}_\alpha,1)$ is

$$V[\phi_2(\bar{y}_\alpha,1)]=\alpha^2\left[\sum_{i=1}^{k}\left\{\frac{1}{n}W_i\left(A_i+S_i^2\right)+V(p_i)\left(\bar{Y}_i^2-\frac{S_i^2}{N_i}\right)-\frac{1}{N}W_iS_i^2\right\}+\sum_{i\neq}^{k}\sum_{j=1}^{k}Cov\left(p_i,\,p_j\right)\bar{Y}_i\bar{Y}_j\right]$$

$$+(1-\alpha)^2\left[\sum_{i=1}^{k}\left\{W_i^2E\left(\frac{1}{n_i}\right)\left(A_i+S_i^2\right)-\frac{1}{N}W_iS_i^2\right\}\right]+2\alpha(1-\alpha)\left[\sum_{i=1}^{k}\left\{\left(\frac{W_i}{n}\right)\left(A_i+S_i^2\right)-\frac{1}{N}W_iS_i^2\right\}\right]$$

Proof: $V\left[\phi_2(\bar{y}_\alpha,1)\right]=E\left[E\left[V\left\{\phi_2(\bar{y}_\alpha,1)\right\}/\,n_i,n_i^{''}\right]\right]$

$$+E\left[V\left[E\left\{\phi_2(\bar{y}_\alpha,1)\right\}/n_i,n_i^{''}\right]\right]+V\left[E\left[E\left\{\phi_2(\bar{y}_\alpha,1)\right\}/\,n_i,n_i^{''}\right]\right]$$

Addition of theorem 3.2, 3.3 and 3.4 provides the proof.

Corollary 3.1: Theorem 3.5 could also be expressed in the form

$$V[\phi_2(\bar{y}_\alpha,1)] = \alpha^2 B_1 + (1-\alpha)^2 B_2 + 2\alpha(1-\alpha)B_3$$

where $B_1 = \sum_{i=1}^{k}\left[\frac{1}{n}W_i\left(A_i + S_i^2\right) + V(p_i)\left(\bar{Y}_i^2 - \frac{S_i^2}{N_i}\right) - \frac{1}{N}W_i S_i^2\right] + \sum_{i\neq}^{k}\sum_{j=1}^{k}\left\{Cov\left(p_i p_j\right)\bar{Y}_i\bar{Y}_j\right\}$

$$B_2 = \sum_{i=1}^{k}\left\{W_i^2 E\left(\frac{1}{n_i}\right)\left(A_i + S_i^2\right) - \frac{1}{N}W_i S_i^2\right\};\ B_3 = \sum_{i=1}^{k}\left\{\left(\frac{W_i}{n}\right)\left(A_i + S_i^2\right) - \frac{1}{N}W_i S_i^2\right\}$$

Note 3.1: Define further

$$B_4 = \sum_{i=1}^{k}\left[V(p_i)\left(\bar{Y}_i^2 - \frac{S_i^2}{N_i}\right)\right],\ B_5 = \frac{(N-n)}{n^2(N-1)}\left[\sum_{i=1}^{k}(1-W_i)\left(A_i + S_i^2\right)\right]$$

Clearly, $B_1 = (B_3 + B_4)$ and $B_5 = (B_2 - B_3)$

THEOREM 3.6: The optimum choice of constant α and optimum variance is

$$\alpha_{opt} = \frac{B_5}{\left(B_4 + B_5\right)},\quad V[\phi_2(\bar{y}_\alpha,1)]_{opt} = \left\{\frac{B_2 B_4 + B_3 B_5}{B_4 + B_5}\right\}$$

THEOREM 3.7: The $[D,\phi_2(\bar{y}_\alpha,1)]$ is efficient over $[D,\phi_2(\bar{y}_0,1)]$ if $\alpha\,\varepsilon\,(0,2\alpha_{opt})$..

Proof: As per corollary 3.1 , we write

$$\begin{aligned} V[\phi_2(\bar{y}_\alpha,1)] &= \alpha^2 B_1 + (1-\alpha)^2 B_2 + 2\alpha(1-\alpha)B_3 \\ &= V[\phi_2(\bar{y}_0,1)] + \alpha^2[B_1 + B_2 - 2B_3] - 2\alpha[B_2 - B_3] \end{aligned}$$

then $V[\phi_2(y_\alpha,1)] < V[\phi_2(\bar{y}_0,1)]$ if $\alpha < \left[\frac{2\mathrm{B}_5}{\mathrm{B}_4 + B_5}\right]$ and $\alpha\,\varepsilon\,[0,1]$

But from theorem 3.6 $\alpha_{opt} = \frac{B_5}{\left(B_4 + B_5\right)} < 1$ and hence the theorem.

4.0 Cost Aspect

Assume C_{0i}, C_{1i}, C_{2i} and C_{3i} are cost involved for the i^{th} strata where

(i) C_{0i} : Cost of including units n_i of the stratum i in sample n.

(ii) C_{1i} : Cost of conducting a pilot enquiry on X for post-stratification of n into n_i.

(iii) C_{2i} : Cost of collecting, editing and processing per n_i units of the response class.

(iv) C_{3i} : Cost of personal interview and processing information per n_i''' units from the non-response class.

Consider the cost functions for the i^{th} strata with total expected cost function

$$C_i^* = \left[(C_{*i})n_i + C_{2i}n_i^{'} + C_{3i}n_i^{'''}\right] and\ E(T_c) = \left(\frac{n}{N}\right)\sum_{i=1}^{k}\left\{(C_{*i})N_i + C_{2i}N_i^{'} + C_{3i}\frac{N_i^{''}}{f_i}\right\}$$

where $C_{*i} = (C_{0i} + C_{1i})$

Let V_0 be prefixed level of variance and for a constant $\lambda \neq 0$, we define a function

$$\delta = E(T_c) + \lambda\left[V\{\phi(\bar{y}_\alpha, 1) - V_0\}\right]$$

(4.0.0)

$$= E(T_c) + \lambda\left[\alpha^2 B_1 + (1-\alpha)^2 B_2 + 2\alpha(1-\alpha)B_3 - V_0\right] \qquad (4.0.1)$$

where λ is a Lagrange multiplier. The objective is to minimize the function δ (i.e. expected cost) subject to the prefix level of variance V_0 of the estimator $\phi(\bar{y}_\alpha, 1)$ [see Cochran (1977, pp. 97)]. To obtain cost-optimum f_i and n, we differentiate δ with respect to λ, f_i and n respectively one-by-one and equate to zero, then respective equations are :

$$\frac{\partial(\delta)}{\partial\lambda} = 0 \Rightarrow V_0 = \left[\alpha^2 B_1 + (1-\alpha)^2 B_2 + 2\alpha(1-\alpha)B_3\right] \qquad (4.1)$$

$$\frac{\partial(\delta)}{\partial f_i} = 0 \Rightarrow \frac{n}{N}\left[\frac{C_{3i}N_i^{''}}{f_i^2}\right] = \lambda\left[\left(\frac{N_i^{''}}{N_i}\right)W_i S_{2i}^2\right]\left[(2\alpha-\alpha^2)\left(\frac{1}{n}\right) + (1-\alpha^2)\ W_i^2 E\left(\frac{1}{n_i}\right)\right] \qquad (4.2)$$

$$\frac{\partial(\delta)}{\partial n} = 0 \Rightarrow \left(\frac{1}{N}\right)\sum_{i=1}^{k}\left[(C_{*i})N_i + C_{2i}N_i^{'} + C_{3i}\left(\frac{N_i^{''}}{f_i}\right)\right]$$

$$= \left(\frac{\lambda}{n^2}\right)\left[\alpha^2 F_1 + (1-\alpha)^2\sum_{i=1}^{k}\left\{\frac{NW_i(n-2) + (2N-n)}{n(N-1)}\right\}(A_i + S_i^2) + 2\alpha(1-\alpha)F_2\right] \qquad (4.3)$$

where, $F_{1=} F_2 + \sum_{i=1}^{k}\left\{\bar{Y}_i^2 - \frac{S_i^2}{N_i}\right\}\left\{\frac{NW_i(1-W_i)}{N-1}\right\} - \sum_{i}^{k}\sum_{j}^{k}\left\{\frac{NW_iW_j}{N-1}\right\}\bar{Y}_i\bar{Y}_j;\ F_2 = \sum_{i=1}^{k}W_i(A_i + S_i^2)$

From (4.2) we get

$$\lambda = \left(\frac{n}{N}\right)\left[\frac{C_{3i}N_i''}{f_i^2}\right] \Big/ \left[\left(\frac{N_i''}{N_i}\right)W_i S_{2i}^2\right]\left[(2\alpha-\alpha^2)\left(\frac{1}{n}\right) + (1-\alpha)^2 W_i^2\ .E\left(\frac{1}{n_i}\right)\right] \qquad (4.3.1)$$

Substitute (4.3.1) in (4.3) to eliminate λ. Now the expression of λ-eliminated form of (4.3) with (4.1) provides the optimum solution for f_i and n for a given choice of α. This could be using any standard technique of the solution of simultaneous equations described in the text of numerical analysis (for eg. Sastry (1998), pp 227). In all there will be (k+1) equations with λ eliminated (4.3) and (4.1), to be solved for (k+1) unknowns $f_1, f_{2,}, f_{3.........}f_k$ and n.

Remark 4.1: Incidentally, (4.3) is not a linear equation in n. So, at many occasions, the cost-optimal solution of f_i and n both may not exist. Therefore, we look for alternative form of optimal solution for f_i and n in order to minimize the expected cost.

5. Conditional Solution for F_i Using (4.2)

In light of remark 4.1, when cost optimal solution does not exist, we impose a restriction on f_i which, in combination with equation (4.2), provides a conditional solution for f_i while given n and α. Assume $\sum_{i=1}^{k} f_i = M$ where M is a prefixed constant related to the total size of sub-sample required for the revisit. From (4.2) we get

$$f_i = \sqrt{\frac{C_{3i}N_i^{''}}{\left[\left\{\lambda\left(\frac{N_i^{''}}{N_i}\right)W_i S_{2i}^2\right\}\left\{\left(2\alpha-\alpha^2\right)\left(\frac{1}{n}\right)+(1-\alpha)^2 W_i^2 E\left(\frac{1}{n_i}\right)\right\}\right]}} \tag{5.0}$$

Put this f_i into the restriction $\sum_{i=1}^{k} f_i = M$ then obtained value of λ is

$$\frac{1}{\sqrt{\lambda}} = \frac{M}{\sum_{i=1}^{k}\sqrt{\frac{C_{3i}N_i''}{\left[\left\{\lambda\left(\frac{N_i^{''}}{N_i}\right)W_i S_{2i}^2\right\}\left\{\left(2\alpha-\alpha^2\right)\left(\frac{1}{n}\right)+(1-\alpha)^2 W_i^2 E\left(\frac{1}{n_i}\right)\right\}\right]}}} \tag{5.1}$$

The substitution of (5.1) in (5.0) eliminates λ and provides conditional cost-optimal f_i for given n and α, denoted by f_i (n. α).

$$f_i(n,\alpha) = \frac{M\sqrt{\frac{C_{3i}N_i^{''}}{\left[\left\{\left(\frac{N_i^{''}}{N_i}\right)W_i S_{2i}^2\right\}\left\{\left(2\alpha-\alpha^2\right)\left(\frac{1}{n}\right)+(1-\alpha)^2 W_i^2 E\left(\frac{1}{n_i}\right)\right\}\right]}}}{\sum_{i=1}^{k}\sqrt{\frac{C_{3i}N_i^{''}}{\left[\left\{\left(\frac{N_i^{''}}{N_i}\right)W_i S_{2i}^2\right\}\left\{\left(2\alpha-\alpha^2\right)\left(\frac{1}{n}\right)+(1-\alpha)^2 W_i^2 E\left(\frac{1}{n_i}\right)\right\}\right]}}} \tag{5.2}$$

Note 5.1: Since (4.2) is an optimum equation obtained by $\frac{\partial(\delta)}{\partial f_i} = 0$, the solution f_i in (5.2) with condition $\sum f_i = M$ would reduce the function δ (i.e. expected cost) subject to a prefixed level of variance V_0. This approach is adopted in Cochran (1977, pp 96—99).

Remark 5.1 : Substitute $f_i^* = f_i(n,\alpha)$ in the ratio of equations (4.2) and (4.4) then the sample size, for given α, denoted by *n(* f_i^* *,* α *)*, is obtained.

6. Alternative Cost Strategy

When (4.1), (4.2) and (4.3) fail to provide cost-optimal solutions, and conditional solution (5.3) too doesnot serves the purpose then an alternative cost strategy is proposed as under :

Setp I: Be limited to *i*th strata only and assume in the population $Y_{ijm} \neq 0$ for ith strata

$Y_{i'jm} = 0$ for all $i' \neq i = 1,2,3,................k$. The expected cost is

$E(T_{ic}^*) = \left(\frac{n}{N}\right)\left[(C_{*i})N_i + C_{2i}N_i^{'} + C_{3i}\left(\frac{N_i^{''}}{f_i}\right)\right]$ with prefixed variance V$_{\text{oi.}}$ (6.0.0)

Setp II: With α consider estimator $\phi_2^*(\bar{y}_\alpha,1)$ for *i*th strata as

$$\phi_{2i}^*(\bar{y}_\alpha,1) = W_{i\alpha}\left[\phi_1(n_i^{'},n_i)\phi_2(\bar{y}_i^{'},1) + \phi_1(n_i^{''},n_i)\phi_2(\bar{y}_i^{'''},1)\right]$$

Such that $[\phi_2(\bar{y}_\alpha,1)] = \sum_{i=1}^{k}\phi_{2i}^*(\bar{\mathrm{y}}_\alpha,1)$ and $\phi_{2\mathrm{i}}^*(\bar{y}_\alpha,1)$ is unbiased for the population mean $(N^{-1})\sum_{j=1}^{2}\sum_{m=1}Y_{ijm}$ under the assumption of step I.

Setp III: Define a function δ_i^* with a constant $\lambda_i \neq 0$

$$\delta_i^* = E(T_{ic}^*) + \lambda_i\left[V\left\{\phi_{2i}^*(\bar{y}_\alpha,1)\right\} - V_{oi}\right]$$

where λ_i is Lagrange multiplier. We want to minimize the expected cost $E(T_{ic}^*)$ subject to condition of the prefixed level of variance *V*$_{oi.}$

Step IV: Differentiate δ_i^* with respect to λ_i, f_i, *n* and equate to zero. We get equations. The solution of these provides a cost-optimal choice of *f*$_i$ [as $f_\alpha^{(i)}$] and *n* [as $n_\alpha^{(i)}$] restricted to the *i*th strata.

Step V: Repeat the above procedure, from step I to IV, one-by-one for all the *k* strata.

Step VI: The outcome of step V provides k cost-optimal values:
$\left(f_\alpha^{(1)}, f_\alpha^{(2)}, f_\alpha^{(3)}............f_\alpha^{(k)}\right)$ and $\left(n_\alpha^{(1)}, n_\alpha^{(2)}, n_\alpha^{(3)},..............n_\alpha^{(k)}\right)$ of *f*$_i$ and *n*.

According to the requirement , any one of the following selection plans, may be chosen :

$$\mathbf{a} : n_{opt} = \sum_{i=1}^{k} n_{\alpha}^{(i)} \qquad \mathbf{b} : n_{\text{opt}} = Max\left(n_{\alpha}^{(1)}, n_{\alpha}^{(2)}, n_{\alpha}^{(3)} ....n_{\alpha}^{(k)}\right),$$

$$\mathbf{c} : n_{opt} = \text{interger value of } \left(k^{-1}\right)\sum_{i=1}^{k} n_{\alpha}^{(i)}$$

Setp VII : When only *r* strata, out of *k* *(k>r)*, provide optimal $n_{\alpha}^{(i)}$ then selection plan of step VI is to restrict among *k* = *r*.

7. Optimum Selection under Alternative Strategy

The variance of $\phi_{2i}^{*}\left(\bar{y}_{\alpha},1\right)$, for prefixed α, is

$$V\left[\phi_{2i}^{*}\left(\bar{y}_{\alpha},1\right)\right] = \alpha^{2}B_{1i}^{'} + (1-\alpha)^{2}B_{2i}^{'} + 2\alpha(1-\alpha)B_{3i}^{'} \qquad (7.0.0)$$

where $B_{1i}^{'} = \left[\left(\frac{W_i}{n}\right)\left(A_i + S_i^2\right) + V\left(p_i\right)\left\{\bar{Y}_i^2 - \frac{S_i^2}{N_i}\right\} - \frac{1}{N}W_i S_i^2\right]$

$$B_{2i}^{'} = \left[\left(W_i^2\right)E\left(\frac{1}{n_i}\right)\left(A_i + S_i^2\right) - \frac{1}{N}W_i S_i^2\right] \quad B_{3i}^{'} = \left[\left(\frac{W_i}{n}\right)\left(A_i + S_i^2\right) - \frac{1}{N}W_i S_i^2\right]$$

According to step III, using (6.0.0) and (7.0.0), we write

$$\delta_i^{*} = E\left[T_{ic}^{*}\right] + \lambda_i\left[V\left\{\phi_{2i}^{*}\left(\bar{y}_{\alpha},1\right)\right\} - V_{oi}\right]$$

$$= \left(\frac{n}{N}\right)\left[\left(C_{*i}\right)N_i + C_{2i}N_i' + C_{3i}\left(\frac{N_i''}{f_i}\right)\right] + \lambda_i\left[\alpha^2 B_{1i}^{'} + (1-\alpha)^2 B_{2i}^{'} + 2\alpha(1-\alpha)B_{3i}^{'} - V_{0i}\right] \quad (7.0.1)$$

Using setp IV , we get following equations :

$$\frac{\partial\left(\delta_i^{*}\right)}{\partial\lambda_i} = 0 \Rightarrow \quad V_{oi} = \alpha^2 B_{1i}^{'} + (1-\alpha)^2 B_{2i}^{'} + 2\alpha(1-\alpha)B_{3i}^{'} \qquad (7.1)$$

$$\frac{\partial\left(\delta_i^{*}\right)}{\partial f_i} = 0 \Rightarrow \quad \left[\frac{n^2 C_{3i}}{S_{2i}^2 f_i}\right] = \lambda_i\left[1 + \frac{(1-\alpha)^2(N-n)(1-W_i)}{n(N-1)W_i}\right] \qquad (7.2)$$

$$\frac{\partial\left(\delta_i^{*}\right)}{\partial n} = 0 \Rightarrow \quad \left(\frac{1}{N}\right)\left[\left(C_{*i}\right)N_i + C_{2i}N_i^{'} + C_{3i}\left(\frac{N_i^{''}}{f_i}\right)\right] = \left(\frac{\lambda_i}{n^2}\right)\left[\alpha^2\left(\bar{Y}_i^2 - \frac{S_i^2}{N_i}\right)\left\{\frac{NW_i(1-W_i)}{(N-1)}\right\}\right.$$

$$\left. + (1-\alpha)^2\left\{\frac{NW_i(n-2) + (2N-n)}{n(N-1)}\right\}\left(A_i + S_i^2\right) + \left(2\alpha - \alpha^2\right)\left\{W_i\left(A_i + S_i^2\right)\right\}\right] \qquad (7.3)$$

Note 7.1: The solution obtained by solving (7.1), (7.2) and (7.3) would minimize the function δ_i^{*} for the prefixed level of variance *V*$_{0i.}$

Remark 7.0: From (7.2), the value of λ_{i} is

$$\lambda_i = \left[\frac{\mathrm{n}^2 C_{3i}}{S_{2i}^2 f_i}\right]\left[1+\frac{(1-\alpha)^2(N-n)(1-W_i)}{n(N-1)W_i}\right]^{-1}$$

(7.3.1)

Substitute (7.3.1) in (7.3) and after some algebraic adjustments, we get

$$\{M_{1i}\}\left[n\left\{(N-1)W_i-(1-\alpha)^2(1-W_i)\right\}+(1-\alpha)^2 N(1-W_i)\right]$$
$$=\alpha^2\{M_{3i}\}+(1-\alpha)^2\{M_{4i}\}\{NW_i(n-2)+(2N-n)\}+\left(2\alpha-\alpha^2\right)\{nM_{2i}\} \quad (7.4)$$

where $M_{1i}=\left[(N-1)\left(f_i^{\,2}S_{2i}^2\right)\left\{(C_{*i})N_i+C_{2i}N_i^{'}+C_{3i}\left(\frac{N_i^{''}}{f_i}\right)\right\}\right]$

$$M_{2i}=\left[N(N-1)^2 W_i^{\,2} C_{3i}\left(A_i+S_i^2\right)\right]$$

$$M_{3i}=\left[N^2(N-1)W_i^{\,2}(1-W_i)\left(\bar{Y}_i^2-\frac{S_i^2}{N_i}\right)\right]; \qquad M_{4i}=\left[N(N-1)W_i C_{3i}\left(A_i+S_i^2\right)\right]$$

Remark 7.1: Equation (7.1) could be rewritten as

$$\left\{Q_{1i}-\alpha^2 Q_{2i}\right\}=\left(A_i+S_i^2\right)\left\{Q_{3i}+\left(1-\alpha^2\right)Q_{4i}\right\} \quad (7.5)$$

where $Q_{1i}=n^2(N-1)\left\{V_{oi}+\frac{1}{N}W_i S_i^2\right\}$; $Q_{2i}=\left[n(N-n)W_i(1-W_i)\left(\bar{Y}_i^2-\frac{S_i^2}{N_i}\right)\right]$

$$Q_{3i}=n(N-1)W_i\,,\ \ Q_{4i}=(N-n)(1-W_i)$$

Remark 7.2: *M*M_{1i}, M_{2i}, M_{3i} and M_{4i} are functions of f_i not containing n and α. Similarly, terms Q_{1i}, Q_{2i}, Q_{3i} and Q_{4i} are functions of n not containing f_i and α.

Remark 7.3: From (7.4), an alternative cost-optimal sample size of the i^th^ strata, in term of given f_i and α, is

$$n_\alpha^{(i)}(f_i)=\left[\frac{\alpha^2 M_{3i}+2(1-\alpha)(N-N_i)M_{4i}-(1-\alpha)^2 N(1-W_i)M_{1i}}{M_{1i}\left\{(N-1)W_i-(1-\alpha)^2(1-W_i)\right\}\ -\left(2\alpha-\alpha^2\right)M_{2i}}\right] \quad (7.6)$$

And similarly, from (7.5) in term of given n and α.

$$f_\alpha^{(i)}(n)=\left[1+\left(\frac{N_i^{'}}{N_i^{''}S_{2i}^2}\right)\left\{\frac{Q_{1i}-\alpha^2 Q_{2i}}{Q_{3i}+(1-\alpha)^2 Q_{4i}}\ -\ S_i^2\right\}\right] \quad (7.7)$$

Remark 7.4: The simultaneous solution of (7.6) and (7.7) provides $f_\alpha^{(i)}$ and $n_\alpha^{(i)}$ optimum values for the ith strata under alternative cost strategy.

Remark 7.5: According to step IV, n_{opt} = $\sum_{i=1}^{k} n_\alpha^{(i)}$ may be a good choice of sample size for given α if (a) is chosen.

Remark 7.6: While we assume $\bar{Y}_i^2 \approx \left(\frac{S_i^2}{N_i}\right)$ or assume a very small value of α or otherwise, terms M_{3i} and Q_{2i} could be ignored, then we have approximation:

$$\left[\, f_\alpha^{(i)}(n)\right]_{appx.} = \left[1+\left(\frac{N_i^{'}}{N_i^{"}S_{2i}^2}\right)\left\{\frac{Q_{1i}}{Q_{3i}+(1-\alpha)^2 Q_{4i}} - S_i^2\right\}\right] \tag{7.8}$$

$$\left[n_\alpha^{(i)}(f_i)\right]_{appx.} = \left[\frac{2(1-\alpha)(N-N_i)M_{4i}-(1-\alpha)^2 N\,(1-W_i)M_{1i}}{M_{1i}\left\{(N-1)W_i-(1-\alpha)^2\,(1-W_i)\right\} \; -\left(2\alpha-\alpha^2\right)M_{2i}}\right] \tag{7.9}$$

Remark 7.7: This is to point out that most preferred (or optimal) value of α is very small therefore the contribution of terms M_{3i} in (7.6) and Q_{3i} in (7.7) may be omitted. Hence (7.8) and (7.9) are good and useful approximations because of being free from prior knowledge of $\bar{Y}_i^2$.

8. Numerical Illustration

A data set (described in appendix A) is considered divide into four strata. Further, each strata has a response group (R) and a non-response group (NR). A random sample of size $n = 80$ is drawn and divided into random ($n_i^{'}$, $n_i^{"}$) units.

Table 8.1 (Cost Data)

Strata	Prefixed Cost				Prefixed Variance level (V_{oi})	Optimum f_i under fixed M = 25 at $\alpha = \alpha_{opt}$ using (5.1) (α_{opt} = 0.014)
	C_{oi}	C_{1i}	C_{2i}	C_{3i}		
I	0.010	0.100	0.500	2.01	4.29	f_1 =10
II	0.011	0.101	0.501	2.02	5.10	f_2 =06
III	0.012	0.102	0.502	2.03	5.60	f_3 =04
IV	0.013	0.103	0.503	2.04	4.90	f_4 =05

Table 8.2 Descriptive Statistics of Population and sample

Population Size N = 400 Mean $\overline{Y} = 81.65$ S^2 = 1823.05					
Type	Division	Strata			
		I	II	III	IV
Sizes of Strata	Response Class (R)	$N_1^{'}$ = 30 $n_1^{'}$ = 6	$N_2^{'}$ =40 $n_2^{'}$ = 8	$N_3^{'}$ = 60 $n_3^{'}$ = 9	$N_4^{'}$ = 70 $n_4^{'}$ = 10
	Non- Response Class (NR)	$N_1^{''}$ = 30 $n_1^{''}$ = 10	$N_2^{''}$ =40 $n_2^{''}$ = 10	$N_3^{''}$ = 60 $n_3^{''}$ = 13	$N_4^{''}$ = 70 $n_4^{''}$ = 14
	Total N$_i$	N$_1$= 60 n_1 = 16	N$_2$ = 80 n_2 =18	N$_3$ = 120 n_3 =22	N$_4$ = 140 n_4 =24
	$n_i^{'''}$	$n_1^{'''}$ = 3	$n_2^{'''}$ = 4	$n_3^{'''}$ = 4	$n_4^{'''}$ = 5
	f_i	f$_1$= 3.33	f$_2$ = 2.50	f$_3$ = 3.25	f$_4$ = 2.80
Mean	$\overline{Y}_i$	$\overline{Y}_1$ =13.48	$\overline{Y}_2$ =47.87	$\overline{Y}_3$ =85.50	$\overline{Y}_4$ =126.08
Popula-tion Mean square	Response Lass (R) (S_{1i}^2)	S_{11}^2 =67.06	S_{12}^2 =156.67	S_{13}^2 =228.01	S_{14}^2 =240.20
	Non- Response Class (NR) (S_{21}^2)	S_{21}^2 =62.97	S_{22}^2 =204.87	S_{23}^2 =248.79	S_{24}^2 =254.74
	Total S_i^2	S_1^2 =64.38	S_2^2 =180.46	S_3^2 =225.63	S_4^2 =245.74
Weight	W$_i$	W$_1$ = 0.15	W$_2$ = 0.20	W$_3$ = 0.30	W$_4$ = 0.35

Table 8.3 Optimum Variance And Preferable Range of α

α	α_{opt} = 0.014
$V[\phi_2(\bar{y}_\alpha,1)]$	(Variance) $_{opt}$ = 6.957 at α_{opt}
Preferable range of α	$\alpha\ \varepsilon$ (0.0, 0.028)

Illustration Under Alternative Cost-Strategy

Table 8.4 Calculation of Equations (7.6), (7.7), (7.8), (7.9)

Strata		Prefix α =0.014				
		Optimum n for given $f_i\left[n_\alpha^{(i)}(f_i)\right]$			Optimum f$_i$ for given $n\left[f_\alpha^{(i)}(n)\right]$	
		f$_i$=1.0	f$_i$=2.0	f$_i$=2.5	n=40	n=80
I	*	12	5	4	16.15	34.4
	**	12	5	4	16.17	34.4
II	*	7	3	2	4.75	9.81
	**	7	3	2	4.76	9.82
III	*	4	2	2	3.04	5.64
	**	4	2	2	3.05	5.65
IV	*	4	2	1	2.22	4.53
	**	4	2	1	2.22	4.54
TOTAL	*	27	12	9		
	**	27	12	9		
Max $\left[n_\alpha^i(f_i)\right]$		12	5	4		
		12	5	4		

* $\left[n_\alpha^{(i)}(f_i)\right]$ and $\left[n_\alpha^{(i)}(n)\right]$; ** $f_i\left[n_\alpha^{(i)}(f_i)\right]_{appx}$ and $f_i\left[n_\alpha^{(i)}(n)\right]_{appx}$

9.0 Conclusions

On recapitulation, we found that the proposed PSPNR scheme is useful for tacking the partial non-response in surveys. Since the post-stratification is a well proved effective scheme for practical solutions, the PSPNR scheme is even more close to the real life due to inclusion of non-response. The proposed estimation strategy is observed useful for estimating $\overline{Y}$ at suitable choices of constant $\alpha > 0$. The proposed one is also a general class of estimators which attains the minimum bound level of variance at the value $\alpha = \alpha_{\text{opt}}$. While $\alpha = 0$, the results of PSNR scheme of Shukla and Dubey (2001) appear as a particular case. In general, the preferable and advisable range of α is when $\alpha \in (0,2\alpha_{opt})$. The proof of several theorems and derivations of corollaries have established a justified basis for the scheme. This is to point out that a very small value of α, near to zero, contributed significantly in terms of gain in efficiency of the estimator. Under cost-function, a set of equations exists to provide cost-optimal selection of f_i and n. But, these do not ensure the optimality every time, on every data set. Therefore, another approach is derived using a linear constraint on the totality of f_i and found easy in determination of cost-optimal f_i. But, even after this, nothing is sure about cost-optimal n. To cope with, an alternative cost strategy is explored and found effective in the determination of f_i and n simultaneously in the presence of cost constraints along with non-response.

REFERENCE

AGRAWAL, M. C. and PANDA, K. B. (1995): On efficient estimation in post-stratification, Metron, 53, 3—4, 107—115.

AGRAWAL, M.C. and PANDA, K.B. (1993): An efficient estimator in post stratification, Metron, 51, 3—4, 179—187.

CASADY, R.J. and VALLIANT, R. (1993): Conditional properties of post-stratified estimators under normal theory, Survey Methodology, 19, 183—192.

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

GROVER and COUPER (1998): Non-response in household surveys, John Wiley and Sons, New York.

HANSEN, M. H. and HURWITZ, W. N. (1946): The problem of non-response in sample suveys, Jour. Amer. Stat. Asso., 41, 517—529.

HINDLE, R. L. and CHAMBERS, R. L. (1991): Non response imputation with multiple sources of non-response, Jour. Official Stat., 7, 169—179.

HOLT, D. and SMITH, T. M. F. (1979) : Post stratification, Jour. Roy. Stat. Soc., A, 142, 33—36.

JACKWAY, P. T. and BOYCE, R. A. (1987): Response including techniques of mail surveys, Aust. Jour. of Stat., 29, 255—263.

JAGERS, P. (1986) : Post stratification against bias in sampling, Int. Stat. Rev., 54, 159—167.

JAGERS, P., ODEN, A. and TRULSSON, L. (1985): Post stratification and ratio estimation, Int. Stat. Rev., 53, 221—238.

JONES, R. (1983): An experimentation of methods of adjusting for non-response to a mail surveys: a mail interview comparison in proceeding surveys, M.G. Madon and L. Olkin (eds), 3, 271—290.

KHARE, B. B. (1987): Allocation in stratified sampling in presence of non-response, Metron, 45, (I/II), 213—221.

KOTT, P. S. (1994) : A note on handling non-response error in surveys, John Wiley and Sons, New York.

LESSLER and KALSBEEK (1992): Non-response error in surveys, John Wiley and Sons, New York.

LITTLE, R. J. A. (1982) : Models for non-response error in surveys, Jour. Amer. Stat. Asso., 77, 237—250.

LITTLE, R. J. A. (1986): Surveys non-response adjustments for estimates of means, Int. Stat. Rev., 54(2), 139—157.

SARNDAL, C. E. and SWENSSON, B. (1987): A general view of estimation for two phase of selection with application to two phase sampling and non-response, Int. Stat. Rev., 55, 279—294.

SASTRY, S. S. (1998): Introductory methods of numerical analysis, Third edition, Printice-Hall Publication, New Delhi.

SHUKLA, D. and DUBEY, JAYANT (2001): Estimation in mail surveys under PSNR sampling scheme, Jour. Ind. Soc. Ag. Stat., 54(3), 288—302.

SMITH, T. M. F. (1991): Post-stratification, The Statisticians, 40, 323.

SUKHATME, P. V. SUKHATME, B. V. SUKHATME, S. and ASOK, C. (1984): Sampling theory of surveys with applications, Iowa State University Press, Indian Society of Agricultural Statistics, New Delhi.

WYWIAL (2001) : Stratification of population after sample selection, Statistics in Transition, 5(2), 327—348.

APPENDIX – A

Population (N = 400)

Strata – I

Response (R)

05	04	02	09	18	03	06	12	11	13	07	06
14	15	17	16	30	22	08	09	01	19	21	23
25	07	29	20	10	02						

Non- Response (NR)

04	22	14	26	08	30	20	15	06	08	02	04
06	08	10	22	14	06	18	20	01	13	25	07
19	21	15	15	27	09						

Strata – II

Response (R)

25	67	59	41	53	35	47	49	50	30	26	38
36	62	34	56	68	42	51	43	31	42	43	64
55	26	47	38	48	58	60	39	58	27	66	55
34	63	52	41								

Non- Response (NR)

41	26	36	64	66	66	70	48	49	30	50	62
25	46	28	40	62	34	63	62	25	28	47	24
55	40	69	47	50	53	55	58	61	66	53	49
67	44	42	70								

Strata – III

Response (R)

60	88	67	93	84	85	60	71	68	79	62	66
65	77	92	80	91	62	72	100	89	98	84	86
78	90	59	74	96	70	66	100	84	102	78	76
70	96	106	88	80	84	74	86	68	81	111	113
105	68	96	71	62	83	94	90	96	47	98	49

Non- Response (NR)

61	103	105	107	109	70	63	84	98	95	90	99	98	67	96
85	74	83	72	101	84	65	67	89	70	106	98	79	105	87
89	91	87	96	72	75	97	80	81	102	108	92	63	90	104
107	88	95	109	120	120	83	62	105	74	107	66	98	109	112

Strata – IV

Response (R)

100	101	112	123	104	108	146	127	148	108	149	122	112	133	110
145	126	117	108	129	120	124	124	136	148	130	102	134	136	148
121	133	105	127	149	131	133	135	137	149	149	131	122	143	144
135	146	107	108	135	150	112	134	106	128	108	106	109	102	140
131	138	143	111	115	150	137	138	129	146					

Non- Response (NR)

139	148	137	130	149	138	127	116	145	124	109	150	147	106	145
144	113	112	141	110	100	104	106	108	140	132	124	146	108	128
121	103	105	127	149	111	118	125	147	109	130	149	108	147	106
125	124	125	102	121	119	107	145	104	143	121	132	108	146	144
132	134	126	148	110	112	134	146	128	150					

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 89—96

A FAMILY OF ESTIMATORS FOR ESTIMATING POPULATION MEAN USING KNOWN CORRELATION COEFFICIENT IN TWO PHASE SAMPLING

Rajesh Singh, Pankaj Chauhan and Nirmala Sawan[1]

ABSTRACT

A family of estimators for estimating the population means of the variable under study, using two auxiliary variables in two-phase sampling, and known correlation coefficient of the second auxiliary character is proposed. It has been shown that the proposed estimators are more efficient than usual unbiased estimator, usual two-phase ratio estimator and Chand (1975) estimator. An empirical study is carried out to illustrate the performance of the constructed estimator.

Key words: Family of estimators, correlation coefficient, bias, mean-squared error.

1. Introduction

The ratio method of estimation is the well-known technique for estimating the population mean of a study character when the population mean of an auxiliary character is known. In the absence of the knowledge on the population mean of the auxiliary character we go for two phase (double) sampling. The two-phase sampling happens to be a powerful and cost effective (economical) technique for obtaining the reliable estimate in first phase sample for the unknown parameters of the auxiliary character and hence has an eminent role to play in survey sampling, for instance, see Hidirogolou and Sarndal (1995,98).

In order to construct an efficient estimator of the population mean of the auxiliary character in first phase sample, Chand (1975) gave a technique of chaining another auxiliary character (which is highly correlated with first auxiliary character but remotely correlated with the study character) with the first auxiliary by using the ratio estimator in the first phase sample. The estimator is

[1] School of Statistics, DAVV, Indore (M.P.), India; rsingh.stat@dauniv.ac.in .

known as chain ratio type estimator. Further, this work was extended by Kiregyera (1980, 1984), Upadhayaya and Singh (2001), Singh et.al. (2004) and many others by proposing several chain type ratio and regression estimators.

It is to be mentioned that the past association with experimental material might provide quite accurate value of the correlation coefficient (ρ) between y and x, for instance, see Sahai and Sahai (1985). On the other hand, in case y and x are the same variable on two previous occasions ρ may be easily obtained, see, Singh and Singh (1984).

In this paper, an attempt has been made to utilize the information on known correlation coefficient (ρ_{xz}) of the second auxiliary character through a simple transformation for estimating the population mean of auxiliary character more precisely in the first-phase (preliminary) sample. Under simple random sampling without replacement (SRSWOR) a family of estimators for estimating the population mean $\overline{Y}$ have been proposed and the expressions of bias and MSE, up to first order of approximation is derived. The performance of the proposed estimator is illustrated with the help of an empirical study.

2. Proposed estimator

Consider a finite population U={U$_1$,U$_2$,….,U$_N$}. Let y and x be the study and auxiliary variable, taking values y_i and x_i respectively for the i_{th} unit U_i. From the population U, a simple random sample of size n is drawn without replacement.

The two-phase sampling ratio estimator of $\overline{Y}$ is given by

$$t_1 = \overline{y}\left(\frac{\overline{x}'}{\overline{x}}\right) \tag{2.1}$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \overline{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x_i$.

Sometimes even if $\overline{X}$ is not known, information on a cheaply ascertainable variable z, closely related to x but compared to x remotely related to y, is available on all units of the population. For instance, while estimating the total yield of wheat in a village, the yield and area under the crop are likely to be unknown, but the total area of each farm may be known from village records or may be obtained at a low cost. Then y, x and z are respectively yield, area under wheat and area under cultivation. Thus assuming that the population mean $\overline{Z}$ of the variable z is known, Chand (1975) proposed a chain ratio-type estimator as

$$t_2 = \overline{y}\left(\frac{\overline{x}'}{\overline{x}}\right)\left(\frac{\overline{Z}}{\overline{z}'}\right) \tag{2.2}$$

where $\overline{z} = \frac{1}{n'}\sum_{i=1}^{n'} z_i$.

Let $v_i = z_i + \rho_{xz}(i = 1,2....,N)$ so that $\overline{v}' = \overline{z}' + \rho_{xz}$ is the sample mean of the transformed variable v in the first-phase sample and $\overline{V} = \overline{Z} + \rho_{xz}$ is the corresponding population mean. We suggest a transformed chain ratio type estimator for the population mean $\overline{Y}$ as

$$t_3 = \overline{y}\left(\frac{\overline{x}'}{\overline{x}}\right)\left(\frac{\overline{Z} + \rho_{xz}}{\overline{z}' + \rho_{xz}}\right) \tag{2.3}$$

To obtain the bias and MSE of t_3, we write

$$\overline{y} = \overline{Y}(1 + e_0), \overline{x} = \overline{X}(1 + e_1), \overline{x}' = \overline{X}(1 + e_1'), \overline{z}' = \overline{Z}(1 + e_2)$$

such that

$$E(e_0) = E(e_1) = E(e_1') = E(e_2) = 0.$$

and

$$\begin{aligned}
&E(e_0^2) = f_1 C_y^2, E(e_1^2) = f_1 C_x^2,\\
&E(e_1'^2) = f_2 C_x^2, E(e_2^2) = f_2 C_z^2,\\
&E(e_0 e_1) = f_1 \rho_{xy} C_y C_x, E(e_0 e_1') = f_2 \rho_{xy} C_y C_x,\\
&E(e_0 e_2) = f_2 \rho_{yz} C_y C_z, E(e_1 e_1') = f_2 C_x^2,\\
&E(e_1 e_2) = f_2 \rho_{xz} C_x C_z, E(e_1' e_2) = f_2 \rho_{xz} C_x C_z,
\end{aligned}$$

Where

$$f_1 = \left(\frac{1}{n} - \frac{1}{N}\right), f_2 = \left(\frac{1}{n'} - \frac{1}{N}\right),$$

$$C_y^2 = \frac{S_y^2}{\overline{Y}^2}, C_x^2 = \frac{S_x^2}{\overline{X}^2}, C_z^2 = \frac{S_z^2}{\overline{Z}^2},$$

$$\rho_{xy} = \frac{S_{xy}}{S_x S_y}, \rho_{xz} = \frac{S_{xz}}{S_x S_z}, \rho_{yz} = \frac{S_{yz}}{S_y S_z},$$

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2, S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2, S_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \overline{Z})^2,$$

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y}), S_{xz} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})(z_i - \overline{Z}),$$

$$S_{yz} = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{Y})(z_i - \overline{Z}).$$

Expressing (2.3) in terms of e's, we have-

$$t = \overline{Y}(1+e_0)(1+e_1')(1+e_1)^{-1}(1+\theta e_2)^{-1} \tag{2.4}$$

where

$$\theta = \frac{\overline{Z}}{\overline{Z}+\rho_{xz}} \tag{2.5}$$

We assume that $|e_0|<1, |\theta e_2|<1$, so that $(1+e_1)^{-1}$ and $(1+\theta e_2)^{-1}$ are expandable.

Expanding the right hand side of (2.2) and retaining terms up to second powers of e's, we have

$$t_3 = \overline{Y}[1+e_0-e_1+e_1'-\theta e_2+\theta^2 e_2^2+\theta e_1 e_2+e_1^2-\theta e_1' e_2-e_1 e_1' \\ -\theta e_0 e_2-e_0 e_1+e_0 e'] \tag{2.6}$$

Taking expectations on both sides in (2.6) and then subtracting $\overline{Y}$ from both sides, we get the bias of the estimator t₃, up to the first order of approximation as

$$B(t_3) = \overline{Y}\{f_3 C_x^2(1+\theta K_{zx}-K_{yx})+f_2\theta C_z^2(\theta-K_{yz})\} \tag{2.7}$$

where

$$f_3 = \left(\frac{1}{n}-\frac{1}{n'}\right),$$

$$K_{yx} = \rho_{yx}\frac{C_y}{C_x}, K_{zx} = \rho_{zx}\frac{C_z}{C_x}, K_{yz} = \rho_{yz}\frac{C_y}{C_z}$$

From (2.6)we have

$$(t_3-\overline{Y}) \cong \overline{Y}\{e_0-e_1+e_1'-\theta e_2\}$$

(2.8)

Squaring both sides of (2.8) and then taking expectation, we get the MSE of the estimator t₃, up to the first order of approximation, as

$$MSE(t_3) = \overline{Y}^2\{f_1 C_y^2+f_2\theta C_x^2(\theta-2K_{yz})+f_3 C_x^2(1-2K_{yx})\} \tag{2.9}$$

3. The suggested family of estimators

We define a general family of estimators of $\overline{Y}$ as

$$t_4 = \overline{y}\left(\frac{\overline{x}'}{\overline{x}}\right)^{\alpha_1}\left(\frac{\overline{Z}+\rho_{xz}}{\overline{z}'+\rho_{xz}}\right)^{\alpha_2}$$

(3.1)

where α_1 and α_2 are unknown constants to be suitably determined.

Using the technique of section 2, the MSE of the estimator t_4 up to the first order of approximation can be written as

$$MSE(t_4)=\bar{Y}^2\left[f_1C_y^2+f_2\alpha_2\theta C_z^2\left(\alpha_2\theta-2K_{yz}\right)+f_3\alpha_1C_x^2\left(\alpha_1-2K_{yx}\right)\right] \quad (3.2)$$

Minimization of (3.2) with respect to α_1 and α_2 yields their optimum values as

$$\left.\begin{aligned}\alpha_1&=K_{yx}=\alpha_{10}\,(\text{say})\\ \alpha_2&=\frac{K_{yz}}{\theta}=\alpha_{20}\,(\text{say})\end{aligned}\right\} \quad (3.3)$$

Substitution of (3.3) in (3.2) yields the minimum value of MSE (t_4) as

$$\min.MSE(t_4)=\bar{Y}^2C_y^2\left[f_1-f_2\rho_{yz}^2-f_3\rho_{yx}^2\right] \quad (3.4)$$

Remark 3.1

It has to be mentioned that the proposed family of estimators t_4 will attain the min. MSE in (3.4) only when the exact optimum values of α_{10} and α_{20} at (3.3) of α_1 and α_2 are known. The optimum values of α_{10} and α_{20} are functions of unknown population parameters such as K_{yx}, K_{yz} and the known constant θ. The values of K_{yx} and K_{yz} can be assessed quite accurately from the past data or experiences gathered in due course of time, for instance, see Srivastava (1967) and Murthy (1967,pp.96-99). Also the prior values of K_{yx} and K_{yz} may be either obtained on the basis of the information from a most recent survey or by conducting a pilot survey, see, Lui (1990,p.3805).

4. Efficiency comparisons

To compare the efficiency of proposed family of estimators t_4 with estimators t_1,t_2,t_3, and MSE's of these estimators up to the first order of approximations are as

$$MSE(t_1)=\bar{Y}^2\left[f_1C_y^2+f_3C_x^2\left(1-2K_{yx}\right)\right] \quad (4.1)$$

$$MSE(t_2)=\bar{Y}^2\left[f_1C_y^2+f_2C_z^2\left(1-2K_{yz}\right)+f_3C_x^2\left(1-2K_{yx}\right)\right] \quad (4.2)$$

$$MSE(t_3)=\bar{Y}^2\left[f_1C_y^2+f_2\theta C_z^2\left(\theta-2K_{yz}\right)+f_3C_x^2\left(1-2K_{yx}\right)\right] \quad (4.3)$$

$$\min.MSE(t_4)=\bar{Y}^2C_y^2\left[f_1-f_2\rho_{yz}^2-f_3\rho_{yx}^2\right] \quad (4.4)$$

$$MSE(t_1)-\min.MSE(t_4)=\bar{Y}^2\left[f_2C_y^2\rho_{yz}^2+f_3\left(C_x-\rho_{yx}C_y\right)^2\right]\geq 0 \quad (4.5)$$

$$MSE(t_2)-\min.MSE(t_4)=\bar{Y}^2\left[f_2\left(C_z-\rho_{yz}C_y\right)^2+f_3\left(C_x-\rho_{yx}C_y\right)^2\right]\geq 0 \quad (4.6)$$

$$\mathrm{MSE}(t_2)-\min.\mathrm{MSE}(t_4)=\overline{Y}^2\left[f_2\left(\theta C_z-\rho_{yz}C_y\right)^2+f_3\left(C_x-\rho_{yx}C_y\right)^2\right]\geq 0 \quad (4.7)$$

The above results lead us to conclude that the proposed estimator t_4 outperforms the other potentially competing estimators t_1, t_2, and t_3.

5. Empirical Study

To see the relative performance of various estimators discussed in the paper we considered two population data used earlier by others. These populations are discussed below:

Population I: Anderson (1958)

N=25
y: head length of second son,
x: head length of first son,
z: head breadth of first son.

$\overline{Y}=183.84$, $\overline{X}=185.72$, $\overline{Z}=151.12$, $C_y=0.0546$, $C_x=0.0526$,

$C_z=0.0488$, $\rho_{yx}=0.7108$, $\rho_{yz}=0.6932$, $\rho_{xz}=0.7346$.

Take n=7 and n'=10.

Population II: Sukhatme and Sukhatme (1977), p.185

N=170.
y: Area under wheat in 1937.
x: Cultivated area in 1931.
z: Area under wheat in 1936.

$\overline{Z}=218.4118, C_y=0.744310, C_z=0.756439, \rho_{yx}=0.816875,$

$\rho_{yz}=0.929922$ $\rho_{xz}=0.83$

Take n'=10 and n=5.

The percent relative efficiency (PRE's) of the estimators $\overline{y}, t_1, t_2, t_3$ and t_4 with respect to $\overline{y}$ have been computed and compiled in table 5.1.

Table 5.1. Percent Relative efficiency of various estimators with respect to $\overline{y}$

Estimator	PRE (.) with respect to $\overline{y}$	
	Population I	Population II
$\overline{y}$	100	100
t_1	123.00	164.21
t_2	180.82	393.29
t_3	181.12	393.71
t_4	196.40	398.26

Table 5.1 clearly indicates that the suggested estimators t_3 and t_4 are more efficient than $\bar{y}$, usual two-phase ratio estimator and Chand (1975) estimator t_2.

REFERENCES:

ANDERSON, T.W. (1958): An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, Inc., New York.

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

HIDIROGLOU, M.A. and SARNDAL, C.E. (1995): Use of auxiliary information for two-phase sampling. Proceedings of the Section on Survey Research Methods, American Statistical Association, Vol.II, 873—878.

HIDIROGLOU, M.A. and SARNDAL, C.E. (1998): Use of auxiliary information for two-phase sampling. Survey Methodology, 24(1), 11—20.

KIREGYERA, B. (1980): A chain ratio type estimators in finite population double sampling using two auxiliary variables. Metrika, 17, 217—223.

KIREGYERA, B. (1984): Regression type estimators using two auxiliary variables and the model of double sampling from finite populations. Metrika, 31, 215—226.

LUI K.J. (1990), Modified product estimators of finite population mean in finite sampling. Communications in Statistics, Theory and Methods, 19(10), 3799—3807.

MURTHY, M. N. (1967): Sampling theory and Methods. Statistical Publishing Society, Calcutta, India.

SAHAI, A. and SAHAI, A. (1985): On efficient use of auxiliary information. Journal of Statistical planning and Inference, 12, 203—212.

SINGH, H.P., UPADHYAYA, L.N. and P.CHANDRA (2004): general families of estimators for estimating population mean using two auxiliary variables in two-phase sampling. Statistics in transition, 6,7, 1055—1077.

SINGH, R.K. and SINGH, G.(1984): A class of estimators for population variance using information on two auxiliary variates. Aligarh Journal of Statistics, 3&4, 43—49.

SRIVASTAVA, S.K. (1967): An estimator using auxiliary information in sample survey. Cal. Statist.Assoc.Bull, 16, 121—131.

SUKHATME, P.V. and SUKHATME, B.V. (1977): Sampling theory of surveys with applications. Iowa State University Press, Ames, U.S.A.

UPADHYAYA, L.N. and SINGH, G.N. (2001): Chain type estimators using transformed auxiliary variable in two-phase sampling. To appear in A.M.S.E. France.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 97—110

PREDICTION OF CHILD SURVIVAL IN INDIA USING DEVELOPED COX PH MODEL: A UTILITY FOR HEALTH POLICY PROGRAMMERS

Rajvir Singh, Shahina Begum, R.K. Ahuja,
Prem Chandra, S. N. Dwivedi[1]

ABSTRACT

Prediction of any health indicator from a developed model can be achieved accurately only when it comes from a properly validated model. The aim of this study was to develop a model using Cox Proportional Hazard method on child survival data and the prediction of survival probabilities of a child at different points of time. A data set of 2118 children of first birth order was drawn from the National Family Health Survey (NFHS, 1992-93), Uttar Pradesh (UP), India. In multivariate analysis variables such as breastfeeding, immunization, premature birth, antenatal care, type of house and father's education were found to be significantly associated with child survival. The validation index shrinkage coefficient was 97 percent, indicating only 3 percent lack of fit in the model, and Somer's D rank correlation (D_{xy}) was –0.65, indicating good correlation between the log hazard and the observed survival time. Best improvement in child survival was found by 7% in first month, 9% in three months and 11% in twelve months when the variables such as breastfeeding, immunization and the mothers received antenatal care during pregnancy used altogether in developed model for prediction. The above findings may be useful for those who are involved in health policy programs for improving child survival and for better public health management.

Key words: Cox PH model, Bootstrapping, Calibration, Discrimination, Prediction.

1. Introduction

Child survival is an important public health indicator of our country (Nath et al, 1994), and also plays a vital role in fertility level. One aim of the National Population Policy is to reduce the infant mortality rate (IMR) to 28 by 2012 per

[1] Department of Biostatistics, All India Institute of Medical Sciences, New Delhi, India

1000 live births and subsequently a reduction in fertility (Planning Commission report, 2004).

It is well known fact that biological factors are responsible for the early neonatal mortality, whereas, the postnatal mortality may be reduced by an intervention. For example, in India, the leading causes of deaths among children are diarrhea and acute respiratory infection (ARI), which can be prevented by making people aware about how to manage the aforesaid diseases through mass media (WHO report, 2005). Women, who are exposed to mass media, are utilizing health services efficiently. Motivation of antenatal care, institutional delivery, breastfeeding, complete immunization of children through mass media may be a great means to reduce the child mortality. Socio-economic characteristics of parents also affect the child survival as poor economic condition may be adversely related to survival due to limited resources and health facilities available. So, in view of the above facts the aim of the study was to find out the factors associated with child survival as well as prediction of child survival probabilities at different points of time using important factors from the developed model.

The development of techniques and strategies to determine the best model to predict an outcome in a large data set with a large number of potential predictive variables has become an increasingly important topic in the last few years. Statistical models developed for prediction need validation and validation of a model can be performed through various available methods such as Data-splitting, Cross validation and Re-sampling methods to check predictive accuracy of a developed model. Efron (1983) has shown that cross validation is relatively inefficient due to high variation of accuracy estimates when the validation process is repeated while data – splitting is far worse due to indices of accuracy vary greatly with different splits. Re- sampling (Bootstrapping) provides nearly unbiased estimates of predictive accuracy that are of relatively low variance where fewer model fits are required than the cross-validation and it has an additional advantage of using the entire data set for model development. Data are too precious to waste (Rocker 1991). Harrell et al (1984) has stated that both reliability and discrimination should be considered in assessing predictive accuracy. Dividing the test population into subgroups according to the predicted outcomes and compared observed against predicted outcomes for each subgroup commonly tests reliability. Predictive discrimination indices have been used for reliability in the study.

In India, studies based on development of a model through Cox PH method and validate the model applying bootstrap method that uses calibration and discrimination indices with prediction of survival probabilities of a child at different points of time are rare. It encourages us to do this study. The information will have great help to policy programmers who are involved in formulating various health policies for better public health management.

2. Material and methods

Data Set: To develop the model, a data set of 2118 children of first birth order born in the 4 years of preceding the survey of NFHS-1992-93 on child survival from Uttar Pradesh (UP), India, was considered. The sample design adopted for the National Family Health Survey (NFHS-1) was Systematic, two-stage stratified random sample of households. First, Census Enumeration Blocks(PSU's) and then followed by selection of household in each of the selected PSU's. The details of the methodology and the objectives of NFHS are given in NFHS report (IIPS, UP, India, 1995).

Variables used under the study:

(a) *Dependent Variable*- Duration of child survival to deaths with status alive and dead.

(b) *Independent Variables*- The following demographic and clinical variables were used as independent variables in the statistical analysis: (i) religion/caste (SC/ST Hindu/ other Hindu/ non-Hindu) (ii) place of residence (rural/ urban) (iii) mother's education (illiterate/ primary/ middle/ ≥ high school) (iv) breast feeding (no/yes) (v) sex of index child (male/ female) (vi) mother's occupation (not working /working) (vii) father's occupation (not working/ working) (viii) type of house (kuchha / semi pucca + pucca) (ix) media exposure (no/ yes) (x) distance from a primary health center (≥ 2 kms / < 2 kms) (xi) antenatal care (no/ yes) (xii) immunization of child (no/ yes) (xiii) place of delivery (at home/ at hospital) (xiv) complication at delivery (no & yes) (xv) premature birth (no & yes) and (xvi) age of mother at index child (in complete years).

In view of its non-linear relationship, mother's age at index child was squared and also added to the model to fulfill the assumption of linearity. Data regarding children from multiple births i.e. triplets or twins, only the first child was included for the analysis. All the variables mentioned above were in the form of fixed covariates with fixed effect, other than age of the mother, which was a time varying covariate with fixed effect.

Statistical Method:

(a) Model Development:

Multivariate Cox regression method was been used and assumptions like linearity, proportionality, interaction effect and multicollinearity were checked through exploratory analysis.

Multivariate Cox Proportional Hazard model has been written in the form of the hazard at time t, $\lambda(t)$, as follows:

$$\lambda(t) = \lambda_0(t) * \exp(\beta_1 X_1 + \beta_2 X_{2+} \ldots\ldots\ldots\ldots\ldots + \beta_p X_p)$$

where, $\lambda_0(t)$ is called baseline hazard function and $\beta_1,\ \beta_2, \ldots\ldots\ldots\ldots\ldots \beta_p$ are unknown regression coefficients(Kleinbaum, 1996).

Regression coefficients were estimated by maximizing likelihood function through Cox Proportional hazards model analysis for all the covariates mentioned above. The stepwise method was used to select variables for inclusion or exclusion from the model in a sequential fashion. For this, a forward with a test for backward elimination was used with probability levels for entry and removal as 0.15 and 0.10 respectively.

(b) Model Validation:

Efron (1983) presented re-sampling method and detailed information is reported in Efron and Tibshirani (1993). A random sample of size 2118 children with replacement from the original sample was called a bootstrap replication, taking T_j , δ_j, X_{j1}, ……X_{jp}, where T_j was the survival time, δ_j was the censoring indicator and X_{j1}, ……,X_{jp} were the covariates of the j^{th} child (j=1,….., 2118). Two hundred bootstrap replications were taken as 200 independent samples. The two important components of predictive accuracy are calibration and discrimination and both the components were used for the validation through bootstrapping method (Harrell et al, 1996). Calibration refers to extent of bias in the model. For example, if the average predicted proportion of children experiencing death within a considered time, for a group of children is as 0.15 which is equal to the observed proportion of the deaths, the predictions are known to be well calibrated. To access calibration accuracy, the suggested methods mainly include calibration curve and shrinkage coefficient.

(i) Calibration Curve:

To obtain a calibration curve, the data were divided into various groups based on the intervals of predicted survival probabilities obtained through developed models at an appropriate time t_j (at 12 months in this study) in such a way that there were 150 children in each group. For each group, the average predicted survival probability (obtained through arithmetic mean of individual probabilities) versus observed survival probability obtained through Kaplan Meier method (Kaplan Meier, 1958) at t_j plotted. The bootstrapping procedure (with 200 re-samples) was thus used to estimate the optimism about the predicted survival probability estimates from the developed Cox model to track the corresponding Kaplan Meier survival probability estimates at time t_j, stratified by grouping children in subsets with 150 children per interval of predicted survival at t_j.

(ii) Shrinkage Coefficient:

The shrinkage coefficient was used to quantify the lack of fit of the model. The heuristic shrinkage estimator has been defined as:

$$\Upsilon = (\text{model } \chi^2 \text{- p })/ \text{ model } \chi^2$$

Where p is the number of regression parameters including all non-linear and interaction effects and the model χ^2 is the total likelihood ratio χ^2 statistics (Van Houwelingen & Le Cessie, 1990).

(iii) Somer's D Rank Correlation (D_{xy}):

Another component of discrimination aspect of the validation of model, was measured through Somer's D_{xy} rank correlation between the predicted log hazard and the observed survival time using 2(C-0.5) formula, where C is the concordance index and was calculated using the following steps (Harrell et al., 1982).

a) Consider all possible pairs of children at least one of them has experienced death;
b) If the predicted survival time was larger for the child whose survival period was longer, the predictions for that pair were said to be concordant with the outcomes;
c) If one child experienced death and the other was known to have not experienced at least till the survival time of the first, the second child was assumed to out live the first;
d) When predicted survivals were identical for a pair of children, 0.5 rather than 1 was added to the count of concordant pairs in the numerator of C. In this case, one was still added to the denominator of C (such children pairs are still considered usable); and
e) Children pair was unusable if both the children experience death at the same time, or if one experienced death and the other still not experienced death, but the child was not followed long enough to determine whether she/he would outlive the one who experienced death.

All the three measures of validation were calculated using bootstrapping method. The detailed description for calculating of the indices and steps for calculating these indices are described by Harrell et al. (1996).

(c) Predictions using the Developed Models:

The developed model in the present study may be used to help the policy planners in better public health management. For this purpose, important clues may be obtained through predicted survival probabilities for a particular variable by holding all other variables at their mean levels. Similarly, these probabilities for a particular combination of variables may also be calculated by holding all remaining variables at their means. The prediction of survival probabilities using Cox hazards model involved various steps as described below (Dickson et al, 1989):

Step 1: The exponential expression of the Cox model, also known as 'Risk score' and generally denoted by R, is defined as follows:

$$R = \beta_1X_1 + \beta_2X_2 + ... + \beta pXp$$

in which X_1 , X_2 , ..., Xp are the considered levels of p predictor variables and β_1 ,β_2 , ….βp are respective unknown regression coefficients. Thus, using maximum likelihood estimates of regression coefficients for the model being used and substituting the observed values of the covariates for each individual, risk score was obtained for every child included in the data analysis. The arithmetic mean of these risk scores provided an average risk score R_1. Obviously, R_1 would be constant for a given data set.

Step 2: Another risk score R_2 was obtained again by using the equation, substituting the estimated values of the regression coefficients and changed levels of the selected variable/set of variables for every child but retaining other variables at their mean levels. This value may, obviously, vary from person to person as a result of variation in the levels of selected variable/set of variables.

Step 3: $S_0(t)$, the baseline survival probabilities at different points of time for a person with average risk score R_1 were worked out using Kaplan Meier method. Thus, $S_0(t)$ at a given point of time was nothing but the survival probability obtained through Kaplan Meier method at that point of time.

Step 4: Therefore, gain in survival probability after adjustment in relation to considered levels of selected covariates, was obtained by

$$S(t) = S_0(t)^{\exp(R_2 - R_1)}$$

A complete analysis under the present study was accomplished with the help of various packages namely BMDP version 7.0, University of California, 1992; S-plus 4.0, 1988-97, Mathsoft Inc., Seatle, WA 98109-3044 USA. These packages were either available in the Department of Biostatistics, All India Institute of Medical Sciences (AIIMS), New Delhi or used after due permission of the concerned authority. Predicted probabilities of survival were performed through Macros on Excel 2000.

3. Results

Table 1 describes the percentage of deaths according to the selected background characteristics. It is found that children belonged to categories such as SC/ST Hindu, rural area, kuchha house, not breast fed, not immunized, illiterate mother, mothers were not exposed to mass media, mothers did not receive antenatal care during pregnancy, mothers had experienced complications during delivery, premature birth, and nearest primary health center was more than 2 kms, experienced proportionately higher deaths than their counterparts.

Table 1. Percent distribution of deaths by selected background chaarakteristics

Variables	Children (Death%)
Religion/caste	
SC/ST Hindu	426 (16.9)
Other Hindu	1366 (12.0)
Non-Hindu	326 (12.9)
Place of residence	
Rural	1699 (14.5)
Urban	419 (07.6)
Mother’s education	
Illiterate	1467 (15.3)
Primary	227 (09.7)

Variables	Children (Death%)
Middle	151 (11.3)
High school comp & above	273 (05.1)
Breastfeeding	
No	161 (72.1)
Yes	1957(08.3)
Sex of index child	
Male	1046 (12.8)
Female	1072 (13.4)
Mother’s occupation	
Not Working	1861 (13.4)
Working	257 (11.3)
Father’s occupation	
Not working	131 (10.7)
Working	1987 (13.3)
Father’s education	
Illiterate	614 (16.3)
Primary	312 (16.4)
Middle	364 (13.2)
High School com & above	828 (09.5)
Type of house	
Kuchha	1105 (15.4)
SemiPucca+Pucca	1013 (09.9)
Media exposure	
No	1206 (15.6)
Yes	912 (09.9)
Distance of primary health	
center	1541 (14.6)
≥2 kms	577 (09.2)
<2kms	
Antenatal care	985 (19.9)
No	1133 (07.2)
Yes	
Immunization of the child	1384 (18.8)
No	734 (02.5)
Yes	
Place of delivery	841 (15.6)
At home	1277 (11.5)
At hospital	
Complication at delivery	1837 (12.4)
No	281 (18.2)
Yes	
Premature birth	2031 (11.9)
No	87 (42.5)
Yes	

Variables	Children (Death%)
Mean age of mother at birth	19.5±2.98
Total	2118 (13.1)

Univariate Analysis:

Table 2 describes the results under univariate analysis that are in the form of Relative Risk, and its 95% confidence interval i.e unadjusted rate ratio (RR) along with 95% C.I. Table suggests that children belonged to categories such as non-Hindu, urban area, mothers educated up to high school & above, breastfed, fathers educated up to high school & above, semi-pucca or pucca house, mothers exposed to mass media, nearest primary health center was less than 2kms from the residence, mothers received antenatal care during pregnancy, and children were immunized were found important for protection of the child survival whereas delivery at home, complications during delivery and premature birth were found risk factors for child death.

Table 2. Unadjusted Relative Risk (RR) of selected Demographic characteristics through Cox PH

Variables	RR	C.I. 95%
Religion/caste		
SC/ST Hindu	1.00	
Other Hindu	0.70	0.54 – 0.93
Non-Hindu	0.75	0.51 – 1.12
Place of residence		
Rural	1.00	
Urban	0.52	0.36 – 0.75
Mother's education		
Illiterate	1.00	
Primary	0.62	0.40 – 0.96
Middle	0.73	0.45 – 1.20
High school comp & above	0.32	0.19 – 0.56
Breastfeeding		
No	1.00	
Yes	0.05	0.04 – 0.07
Sex of index child		
Male	1.00	
Female	1.04	0.82 – 1.32
Mother's occupation		
Not Working	1.00	
Working	0.82	0.56 – 1.21
Father's occupation		
Not working	1.00	

Variables	**RR**	**C.I. 95%**
Working	1.20	0.70 – 2.06
Father's education		
Illiterate	1.00	
Primary	0.99	0.71 – 1.39
Middle	0.80	0.57 – 1.13
High School com & above	0.57	0.43 – 0.77
Type of house		
Kuchha	1.00	
SemiPucca+Pucca	0.69	0.54 – 0.87
Media exposure		
No	1.00	
Yes	0.62	0.49 – 0.80
Distance of primary health center	1.00	
≥2 kms	0.61	0.46 – 0.83
<2kms		
Antenatal care	1.00	
No	0.35	0.27 – 0.45
Yes		
Immunization of the child	1.00	
No	0.11	0.07 – 0.18
Yes		
Place of delivery	1.00	
At hospital	1.36	1.08 – 1.72
At home		
Complication at delivery	1.00	
No	1.54	1.14 – 2.08
Yes		
Premature birth	1.00	
No	4.70	3.32 – 6.65
Yes		

Multivariate Analysis:

Due to non-linear relation of mother's age at index child, square of mother's age at index child was considered in the model. All the covariates satisfied the proportionality assumption; therefore, consideration of each covariate in the data analysis was done in the form of their fixed effects. Subsets of variables entered into the final Cox PH model related to the survival of children. Some of the variables entered into the model partially. For meaningful presentation, partially entered variables were considered fully in the presentation of final model. The variables breastfeeding, immunization, antenatal care, type of house and father's education are found protective factors for child survival whereas premature birth is found a risk factor. Risk ratio and their 95% C.I. are presented in Table 3.

Table 3. Adjusted Relative Risk (RR) of selected demographic characteristics through Cox PH

Variable	R.R.	95% C.I.
Mother's age at index child (cont.)	0.93	0.72 – 1.21
Mother's age^2 at index child (cont.)	1.00	0.99 – 1.00
Breastfeeding	1.00	
No	0.07	0.05 – 0.09
Yes		
Father's education	1.00	
Illiterate	1.35	0.60 – 1.89
Primary	1.06	0.74 – 1.50
Middle	0.87	0.64 – 1.19
≥High school		
Type of house	1.00	
Kuchha	0.78	0.61 – 1.01
Pucca		
Immunization	1.00	
No	0.17	0.11 – 0.28
Yes		
Premature birth	1.00	
No	2.97	2.06 – 4.28
Yes		
Antenatal care	1.00	
No	0.56	0.43 – 0.74
Yes		

Validation of the Model:

The bootstrapping procedure (with 200 re-samples) was used to estimate the optimism regarding the ability of the predicted survival probability estimates from the developed Cox PH model. Shrinkage coefficient was obtained through bootstrapping to quantify lack of fit the model whereas discrimination aspect of the validation of the model was measured by Somer's D_{xy} rank correlation between the log hazard and the observed survival time through bootstrapping. A calibration curve was also performed to see the extent of bias in the model. Shrinkage coefficient was 97 percent, indicating only 3 percent lack of fit in the model whereas Somer's D rank correlation D_{xy} was –0.65, indicating very good correlation between the log hazard and the observed survival time, are presented in Table 4.

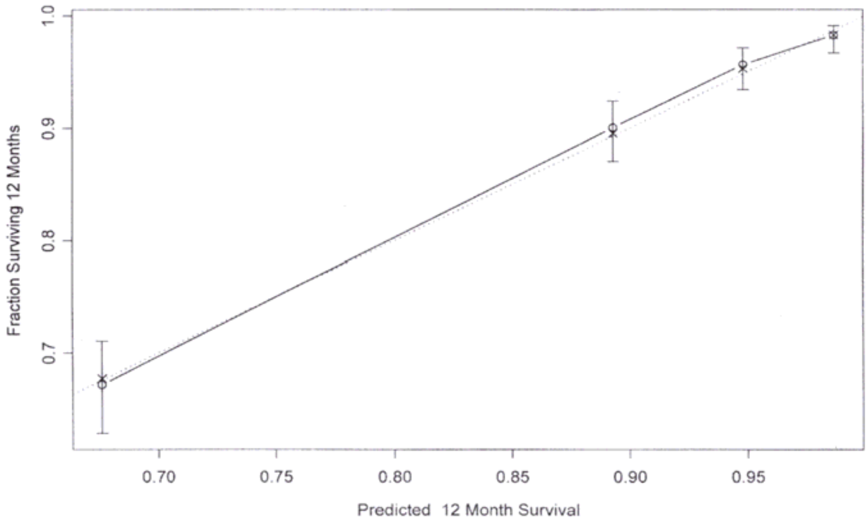
Table 4. Validity Indices of the Model

	Index Original	Training	Test	Optimism	Index Corrected	Re-sample
Birth order I						
Shrinkage Coefficient	1.00	1.00	0.97	0.03	0.97	200
Dxy	-0.66	-0.66	-0.66	-0.01	-0.65	200

D_{xy}: Somer`s D-rank correlation.

Similar pattern was also suggested by calibration curve, which was used to assess the accuracy of the model for prediction where dots correspond to apparent predictive accuracy and X marks the bootstrap corrected estimates, as shown in Figure 1.

Figure 1. Calibration Curve.



Predictions of survival probabilities at various points of time:

Survivals probabilities in relation to R1 were listed under first row where as those related to R2 were listed in successive rows. Thus, the difference between the two probabilities provides gain as a result of proposed changes in the levels of selected variable/set of variables. Children those who are breastfed, immunized and their mother received antenatal care during pregnancy have highest improvement in child survival. It is 7 percent (0.9815-0.9094=0.0721) in first month, 9 percent in three month, and 11 percent in 12 months respectively. Other predicted survival probabilities are presented in Table 5.

Table 5. Predicted Probabilities of Child Survival.

Characteristics	*Probability of survival at months*				
	1	3	6	9	12
Model (R_1)	0.9094	0.8893	0.8893	0.8793	0.8692
Breast feeding	0.9264	0.9098	0.9098	0.9015	0.8932
Immunization	0.9700	0.9631	0.9631	0.9596	0.9560
Antenatal care	0.9301	0.9143	0.9144	0.9065	0.8985
Father's education high school & above	0.9206	0.9028	0.9028	0.8939	0.8850
Breastfeeding + immunization	0.9758	0.9701	0.9701	0.9673	0.9644
Breastfeeding+ antenatal care	0.9433	0.9304	0.9304	0.9239	0.9174
Breastfeeding+ immunization + antenatal care	0.9815	0.9771	0.9771	0.9749	0.9727
Breastfeeding+ father's education (high school & above)	0.9355	0.9209	0.9209	0.9136	0.9063
Immunization+ father's education (High school and above)	0.9738	0.9677	0.9677	0.9647	0.9616

4. Discussion and conclusions

The factors affecting the child survival are essential components for population dynamics and can never be ignored. Breastfeeding, immunization, premature birth, antenatal care, and type of house and father's education are found statistically significant. There is a need to improve child survival by controlling the factors that adversely effect the child survival in developing countries like India and thus consequently the population dynamics.

Harrell et al (1984) has proposed a general c-index of predictive discrimination to measure the ability of a model to predict survival of patients having coronary heart disease. This index is a measure of concordance between the predicted and the observed outcome and can be applied to different statistical models. The advantage of using bootstrapping procedure for internal validation is that the entire data set could be used for development of model. Also, this procedure provides unbiased estimates of predictive accuracy that are of relatively low variance. The validation indices i.e. shrinkage coefficient and D rank correlation is 97 percent, indicating only 3 percent lack of fit in the model and Dxy= –0.65, indicating good correlation between the log hazard and the observed survival time respectively. Therefore, developed model is good enough to describe the predictive accuracy for the data set. In the present study we have demonstrated the predicted survival probabilities at different points of time. These probabilities have shown improvement in child survival. The above results may be useful for those who are involved in health policy programs for improving child survival policy and for better public health management. Thus, it will be of great help to policy planners who are involved in various health policies for better public health management.

Acknowledgement

We would like to thank Prof. Frank E. Harrell Jr, Division of Biostatistics and Epidemiology, Department of Health Evaluation Sciences, School of Medicine, University of Virginia, USA for the permission to use Library in S-Plus (Harrell, 1994). We are also thankful to Mr. V.A.V Guru Prasad, Data Entry Operator in the Department of Biostatistics, AIIMS, New Delhi for his valuable contribution in formatting the manuscript.

Corresponding Author:
Dr. Rajvir Singh
Senior Scientist
Department of Biostatistics
All India Institute of Medical Sciences (AIIMS)
New Delhi-110029, INDIA
Phone: 91-011-26593240; Fax: 91-011-26588663
Email:rajvir_aiims@yahoo.com

REFERENCES

BMDP 7.0 Statistical Software, Inc. (1992), 1440 Sepulveda Boulevard Suite 316, Los Angeles, CA 90025.

EFRON, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation, *Journal of the American Statistical Association*, 78, 316—331.

EFRON, B., TIBSHIRANI RJ. (1993). *An Introduction to Bootstrap*, New York: Chapman and Hall.

HARRELL, FE JR., CALIFF RM., PRYOR DB., LEE KL., ROSATI RA. (1982). Evaluating the yield of medical tests, *Journal of the American Medical Association*, 247, 2543—2546.

HARRELL, FE JR, LEE, KL, CALIFF, RM et al. (1984). *Regression modeling strategies for improved prognostic prediction.* Statistics in Medicine, 3, 143—152.

HARRELL FE JR. (1994). *Design: S-Plus functions for Biostatistica/ /Epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by sorting enhanced model design attributes in the fit.* Programs available from statlib@lib.stat.cmu.edu.

HARRELL, FE JR., LEE KL., MARK DB. (1996). Tutorial in Biostatistics Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine*, 15, 361—387.

Human Development Report 2004, Planning Commission, India,2004.

International for Population Sciences (IIPS). (1995). *National Family Health Survey, Uttar Pradesh (UP), India, 1992—1993*. Bombay: IIPS.

KLEINBAUM DG. (1996). *Survival Analysis: Statistics in Health Sciences*, New York, Inc. Springer-Verlag.

MEIER KAPLAN E. (1958). Non parametric estimation from incomplete observations. *J Amer Statist Assoc*, 53, 457—481.

NATH DC, KENNETH CL, & TALUKDAR. (1994). *Most recent birth intervals I a non-contraception Indian Population: An evolutionary ecology approach.* J Biosoc Sci: 32, 343—354.

ROLLAND DICKSON, E., PATRICA M GRAMBSCH, THOMAS R., FLEMING LLOYD DISHER AND ALICE LANGWORTHY. (1989). Prognosis in primary biliary cirrohsis: Model for decision making. *Hepatology*, 10(1), 1—7.

ROCKER, EB. (1991). Prediction error and its estimation for subset-selected models. *Technometrics*, 33, 459—468.

The World Health Report, 2005. Make Every Mother Count. WHO, 121 Geneva, 27, Switzerland.

VAN HOUWELINGEN, JC., LE CESSIE S. (1990). Predictive value of statistical models, *Statistics in Medicine*, 8, 1303—1325.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 111—123

ON SYNTHETIC AND COMPOSITE ESTIMATORS FOR SMALL AREA ESTIMATION UNDER LAHIRI — MIDZUNO SAMPLING SCHEME

G.C. Tikkiwal[1] and K. K. Pandey[2]

ABSTRACT

This paper studies performance of synthetic ratio estimator and composite estimator, which is a weighted sum of direct and synthetic ratio estimators, under Lahiri – Midzuno (L-M) sampling scheme. The synthetic estimator under L-M scheme is unbiased and consistent if the assumption of synthetic estimator is satisfied. Further, this paper compares performance of the synthetic and composite estimators empirically under L-M and SRSWOR schemes for estimating crop acreage for small domains. The study shows that both the estimators perform better under L-M scheme as having comparatively smaller absolute relative biases and relative standard errors.

Key words: Composite estimators, Synthetic ratio estimators, Small domains, Lahiri – Midzuno sampling design, SICURE model .

1. Introduction

Gonzalez and Wakesberg (1973) and Schaible, Brock, Casady and Schnack (1977) compare errors of synthetic and direct estimators for standard Metropolitan Statistical Areas and Counties of U.S.A. The authors of both the papers conclude that when in small domains sample sizes are relatively small the synthetic estimator out performs the simple direct, whereas, when sample sizes are large the direct outperforms the synthetic. These results suggest that a weighted sum of these two estimators, known as composite estimator, can provide an alternative to choosing one over the other. Tikkiwal, B.D. and Tikkiwal G.C. (1998) and Tikkiwal G.C. and Ghiya (2004) define a generalized class of composite estimators for small domains using auxiliary variable, under simple

[1] Deptt. of Mathematics & Statistics, J.N.V. University, Jodhpur-342 011, India;
E-mail : gctikkiwal@yahoo.com
[2] Banasthali Vidyapith, P.O. Box Banasthali Vidyapith - 304022, India

random sampling and stratified random sampling schemes. Further, the authors compare the relative performance of the estimators belonging to the generalized class with the corresponding direct and synthetic estimators. The study suggests the use of composite estimator, combining direct and synthetic ratio estimators, as it has smaller relative bias and standard error.

In this paper we study the performance of synthetic ratio estimator, and composite estimator belonging to the generalized class of composite estimators for small domains, under Lahiri-Midzuno scheme of sampling. The study suggests that the estimators perform better under Lahiri-Midzuno scheme of sampling than, under SRSWOR scheme.

2. Notations

Suppose that a finite population U = (1, ... , i, ... , N) is divided into 'A' non overlapping small domains Ua of given size Na (a = 1, ... , A) for which estimates are required. We denote the characteristic under study by 'y'. We further assume that the auxiliary information is available and denote this by 'x'. A random sample s of size n is selected through Lahri-Midzuno sampling scheme (1951, 52) from population U such that na units in the sample 's' comes from small domain Ua (a = 1, ... , A).

Consequently,

$$\sum_{a=1}^{A} N_a = N \quad and \quad \sum_{a=1}^{A} n_a = n$$

We denote the various population and sample means for characteristics Z = X, Y by

$\overline{Z}$ = mean of the population based on N observations.

$\overline{Z}a$ = population mean of domain 'a' based on Na observations.

$\overline{z}$ = mean of the sample 's' based on n observations.

$\overline{z}a$ = sample mean of domain 'a' based on na observations.

Also, the various mean squares and coefficient of variations of the population 'U' for characteristics Z are denoted by

$$S_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(z_i - \overline{Z}\right)^2, \qquad C_z = \frac{S_z}{\overline{Z}}$$

The coefficient of covariance between X and Y is denoted by

$$C_{xy} = \frac{S_{xy}}{\overline{X}\,\overline{Y}}$$

where,

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right)$$

The corresponding various mean squares and coefficient of variations of small domains Ua are denoted by

$$S_{z_a}^2 = \frac{1}{N_a - 1}\sum_{i=1}^{N_a}\left(Z_{a_i} - \overline{Z}_a\right)^2 \; , \; C_{z_a} = \frac{S_{z_a}}{\overline{Z}_a} \quad \text{and } C_{x_a y_a} = \frac{S_{x_a y_a}}{\overline{X}_a \overline{Y}_a}$$

where,

$$S_{x_a y_a} = \frac{1}{N_a - 1}\sum_{i=1}^{N_a}\left(y_{a_i} - \overline{Y}_a\right)\left(x_{a_i} - \overline{X}_a\right)$$

and zai (a = 1, ... , A and i = 1, ... , Na) denote the i-th observation of the small domain 'a' for the characteristic Z = X, Y.

3. Synthetic Ratio Estimator

We consider here synthetic ratio estimator of population mean $\overline{Y}_a$, based on auxiliary information 'x' under Lahiri-Midzuno sampling scheme, as described in previous section. The synthetic ratio estimator of population mean $\overline{Y}_a$ of small area 'a' is defined as follows :

$$\overline{y}_{syn,a} = \frac{\overline{y}}{\overline{x}}\overline{X}_a \tag{3.1}$$

This estimator may be heavily biased unless the following assumption is satisfied

$$\left(\overline{Y}_a \; / \; \overline{X}_a\right) \doteq \left(\overline{Y} \; / \; \overline{X}\right) \quad \text{for all a} \in \text{A} \tag{3.2}$$

3.1.Bias and Mean Square Error

Under Lahiri-Midzuno sampling scheme

$$\begin{aligned} E\left(\overline{y}_{syn,a}\right) &= E\left(\frac{\overline{y}}{\overline{x}}\overline{X}_a\right) \\ &= \frac{\overline{X}_a}{\overline{X}} E\left(\frac{\overline{y}}{\overline{x}}\overline{X}\right) \\ &= \frac{\overline{X}_a}{\overline{X}}\overline{Y} \end{aligned} \tag{3.3}$$

Therefore, design bias of $\overline{y}_{syn,a}$ is

$$B\left(\bar{y}_{syn,a}\right)=\left(\frac{\bar{Y}}{\bar{\bar{X}}}\,\bar{X}_a-\bar{Y}_a\right)=B_1\,(say) \tag{3.4}$$

The mean square error of $\bar{y}_{syn,a}$ is given by

$$\mathrm{MSE}\left(\bar{y}_{syn,a}\right)=\frac{\bar{X}_a^2}{\bar{\bar{X}}^2}V\left(\frac{\bar{y}}{\bar{\bar{x}}}\bar{X}\right)+B_1^{\;2}$$

$$=\frac{\bar{X}_a^2}{\bar{\bar{X}}}\left[\frac{1}{\binom{N}{n}}\sum_c\left(\frac{\bar{y}^2}{\bar{\bar{x}}}\right)_c-\bar{Y}^2\right]+B_1^{\;2} \tag{3.5}$$

where, $\sum_c$ stands for summation over all possible samples.

Remark 3.1

The above expression of MSE $\left(\bar{y}_{syn,a}\right)$ is not in analytical form.

Remark 3.2

If the synthetic assumption given in Eq. (3.2) satisfies then the $B_1=B\left(\bar{y}_{syn,a}\right)=0$ and hence consistent estimator of MSE $\left(\bar{y}_{syn,a}\right)$ is given by

$$\mathrm{mse}\left(\bar{y}_{syn,a}\right)=\frac{\bar{X}_a^{\;2}}{\bar{X}^2}\,v\left(\bar{y}_R\right)$$

$$=\frac{\bar{X}_a^{\;2}}{\bar{X}^2}\left[\bar{y}_R^2-\frac{\bar{X}}{\bar{\bar{x}}}\left\{\bar{y}^{\;2}-\left(\frac{1}{n}-\frac{1}{N}\right)s_y^2\right\}\right] \tag{3.6}$$

where, $\bar{y}_R=\frac{\bar{y}}{\bar{\bar{x}}}\bar{X}$

3.2 Comparison under SRSWOR

The expressions of Bias and Mean square error of synthetic ratio estimator under SRSWOR scheme is given by Tikkiwal & Ghiya (2000), while discussing the properties of generalized class of synthetic estimator, as under

$$B_2=B\left(\bar{y}_{syn,a}\right)=\frac{\bar{Y}}{\bar{\bar{X}}}\bar{X}_a\left[1+\frac{N-n}{Nn}\left(C_x^2-C_{xy}\right)\right]-\bar{Y}_a \tag{3.7}$$

and

$$MSE\left(\bar{y}_{syn,a}\right)=\left(\frac{\bar{Y}}{\bar{X}}\bar{X}_a\right)^2\left[1+\frac{N-n}{Nn}\left\{3C_x^2+C_y^2-4C_{xy}\right\}\right]$$

$$-2\bar{Y}_a\left(\frac{\bar{Y}}{\bar{X}}\bar{X}_a\right)\left[1+\frac{N-n}{Nn}\left(C_x^2-C_{xy}\right)\right]+\bar{Y}_a^2 \qquad (3.8)$$

Comparing the expression of biases B1 and B2 of $\bar{y}_{syn,a}$ under L-M & SRSWOR schemes, we get from Eqs. (3.4) and (3.7)

$$B_2-B_1=\frac{N-n}{Nn}\frac{\bar{Y}}{\bar{X}}\bar{X}_a\left(C_x^2-C_{xy}\right)$$

(3.9)

$$So, \qquad B_2\geq B_1 \;\; if$$

$$C_x^2-C_{xy}\geq 0 \;\Rightarrow \rho\frac{C_y}{C_x}\leq 1$$

Remark 3.3

If the synthetic assumption given in Eq. (3.2) satisfies then the expression of bias B2 given in Eq. (3.7) reduces to

$$B_2=\frac{N-n}{Nn}\left(C_x^2-C_{xy}\right) \qquad (3.10)$$

That is, B2 ≠ 0 even if synthetic assumption is satisfied. Whereas under this condition B1 = 0.

Remark 3.4

If the synthetic assumption is satisfied than the expressions of MSE $\left(\bar{y}_{syn,a}\right)$ given in Eq. (3.5) and Eq. (3.8) reduces respectively to

$$\mathrm{M}_1=\mathrm{MSE}\left(\bar{\mathrm{y}}_{\mathrm{syn,a}}\right)=\frac{\bar{\mathrm{X}}_{\mathrm{a}}^2}{\bar{\mathrm{X}}}\left[\frac{1}{\binom{\mathrm{N}}{\mathrm{n}}}\sum_{\mathrm{c}}\left(\frac{\bar{\mathrm{y}}^2}{\bar{\mathrm{x}}}\right)_{\mathrm{c}}-\bar{\mathrm{Y}}^2\right] \qquad (3.11)$$

and

$$M_2=MSE\left(\bar{y}_{syn,a}\right)=\frac{N-n}{Nn}\left(C_x^2+C_y^2-2\,C_{xy}\right) \qquad (3.12)$$

It may be noted here that the expression M1 under L-M design is still not in analytical form, therefore, a theoretical comparison of expressions M1 and M2 is not possible.

4. Composite Estimator

We consider in this section a composite estimator $\left(\overline{y}_{c,a}\right)$, which is a combination of direct ratio $\left(\overline{y}_{d,a}\right)$ and synthetic ratio $\left(\overline{y}_{syn,a}\right)$ estimators, under L-M design.

That is,

$$\overline{y}_{c,a} = w_a\ \overline{y}_{d,a} + \left(1 - w_a\right)\overline{y}_{syn,a} \tag{4.1}$$

Where, $\overline{y}_{d,a} = \frac{\overline{y}_a}{\overline{x}_a}\overline{X}_a$ and wa is suitably chosen constant.

It may noted that the estimator $\overline{y}_{d,a}$ is design biased whereas $\overline{y}_{syn,a}$ is design unbiased estimator under L-M design. When domain size Na is known, an estimator of the approximate variance of $\overline{y}_{d,a}$ under L-M design is given by

$$\hat{V}(\overline{y}_{d,a}) = \left\{\frac{\overline{X}_a}{\sum\limits_{i\in s_a}(x_i/\pi_i)}\right\}^2 \sum\sum_{i,j\in s_a}\Delta_{ij}\left(\frac{y_i - \hat{B}_a x_i}{\pi_i}\right)\left(\frac{y_j - \hat{B}_a x_j}{\pi_j}\right) \tag{4.2}$$

Where, sa = Ua ∩ s. That is, sa is the subset of sample s that falls in the domain Ua.

$$\Delta_{ij} = \frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}, \qquad \hat{B}_a = \frac{\sum\limits_{i\in s_a}(y_i/\pi_i)}{\sum\limits_{i\in s_a}(x_i/\pi_i)}$$

$$\pi_i = \frac{N-n}{N-1}p_i + \frac{n-1}{N-1}$$

$$\pi_{ij} = \begin{cases}\left(\frac{n-1}{N-1}\right)\left\{\left(\frac{N-n}{N-2}\right)(p_i + p_j) + \left(\frac{n-2}{N-2}\right)\right\} & forall\quad i \neq j\\[2ex] \frac{N-n}{N-1}p_i + \frac{n-1}{N-1} & forall\quad i = j\end{cases}$$

and $p_i = x_i / X$
[cf. Sarandal et al. (1992) Eq. (10.6.3)]

4.1 Estimation of Weights

The optimum values w'_a of wa may be obtained by minimizing the mean square error of $\bar{y}_{c,a}$ with respect to wa and it is given by

$$w'_a = \frac{MSE(\bar{y}_{syn,a}) - E(\bar{y}_{d,a} - \bar{Y}_a)(\bar{y}_{syn,a} - \bar{Y}_a)}{MSE(\bar{y}_{d,a}) + MSE(\bar{y}_{syn,a}) - 2E(\bar{y}_{d,a} - \bar{Y}_a)(\bar{y}_{syn,a} - \bar{Y}_a)}$$

Under the assumption that $E(\bar{y}_{d,a} - \bar{Y}_a)(\bar{y}_{syn,a} - \bar{Y}_a)$ is small relative to $MSE(\bar{y}_{syn,a})$, the w_a reduced to

$$w^*_a = \frac{MSE(\bar{y}_{syn,a})}{MSE(\bar{y}_{d,a}) + MSE(\bar{y}_{syn,a})} \tag{4.3}$$

Since $\bar{y}_{syn,a}$ is not an unbiased estimator, therefore, an unbiased estimator of MSE($\bar{y}_{syn,a}$) under the assumption that Cov($\bar{y}_{d,a}$, $\bar{y}_{syn,a}$) = 0, is given by [cf. Rao (2003), Eq. 4.2.12)]

$$mse(\bar{y}_{syn,a}) = (\bar{y}_{syn,a} - \bar{y}_{d,a})2 - v(\bar{y}_{d,a}) \tag{4.4}$$

Now, using Eq. (3.6) the weights w^*_a can be estimated as follows:

$$\hat{w}^*_a = \frac{mse(\bar{y}_{syn,a})}{(\bar{y}_{syn,a} - \bar{y}_{d,a})^2} \tag{4.5}$$

But this estimator of w^*_a can be very unstable. Schaible (1978) proposes an average weighting scheme based on several variables or "similar" areas or both, to overcome this difficulty. In our empirical study presented in next section, we take average of $\hat{w}^*_a$ over "similar" areas.

5. Crop Acreage Estimation for Small Domains — A Simulation Study

In this section we compare the relative performance of $\bar{y}_{syn,a}$ and $\bar{y}_{c,a}$ under L-M and SRSWOR sampling schemes, through a simulation study, as the mean square errors of $\bar{y}_{d,a}$ and $\bar{y}_{syn,a}$, and hence of $\bar{y}_{c,a}$ are not in analytical form. This we do by taking up the State of Rajasthan, one of the states in India, for our case study.

5.1 Existing methodology for estimation

In order to improve timelines and quality of crop acreage statistics, a scheme known as Timely Reporting Scheme (TRS) has been in vogue since early seventies in most of the States of India. The TRS has the objective of providing quick and reliable estimates of crop acreage statistics and there-by production of the principle crops during each agricultural season. Under the scheme the Patwari (Village Accountant) is required to collect acreage statistics on a priority basis in a 20 percent sample of villages, selected by stratified linear systematic sampling design taking Tehsil (a sub-division of the District) as a stratum. These statistics are further used to provide state level estimates using direct estimators viz. Unbiased (based on sample mean) and ratio estimators.

The performance of both the estimators in the State of Rajasthan, like in other states, is satisfactory at state level, as the sampling error is within 5 percent. However, the sampling error of both the estimators increases considerably, when they are used for estimating acreage statistics of various principle crops even at district level, what to speak of levels lower than a district. For example, the sampling error of direct ratio estimator for Kharif crops (the crop sown in June-July and harvested in October- November every year) of Jodhpur district (of Rajasthan State) for the agricultural season 1991-92 varies approximately between 6 to 68 percent. Therefore, there is need to use indirect estimators at district and lower levels for decentralized planning and other purposes like crop insurance, bank loan to farmers.

5.2 Details of the simulation study

For collection of revenue and administrative purposes, the State of Rajasthan, like most of the other states of India, is divided into a number of districts.

Further, each district is divided into a number of Tehsils and each Tehsil is also divided into a number of Inspector Land Revenue Circles (ILRCs). Each

ILRC consists of a number of villages. For the present study, we take ILRCs as small domains.

In the simulation study, we undertake the problem of crop acreage estimation for all Inspector Land Revenue Circles (ILRCs) of Jodhpur Tehsil of Rajasthan. They are seven in number and these ILRCs contain respectively 29, 44, 32, 30, 33, 40 and 44 villages. These ILRCs are small domains from the TRS point of view. The crop under consideration is Bajra (Indian corn or millet) for the agriculture season 1993-94. The bajra crop acreage for agriculture season 1992-93 is taken as the auxiliary characteristic x.

We consider the following estimators of population total Ta of small domain 'a' for a = 1,2,..., 7

Synthetic ratio estimator $t_{1,a} = N_a \left(\frac{\bar{y}}{\bar{x}} \right) \bar{X}_a$

and

Composite estimator t2,a = Na $\bar{y}_{c,a}$

To assess the relative performance of the estimators under two different sampling schemes viz. L-M and SRSWOR, their Absolute Relative Bias (ARB) and Simulated relative standard error (Srse) are calculated for each ILRC as follows :

$$\mathrm{ARB}(t_{k,a}) = \frac{\left| \frac{1}{500} \sum_{s=1}^{500} t_{k,a}^{s} - T_a \right|}{T_a} \mathrm{x}100 \qquad (5.2.1)$$

and

$$\mathrm{Srse}(t_{k,a}) = \frac{\sqrt{\mathrm{SMSE}(t_{k,a})}}{T_a} \mathrm{x}100 \qquad (5.2.2)$$

where

$$\mathrm{SMSE}(t_{k,a}) = \frac{1}{500} \sum_{s=1}^{500} (t_{k,a}^{s} - T_a)^2 \qquad (5.2.3)$$

for k = 1, 2 and a = 1, ..., 7

5.3 Results

We present the results of ARB (in %) synthetic ratio estimator $\left(\bar{y}_{syn,a}\right)$ in Table 5.3.2 and of composite estimator $\left(\bar{y}_{c,a}\right)$ in Table 5.3.3. The Srse (in %) of composite estimator are presented in Table 5.3.4 and Table 5.3.5. The total number of villages in Jodhpur Tehsil is 252. We take n = 25, 50, 63 and 76 i.e. samples, approximately, of 10%, 20%, 25% and 30% villages. It may be noted that a sample of 20% villages are presently adopted in TRS. Before simulation, we first examined the validity of synthetic assumption given in Eq. (3.1) . The results of these are presented in Table 5.3.1. From this we note that the assumption closely meets for ILRCs (3), (5) and (7) . Where as, the assumption deviate moderately for ILRC (4) , and deviate considerably for ILRCs (1) and (2). In case of composite estimators, we estimate the weights for each small domain using Eq. (4.5) but for estimating total of small domains of ILRCs (3), (5) and (7) we take average of $\hat{w}_a^*$ over these domains, being "similar".

We observe from Table 5.3.2 to Table 5.3.5 (specially for n=50 i.e. a sample of 20% villages that is being selected under TRS scheme) that both the estimators perform well in ILRCs (3) , (5) and (7) under both the sampling schemes, where synthetic assumption closely satisfied . But the composite estimator $\left(\bar{y}_{c,a}\right)$ performs better than the synthetic ratio estimator. The ARB of both the estimators under consideration is much smaller in case of L-M scheme than in case of SRSWOR. Also the Srse of both the estimators reduces under L-M scheme and is about 5%. Here we suggest that when the synthetic assumption is not valid one should look for other types of estimators such as those obtained through the SICURE MODEL [B.D.Tikkiwal (1993)] or presented in Ghosh and Rao (1994).

Table 5.3.1 Absolute Differences (Relative) under Synthetic Assumption of Synthetic Ratio Estimator for Various ILRCs.

ILRC	$\bar{Y}_a / \bar{X}_a$	$\bar{Y} / \bar{X}$	$\left\lfloor \left\lvert \bar{Y}_a / \bar{X}_a - \bar{Y} / \bar{X} \right\rvert / \left(\bar{Y}_a / \bar{X}_a\right) \right\rfloor X100$
(1)	.7303	.8675	18.17
(2)	.7402	.8675	17.19
(3)	.8663	.8675	0.13
(4)	.9416	.8675	7.86
(5)	.8595	.8675	0.91
(6)	.9666	.8675	10.25
(7)	.8815	.8675	1.58

Source: own calculations.

Table 5.3.2 Absolute Relative Biases (in %) of Synthetic Ratio Estimator under L-M and SRSWOR Designs for different sample sizes.

ILRC	For n = 25		For n = 50		For n = 63		For n = 76	
	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR
(1)	17.06	18.01	15.88	17.90	14.01	17.68	13.65	18.02
(2)	18.79	19.65	9.01	19.5	8.94	19.32	7.05	19.66
(3)	0.59	0.62	0.016	0.72	0.011	0.895	0.008	0.61
(4)	1.06	8.57	1.28	8.66	1.13	8.81	1.11	8.55
(5)	0.132	0.156	0.021	0.55	0.014	0.11	0.012	0.17
(6)	8.34	10.94	7.79	11.03	5.83	11.18	5.14	10.93
(7)	0.96	1.12	0.34	1.02	0.26	0.85	0.22	1.13

Source: own calculations.

Table 5.3.3 Absolute Relative Biases (in %) of Composite Estimator under L-M and SRSWOR Designs for different sample sizes.

ILRC	For n = 25		For n = 50		For n = 63		For n = 76	
	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR
(1)	9.68	10.72	8.10	8.40	7.65	8.01	4.63	5.18
(2)	11.53	12.6	8.76	10.02	5.43	7.60	5.15	6.42
(3)	0.36	1.98	0.009	0.50	0.006	0.53	.008	0.28
(4)	6.97	7.57	1.19	6.30	2.19	5.20	2.08	4.73
(5)	0.105	0.01	0.019	0.38	0.008	0.29	0.007	0.41
(6)	7.14	7.60	3.45	4.60	4.19	4.60	3.01	3.51
(7)	0.83	1.53	0.24	1.20	0.18	1.01	0.17	1.40

Source: own calculations.

Table 5.3.4 Simulated Relative Standard Error (Srse in %) of Synthetic Ratio Estimator under L-M and SRSWOR Designs for different sample sizes.

ILRC	For n = 25		For n = 50		For n = 63		For n = 76	
	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR
(1)	19.87	20.15	18.34	19.11	17.67	18.07	19.78	18.67
(2)	21.34	22.34	19.39	20.67	19.81	20.01	18.54	19.98
(3)	7.15	7.67	5.01	5.71	5.03	5.15	5.51	5.01
(4)	10.13	11.08	9.87	10.10	9.81	10.01	8.31	9.87
(5)	7.65	8.14	5.14	5.91	5.01	5.05	4.98	5.01
(6)	16.01	15.13	11.13	12.14	12.15	13.14	11.98	13.06
(7)	6.85	7.97	5.36	5.85	4.98	5.18	5.11	5.08

Source: own calculations.

Table 5.3.5 Simulated Relative Standard Error (Srse in %) of Composite Estimator under L-M and SRSWOR Designs for different sample sizes.

ILRC	For n = 25		For n = 50		For n = 63		For n = 76	
	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR	LM	SRSWOR
(1)	17.65	18.93	13.67	16.48	14.65	15.83	15.01	16.71
(2)	14.98	15.61	11.81	13.48	12.74	13.01	11.82	14.63
(3)	6.08	6.81	4.34	4.78	4.11	4.54	4.08	4.89
(4)	11.98	12.34	9.16	10.15	8.84	9.71	8.01	8.76
(5)	6.34	6.98	4.73	5.01	4.25	4.98	4.13	4.31
(6)	9.24	9.89	7.63	8.13	8.01	7.63	6.79	7.01
(7)	7.11	7.63	5.14	5.44	4.91	5.31	4.16	5.28

Source: own calculations.

ACKNOWEDGEMENT

The authors are grateful to referees for making valuable observations; which helped in improving the paper.

REFERENCE

GHOSH, M. and RAO, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 91, 55—93.

GONZALEZ, N.E. and WAKSBERG, J.(1973). Estimation of the error of synthetic estimates. Paper presented at first meeting of international association of survey statisticians, Vienna, Austria, 18—25.

LAHIRI, D.B. (1951). A method of sample selection providing unbiased ratio estimates. Bull. Int. Stat. Inst. 3, 133—40.

MIDZUNO, H. (1952). On the sampling system with probability proportional to sum of sizes, Ann. Inst. Stat. Math., 3, 99—107.

Rao, J. N. K. (2003).Small area estimation. Wiley Interscience

Sarndal, C. E., Swensson, B. and Wretman, J. (1992). Model assisted survey sampling, Springer - Verlag.

SCHAIBLE, W.L. (1978). Choosing weights for composite estimators for small area statistics. Proceedings of the survey research methods section, Amer. Statist. Assoc., Washington. D.C., 741—746.

SCHAIBLE, W.L., BROCK, D.B., CASADY, R.J. and SCHNACK, G.A. (1977). An empirical comparison of the simple synthetic and composite estimators for small area statistics. Proceedings of Amer. Statist. Assoc., Social Statistics section 1017—1021.

TIKKIWAL, B.D. (1993). Modeling through survey data for small domains. Proceedings of International Scientific Conference on small Area Statistics and Survey Design (An invited paper), held in September 1992 at Warsaw, Poland.

TIKKIWAL, G.C. and GHIYA A. (2000). A generalized class of synthetic estimators with application of crop acreage estimation for small domains. Biom, J. 42, 7, 865—876.

TIKKIWAL, G.C. and GHIYA A. (2004). A generalized class of composite estimators with application to crop acreage estimation for small domains. Statistics in Transitions (6), 5, 697—711.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 125—137

SIMULATION ANALYSIS OF ACCURACY ESTIMATION OF POPULATION MEAN ON THE BASIS OF STRATEGY DEPENDENT ON SAMPLING DESIGN PROPORTIONATE TO THE ORDER STATISTIC OF AN AUXILIARY VARIABLE*

Janusz Wywiał[1]

ABSTRACT

The paper deals with an analysis of the accuracy of the strategies for estimating the mean as well as the total value of a variable under study in a fixed and finite population. Positive valued auxiliary variable is involved. Three strategies called quantile types are compared with a simple sample mean, with an order ratio estimator from the simple sample mean as well as with the order ratio estimator from a sample drawn according to the sampling design proportional to the sample mean of an auxiliary variable. The proposed ratio type strategy is dependent on the sampling design proportional to the value of the order statistic of the auxiliary variable, as proposed by Wywiał (2006). The comparison of the strategies' accuracy has been based on a computer simulation. In the case of a small size population, the mean square errors have been evaluated on the basis of all possible samples which can be selected. In the case of a larger population, the samples have been drawn according to the considered sampling schemes. Finally, appropriate conclusions have been formulated.

Key words: sampling design, order statistic, auxiliary variable, sampling scheme, estimation, strategy, accuracy comparison, relative efficiency

1. Sampling designs and schemes

Let $U=(1,2,...,N)$ be a fixed population of the size N. An observation of a variable under study (an auxiliary variable) attached to the i-th population element will be denoted by y_i $(x_i>0)$, $i=1,...,N$. The sample of size n, drawn without

* The paper was presented at the conference *Survey Sampling in Economic and Social Research*, Katowice, September 11-12, 2006.

[1] University of Economics, Katowice, Bogucicka 3, Poland; E-mail:wywial@ae.katowice.pl

replacement from the population, will be denoted by s. The sampling design is denoted by $P(s)$ and inclusion probabilities of the first and second orders - by π_k, for k=1,....,N and $\pi_{k,t}$ for k≠t, k=1,...N, t=1,...N, respectively. Let $\boldsymbol{S}$ be the sample space of the samples of size n, drawn without replacement. We are going to consider the following sampling designs of simple samples drawn without replacement: $P_0(s) = \binom{N}{n}^{-1}$ for all $s \in \boldsymbol{S}$. The sampling design proportional to the sample mean of the auxiliary variable:

$$P_1(s) = \frac{\bar{x}_s}{\bar{x}} \binom{N}{n}^{-1} \qquad \text{for all } s \in \boldsymbol{S} \tag{1}$$

where $\bar{x}_s = \frac{1}{n}\sum_{i\in s} x_i$, $\bar{x} = \frac{1}{N}\sum_{k\in U} x_k$. In order to find the inclusion probabilities see Brewer and Hanif (1983), Wywiał (1991, 2000, 2003). Lahiri (1951) and Midzuno (1952) proposed the following sampling scheme implementing the $P_1(s)$ sampling design. The first element of the sample is selected with the probability $p(k) = \frac{x_k}{N\bar{x}}$, $k=1,...,N$ and the next $(n-1)$ elements are selected in the same way as the simple sample of size $(n-1)$, drawn without replacement.

Let $X_{(r)}$ be the r-th order statistic from a simple sample drawn without replacement. Let $\alpha \in (0;1)$ and $[n\alpha]$ is the integer part of the value $n\alpha$. The sample quantile of the order α is defined as $Q_{s,\alpha} = X_{(r)}$ where $r = [n\alpha]+1$ and $(r-1)/n \le \alpha <$ r/n. Let $G(i,r)=\{s: X_{(r)}=x_i\}$ be the set of all samples s whose r-th order statistic of the auxiliary variable is equal to x_i where $r \le i \le N-n+r$. Let $U_1=(1,...,i-1)$ be a subpopulation of the population U and let s_1 be the simple sample of size $(r-1)$, drawn without replacement from U_1. Similarly, let $U_2=(i+1,...,N)$ be a subpopulation of the population U and let s_2 be a simple sample of size $(n-r)$, drawn without replacement from U_2. The sample space of the samples of the type s_1 will be denoted by $\mathbf{S}(U_1,s_1)$ and the sample space of the samples of the type s_2 will be denoted by $\mathbf{S}(U_2,s_2)$. Let $s = (s_1 \cup \{i\} \cup s_2)$ be such a sample that the value of the r-th order statistic of an auxiliary variable - observed in the sample - equals x_i. Hence, the sample space of such a sample s is as follows: $\mathbf{S}_{r,i} = \mathbf{S}(U_1,s_1) \times \{i\} \times \mathbf{S}(U_2,s_2)$, where $\times$ is the symbol of the Cartesian product. So, the sample space for all $i=r,...,N-n+r$ is as follows $\mathbf{S} = \bigcup_{i=r}^{N-n+r} \mathbf{S}_{r,i}$.

The probability distribution of the order statistic from the simple sample drawn without replacement (see e.g. Wilks (1962)) is as follows:

$$P_0(X_{(r)} = x_i) = \frac{\binom{i-1}{r-1}\binom{N-i}{n-r}}{\sum_{j=r}^{N-n+r}\binom{j-1}{r-1}\binom{N-j}{n-r}} \quad \text{for r} \le \text{i} \le \text{N-n+r} \qquad (2)$$

Wywiał (2006) proposed the following sampling design proportional to the x_i value of the $X_{(r)}$ statistic.

$$P_2(s \mid r) = \frac{X_{(r)}}{\sum_{j=r}^{N-n+r}\binom{j-1}{r-1}\binom{N-j}{n-r}x_j} \qquad \text{for } s \in \mathbf{S},$$

or

$$P_2(s \mid r) = \frac{x_i}{\sum_{j=r}^{N-n+r}\binom{j-1}{r-1}\binom{N-j}{n-r}x_j} \qquad \text{for } s \in \boldsymbol{S}_{r,i},\ i=r,\dots,N\text{-}n\text{+}r. \qquad (3)$$

Particularly, if $x_i = c>0$ for all $i=1,...,N$, the above sampling design is reduced to the $P_0(s)$ simple sampling design.

The conditional version of the sampling design is as follows.

$$P_2(s \mid x_u \le X_{(r)} \le x_v) = P_2(s \mid r; u, v) = \frac{X_{(r)}}{\sum_{j=u}^{v}\binom{j-1}{r-1}\binom{N-j}{n-r}x_j} \quad \text{for } s \in \boldsymbol{S}_{r/u,v},$$

where $\mathbf{S}_{\mathbf{r}/u,v} = \bigcup_{i=u}^{v}\mathbf{S}_{r,i}$. Equivalently,

$$P_2(s \mid r; u, v) = \frac{x_i}{\sum_{j=u}^{v}\binom{j-1}{r-1}\binom{N-j}{n-r}x_j} \qquad \text{for } s \in \boldsymbol{S}_{r,i},\ i=r,\dots,N\text{-}n\text{+}r. \quad (4)$$

for s∈ G(i,r) and $r \le u \le i \le v \le N\text{-n+r}$.

Let us note that $P_2(s|r) = P_2(s|r;r,N\text{-}n\text{+}r)$. The $P_2(s|r;u,N\text{-}n\text{+}r)$ sampling design will be called the left limited one when $u > r$. If $v < N\text{-}n\text{+}r$, $P_2(s|r;r,v)$ will be called the right limited one. Finally, $P_2(s|r;u,v)$ will be named limited when both $u > r$ and $v < N\text{-}n\text{+}r$.

Let us define such a function $\delta(t)$ that if $t<0$, $\delta(t)=0$ else $\delta(t)=1$. Moreover, let

$$z_r(u,v) = \sum_{j=u}^{v} x_j \binom{j-1}{r-1}\binom{N-j}{n-r}$$

The inclusion probabilities of the first order are as follows:

$$\pi_k = \frac{\delta(u-k)\delta(r-1)\delta(v-1)\delta(u-1)}{z_r(u,v)}\sum_{i=u}^{v}\binom{i-2}{r-2}\binom{N-i}{n-r}x_i +$$

$$+\frac{\delta(k-u+1)\delta(v-k+1)}{z_r(u,v)}\left(\delta(n-r)\delta(k-u)\delta(k-1)\sum_{i=u}^{k-1}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1}x_i + \binom{k-1}{r-1}\binom{N-k}{n-r}x_k +\right.$$

$$\left.+\delta(r-1)\delta(v-k)\sum_{i=k+1}^{v}\binom{i-2}{r-2}\binom{N-i}{n-r}x_i\right) + \frac{\delta(k-v)\delta(n-r)\delta(N-v)}{z_r(u,v)}\sum_{i=u}^{v}\binom{i-1}{r-1}\binom{N-i-1}{n-r-1}x_i \quad (5)$$

for k=1,...,N. The probabilities of the second order are derived by Wywiał (2006).

The sampling scheme implementing the *P₂(s|r;u,v)* conditional sampling design is as follows. Firstly, population elements are ordered according to the increasing values of the auxiliary variable. Next, the *i*-th element of the population is drawn with this probability:

$$p_2(i \mid r;u,v) = \frac{\binom{i-1}{r-1}\binom{N-i}{n-r}x_i}{\sum_{j=u}^{v}\binom{j-1}{r-1}\binom{N-i}{n-r}x_j}, \quad \text{i=u,...,v} \qquad (6)$$

Next the simple sample s_1 of size *(r-1)* is drawn from the subpopulation U_1 and the simple sample s_2 of size *(n-r)* from the subpopulation U_2. Let us note that the expression (6) shows us the truncated distribution of the order statistic $X_{(r)}$ from the sample drawn according to the sampling design $P_2(s|r;u,v)$. Hence, $p_2(i \mid r;u,v) = P_2(X_{(r)} = x_i \mid x_u \le X_{(r)} \le x_v)$.

Particularly, when u=v,

$$P_2(s \mid r;u,u) = \frac{P_2(s \mid r)}{P_0(s \mid X_{(r)} = x_u)} = \frac{1}{\binom{u-1}{r-1}\binom{N-u}{n-r}}$$

In this case p₂(u|r;u,u)=1. This is a case of the well known stratified sampling design.

2. Estimators and strategies

The well known Horvitz-Thompson (1952) estimator is as follows.

$$t_{HTS} = \frac{1}{N}\sum_{k=1}^{N}\frac{a_k y_k}{\pi_k} \qquad (7)$$

where $a_k = 1$, if the *k*-th population element was drawn to a sample. When $a_k = 0$, the *k*-th element was not drawn to the sample. It is well known that the strategy

$(t_{HTS}, P(s))$ is unbiased for the population mean when all inclusion probabilities are positive. Moreover, the strategy $(t_{HTS}, P_0(s)) = (\bar{y}_S, P_0(s))$ is called a simple sample mean. In the next paragraph, we are going to consider the strategy $(t_{HTS}, P_2(s \mid r; u, v))$.

The ordinary ratio estimator is as follows:

$$t_{RS} = \bar{y}_S \frac{\bar{x}}{\bar{x}_S} \tag{8}$$

The strategy $(t_{RS}, P_1(s))$ is unbiased for the population mean. As it is well known, the strategy $(t_{RS}, P_0(s))$ is almost unbiased for population mean when the sample size is sufficiently large.

On the basis of the expression (2) we evaluate the following expected value of the truncated distribution of the *r*-th order statistic of the auxiliary variable.

$$E_0(X_{(r)} \mid u, v) = \sum_{i=u}^{v} x_i P_0(X_{(r)} = x_i \mid u, v) \tag{9}$$

where

$$P_0(X_{(r)} = x_i \mid u, v) = \frac{P_0(X_{(r)} = x_i)}{P_0(x_u \leq X_{(r)} \leq x_v)} \tag{10}$$

Particularly:

$$E(X_{(r)} \mid r, N-n+r) = E(X_{(r)}) = \binom{N}{n}^{-1} \sum_{i=r}^{N-n+r} x_i \binom{i-1}{r-1}\binom{N-i}{n-r}.$$

Wywiał (2006) proposed the following estimator.

$$t_{qS} = \bar{y}_S \frac{E_0(X_{(r)})}{X_{(r)}} \tag{11}$$

This estimator can be named as the order-statistic-ratio estimator or the quantile-ratio estimator. The $(t_{qS}, P_2(s \mid r))$ strategy is unbiased for the population mean. The strategy $(t_{qS}, P_2(s \mid r, u, v))$ can be biased if $r < u < v < N - n + r$.

The next estimator is as follows.

$$t_{qS,1} = \bar{y}_{HT,S} \frac{E_0(X_{(r)})}{X_{(r)}} \tag{12}$$

The parameters of the $(t_{qS,1}, P_2(s \mid r, u, v))$ strategy will be considered in the next part of this paper.

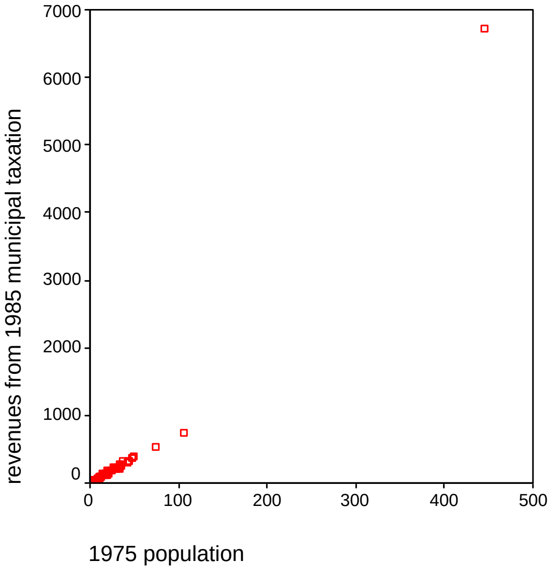
The $\left(t_{RS};P_0(s)\right)$ and $\left(t_{RS};P_1(s)\right)$ strategies will be named moment dependent strategies. The quantile dependent strategies are: $\left(t_{HTS};P_2(s)\right)$, $\left(t_{qS};P_2(s)\right)$ and $\left(t_{qS,1};P_2(s)\right)$.

3. Accuracy comparison of estimation strategies

3.1.The case when there is a single outlier

The population in the demonstration consists of the municipalities in one region in Sweden. The auxiliary variable x is: *1975 municipal population (in thousands)* and the variable under study y is: *1985 municipal taxation revenues (in millions of kronor)*. Their observations have been published by Särndal, Swenson and Wretman (1992). The size of this population is 56 municipalities. There is one outlier observation of the variables, see Figure 1.

Figure 1. Scatterplot with the outlier.



The accuracy of the *(t_S,P(s))* estimation strategy was measured by means of the relative efficiency - *deff*:

$$deff(t_S, P(s)) = \frac{MSE(t_S, P(s))}{D^2(\bar{y}_S, P_0(s))} 100\%$$

Let deff1=deff$\left(t_{RS};P_0(s)\right)$, deff2=deff$\left(t_{RS};P_1(s)\right)$, deff3=deff$\left(t_{qS};P_2(s\mid r;u,v)\right)$, deff4=deff$\left(t_{HTS};P_2(s\mid r;u,v)\right)$, deff5=deff5$\left(t_{qS,1};P_2(s\mid r;u,v)\right)$.

In tables 1—2 and figure 2 accuracy coefficients of the above strategies are evaluated on the basis of all possible samples which can be drawn without replacement from the population.

Table 1 demonstrates the accuracy of the three strategies for estimation of mean values on the basis of the samples of size 3, drawn without replacement. Two cases were considered. The first of them concerns already introduced population of size 56. The outlier observation of the variables *x* and *y* is attached to one element of this population. The next population is obtained after removing this element with the outlier observation.

Table 1. The strategies dependent on order statistics.

deff	*Data with autlier*			*Data without autlier*		
	r=3	*r=2*	*r=1*	*r=3*	*r=2*	*r=1*
Deff3	*1.11*	*87.85*	*103.20*	*21.12*	*61.86*	*99.14*
Deff4	*5.5*	*56.68*	*78.18*	*25.80*	*41.24*	*70.61*
Deff5	*10.14*	*33.49*	*65.94*	*55.89*	*49.71*	*73.16*

Let n=3. In the case of existing of the outlier *deff1*=1.99, *deff2*=3.52 and $D^2(\bar{y}_S,P_0(s))=248725$. When the outlier is omitted *deff1*=1.79, *deff2*=1.49 and $D^2(\bar{y}_S,P_0(s))=6074$.

The analysis of table 1 leads to the conclusions that the quantile type estimators are the most efficient when the rank of the order statistic is equal to the sample size *n=3*. The $\left(t_{qS};P_2(s\mid r)\right)$ strategy is the most accurate among the considered ones in the case of existing the single outlier. Moreover, let us note that the strategy $\left(t_{qS,1};P_2(s\mid r)\right)$ is less efficient among the considered ones for *r=3*.

3.2.Left limited sampling design

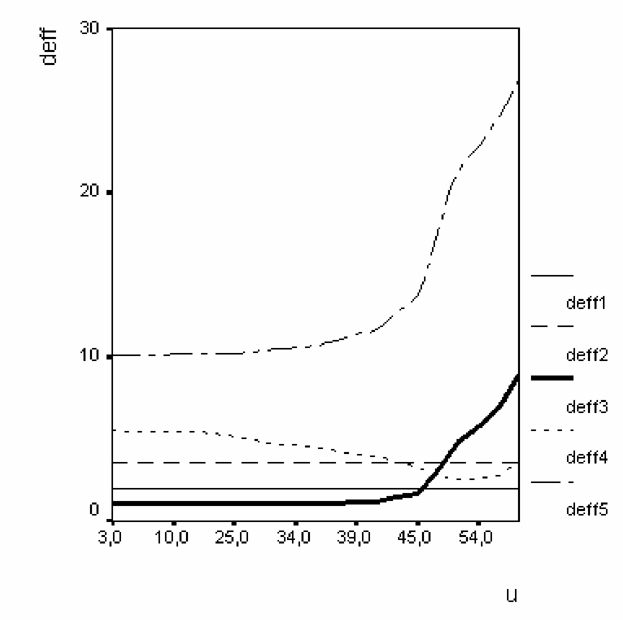
The above described data with the outlier are considered here, too. The left limited sampling design will be taken into account. . Let the parameter *u* identify the left truncation point x_u of the distribution of the order statistic given by the expression (10). Table 2 represents the evaluated *deff* coefficients of the strategies for n=r=3.

Table 2 as well as figure 2 lead to conclusion that the $\left(t_{qS};P_2(s\mid r;u,N-n+r)\right)$ strategy is the best among the ones considered for *u≤40*. The $\left(t_{qS,1};P_2(s\mid r;u,N-n+r)\right)$ strategy is the worst for all the values *u∈[5;56]*. The cases when *r<3* were not considered because we can expect that only the strategies based on moments will be the most accurate.

Table 2. Efficiency of the strategies for *n=r=3*.

u	*deff3*	*deff4*	*deff5*
5	*1,11*	*5,49*	*10,1*
10	*1,11*	*5,48*	*10,2*
20	*1,08*	*5,35*	*10,2*
30	*1,06*	*4,94*	*10,3*
40	*1,22*	*3,95*	*11,7*
45	*1,70*	*3,29*	*13,7*
55	*6,96*	*2,79*	*24,6*
56	*8,72*	*3,60*	*26,8*

Figure 2. Efficiency for *n=r=3*. The case with outlier.



3.3. Accuracy of the strategies dependent on degree of the order statistic

Now, observations of the *(x,y)* variables in the population of the *284* municipalities in Sweden are taken into account (Särndal, Swenson and Wretman (1992)). In this population three outlier observations exist. So, the first population (with outliers, see figure 3) is of size *284* and the second one (without outliers) is of size *281*. In the case of the first population *deff1=4,89*, $D^2(\bar{y}_S; P_0(s))$*=6536,014* and in the case of the second one *deff1=11,01*, $D^2(\bar{y}_S; P_0(s))$*=117282,4*. The analysis of estimation accuracy was based on a computer simulation. The samples of size 6 were drawn 5,000 times from each population and for *r=1,...,6*. The sampling scheme defined by the expression (6) was used in order to select samples. Let expression (5) evaluate the value of the inclusion probabilities which were used to determine the value of the Horvitz-Thompson estimator.

Figure 3. Scatterplot of the *x* and *y* variables; the population with three outliers.

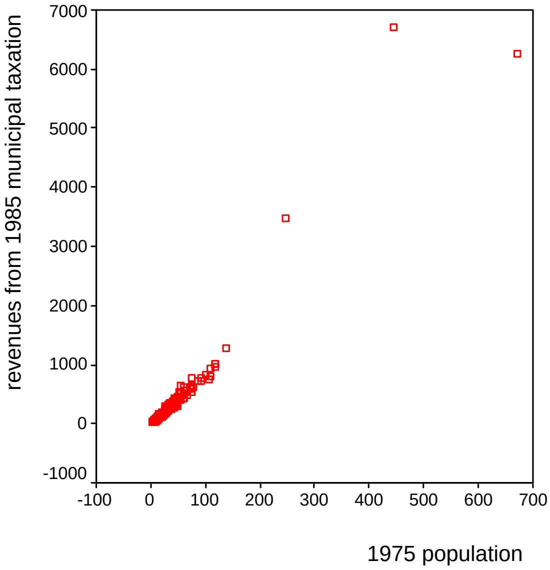


Figure 4. The *deff* dependent on order stat. for *n=6* in the case of the pop. without outliers

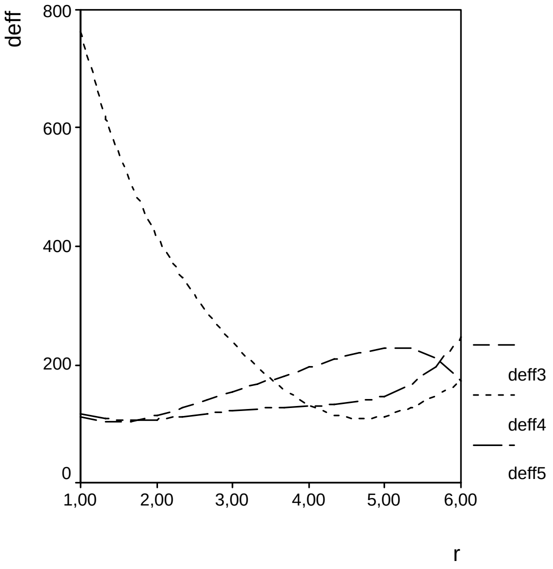
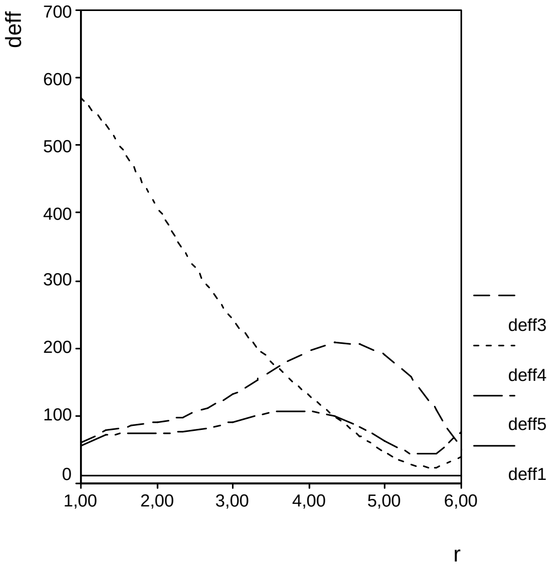


Figure 5. The *deff* dependent on order stat. for *n=6* in the case of the pop. with outliers.



On the basis of the contents of table 3 and figures 4 and 5, we infer that the quantile type strategies are less accurate than those based on moments. Only in the case when r=n=6 and the outlier exists the quantile type strategies can be more accurate than the simple sample mean.

Table 3. The *deff* of the strategies dependent on order statistics for *n=6*.

r	*Data with outliers*			*Data without outliers*		
	deff3	*deff4*	*deff5*	*deff3*	*deff4*	*Deff5*
1	*59,69*	*571,22*	*55,18*	*112,54*	*762,93*	*117,06*
2	*90,19*	*409,46*	*74,19*	*113,54*	*417,26*	*107,25*
3	*131,12*	*241,59*	*91,33*	*153,60*	*237,38*	*121,56*
4	*196,69*	*129,28*	*107,43*	*194,84*	*131,20*	*129,48*
5	*190,42*	*45,63*	*62,41*	*227,67*	*111,99*	*146,89*
6	*49,75*	*40,52*	*76,28*	*172,95*	*169,87*	*245,80*

3.4. Dependence of relative efficiency on sample size

The observations of the *(x,y)* variables in the population of the *284* municipalities (with outliers) in Sweden - analysed in the previous paragraph - will be considered here, too.

Figure 6. The *deff* for several sample sizes equal to the *r* degree of the order statistic in the case of a population with outliers.

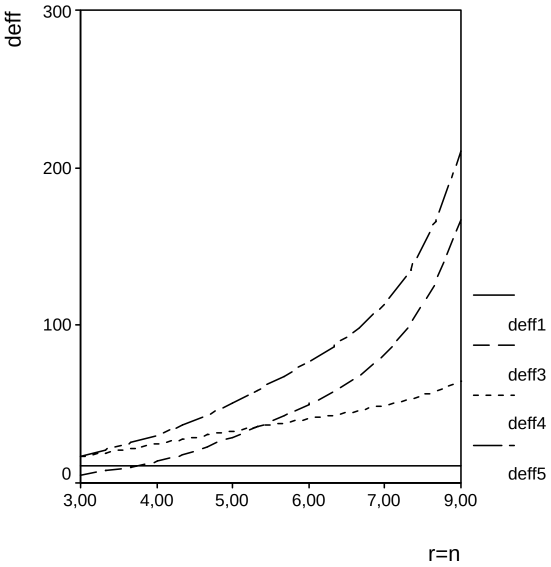


Table 4 and figure 6 let us infer that the quantile dependent strategies are more accurate than simple sample mean only for small sample sizes. Among those strategies, only the $\left(t_{qS};P_2(s\mid r)\right)$ is more accurate than the order ratio estimator from the simple sample - only when *n=r=3*.

Table 4. The *deff* for several sample sizes n= *r* in the case of the population with outliers.

r=n	*deff3*	*deff4*	*deff5*
3,00	*4,79*	*16,72*	*17,03*
6,00	*49,95*	*40,66*	*76,64*
7,00	*81,41*	*49,18*	*113,97*
12,00	*369,79*	*89,71*	*431,49*
15,00	*664,86*	*114,75*	*745,19*
20,00	*1410,0*	*157,06*	*1524,4*
30,00	*3893,5*	*243,45*	*4078,7*

4 Conclusions

The results of the comparison of the introduced quantile type strategies with the well known moment-based ratio strategies are equivocal. Generally, the quantile type strategies can be more precise than the simple sample means when the degree of the order statistic is large and the sample size is small. Among the quantile type strategies, the $\left(t_{qS};P_2(s\mid r;u,v)\right)$ strategy can be preferred especially in the case when outliers (too large values) of a variable under study and an auxiliary variable exist. In this case, their precision can be even better than that of the moment-based strategies but only for a small size of the sample and a large degree r of the order statistic. Hence, it can be eventually useful as sampling design on the second stage of two-stage sampling design.

The received results are valid of course when we estimate the total value of a variable under study. In this case estimators are obtained through multiplying the considered ones by the size of a population.

Acknowledgement

The research was supported by the grant number 1 H02B 018 27 from the Ministry of Science and Higher Education.

REFERENCES

BREWER K. R.W., HANIF M. (1983). *Sampling with unequal probabilities*. Springer Verlag, New York-Heidelberg-Berlin 1983.

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

LAHIRI G. W. (1951): A method for sample selection providing unbiased ratio estimator. *Bulletin of the International Statistical Institute*, vol. 33, s. 133—140.

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

SÄRNDAL C. E., B. SWENSSON, J. WRETMAN: (1992): *Model Assisted Survey Sampling.* Springer Verlag, New York-Berlin-Heidelberg- London-Paris-Tokyo-Hong Kong- Barcelona-Budapest.

WILKS S. S. (1962). Mathematical Statistics. John Wiley and Sons, Inc. New York, London.

WYWIAŁ J. (2006). Sampling design proportional to order statistic of auxiliary variable. *Statistical Papers,* http://www.springerlink.com/content/1613-9798/?sortorder=asc&Content+Status=Accepted.

WYWIAŁ J. (2003). *Some Contributions to Multivariate Methods in Survey Sampling*. Katowice University of Economics, Katowice.

WYWIAŁ J. (2000). On precision of Horvitz-Thompson strategies. *Statistics in Transition* vol. 4, nr 5, pp. 779—798.

WYWIAŁ J. (1991). On sampling design proportional to mean value of an auxiliary variable (in Polish). *Wiadomości Statystyczne,* nr 6, 1991, pp. 21—23.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 139—153

EXTREME VALUE TREATMENT FOR SAMPLES FROM SKEW INCOME DISTRIBUTIONS

Marek Balog[1] and Daniel Thorburn[2]

ABSTRACT

Income distributions, as well as many other economic variables, are often quite skew with a few very large values. When a sample is taken one may get too many or too few extreme observations. Estimates of totals and means, as well as inequality measures like the Gini Coefficient, can be strongly influenced when ordinary design-based methods are used, if the outliers are weighted according to their inclusion probabilities. In this paper we use a model for the tail of the distribution. The parameters of the tail distribution are estimated and the estimated tail distribution is used for estimation of means and inequality measures. In this way the possibility of extreme values are taken into account even when there is no extreme value in the sample and the impact of the extreme values are decreased if there are too many of them in the tail. The methods are applied to the full income distribution in Sweden. It is shown that this method gives better results than both classical design-based and classical outlier methods.

Key words: Outliers, Gini coefficient, Income share ratio, OBRE algorithm, Parametric tail, Pareto distribution, Survey sampling, Welfare measures

1. Introduction

There is a need for finding good methods to estimate welfare indicators from surveys, which can serve as reliable methods for comparisons between individual groups or countries. A big problem is that the distributions are skew and that one may get too many or none of the really extreme values in the sample. One should thus want a method that, loosely speaking, imputes a few extreme observations when there are too few in the sample and remove some when there are too many. The problem is that the true number of extremes is unknown. We suggest that one

[1] Dep. of Statistics, University of Economics, Dolnozemská cesta 1, Bratislava, Slovakia, balogmarek@yahoo.com

[2] Dep. of Statistics, Stockholm University, S-106 91 Stockholm, Sweden, Daniel.Thorburn@stat.su.se

should model the tail of the distribution, using part of the upper tail in the sample. Using that model one may estimate what the impact should be of the most extreme values.

P. Van Kerm (2006) says: *“Key social indicators are poverty and inequality indicators which inform about the dispersion of citizen’s incomes and equally show how much are people lagging behind societal standards”*. Necessarily if we want to make sure that estimation of welfare indices will be as accurate as possible we must focus attention on methods for the outliers treatment in our data on income distribution.

We will study some methods to estimate some common indicators from samples of the Swedish populations. We have access to a register of the disposable incomes in Sweden during 2004, based on the tax authorities` figures. Thus we can compare our estimates with the true figures based on the full data. Lee (1995) and Chambers and Kokic (1986) give good overviews of outlier handling in economic and business surveys.

2. The problem of outliers in surveys with skew data

One may ask why one should take the outlier problem into consideration at all when doing sample surveys without measurement errors. The outliers are true values of a real object in the frame. In ordinary robust statistics there is a risk that the data is contaminated and methods are developed so that erroneous observations do not have undue influence. In our case one might argue that one should not do any omissions or corrections of the original data if one wants to make estimates based on true data. But since the weights may be quite large, the estimates may change a lot if one of the largest units is included in the data or not (The weights are usually define by $w_i=1/\pi_\iota$, where π_ι is the inclusion probability). Is it acceptable that a few randomly selected data points have a large leverage on the estimation of national indicators? So, how much do extreme data affect conclusions that are made from estimated welfare measures? We suggest that one should use ordinary design based estimation technique for the main bulk of the data, but for the extreme values one should use a model.

Statistics Sweden has kindly provided us with true data on the incomes of all Swedish inhabitants from the Swedish tax authorities for the fiscal year 2004. We will study the effect of different estimation methods on some well-known income inequality measures. We will see how much the different adjustments affect them, both in the bias, standard error and mean square error. Uncertainty and unreliability due to sampling variation should be kept to a minimum. If extreme values are present in the income distribution either by mistake or genuine randomness the income data are still suspect and it is required to control the impact of outliers on the estimates.

3. Methods of outlier treatment

There are several strategies for dealing with outliers in measured data sets. We will discuss a few.

a) Classical adjustments:

- **Trimming -** trimming a fraction of the data – it means, discarding the top and/or bottom *p* percent of the data. The basic idea behind trimming is that outliers are so dubious that they in fact do not give any information.
- **Winsorising -** is close to trimming, with difference that the extreme data are not removed from the datasets but are replaced by the value of the trimming thresholds. In other words it is a transformation of the outliers that replaces all the top and/or bottom *p* percent of the data by the corresponding percentiles of the data; for example a 90 % winsorisation of the upper part replaces the 10% highest values by the 90th percentile. The basic idea behind winsorising is that the outliers represent true objects but their magnitude is probably exaggerated. That is it would be sensible to replace them by a not quite so extreme value.

Variants of classical adjustments:

In sampling there are two variants of both trimming and winsorising. In one version of trimming the extreme object is left out completely. In another version the object is kept but its weight is set to 1 (instead of$1/\pi_i$), i. e. the object is assumed to be true, but so rare that it should not affect the estimates of the remaining ones (Dalén, 1987). In the corresponding version of winsorising the object is replaced by two values: once with the observed value and sampling weight 1 and once with the specified percentile and weight ($1/\pi_i$)-1.

b) Parametric approaches:

Drawing from parametric tails — this approach is based on modelling parametrically or semi-parametrically the income distribution and estimating parameters with estimators that are robust to data contamination. Thus social indicators can be estimated indirectly from the estimated distribution. Since most welfare indicators are sensitive to the outliers it may be reasonable to use robust parametric models for data adjustments. The concept is based on imputation of extreme income values by replacing the actual observed values occurring in distribution tail by random draws from the robustly estimated parametric distribution. Certain uncertainty is introduced by the random draws as for simulation. It can be easily avoided by taking set of income data replication and estimating welfare indicators for each replication and individual indicator result would be an average over the replications of each indicator.

In the literature several skew distributions have been proposed for income data. (Chambers & Kokic, 1986), mainly Pareto, lognormal and Gamma but also

Weibull and other extreme values distributions have been suggested. Thorburn (1991) and Karlberg (1999, 2000) used the lognormal. In this paper we will follow van Kerm (2006) and model only the tail. The Pareto distribution is a good choice in this case since it is straightforward to estimate the parameters using and modelling only the tail.

The idea of our parametric tail approach (cf. van Kerm 2006) is to assume that the distribution can be described by a model but only for observations above a certain limit. Those below that limit are estimated non-parametrically by their mean. Suppose that we have a sample of n, say 10 000, observations. First divide this sample into two parts S_- and S_+, where the latter contains only the k, say 100, largest observations. The latter are modelled to come from the tail of a distribution F_θ with parameter θ, which is estimated from the data. Van Kerm (2006) suggested drawing several new sets, each with k new tail observations from $F_{\theta*}$ to replace the observed S_+. In this way an extreme outlier in the sample will not necessarily occur in the sample next time. On the other hand if there were no really extreme in the sample there may pop up one in the resampled new set. Van Kerm hopes that the proportion of extremes will be more correct in the combined sample compared to the original sample. He used eight resamples. We will use ten. To make the estimate even better we will in the simulation part use a numerical resampling method that completely eliminates the random error due to resampling.

We choose the Pareto distribution since it is often used for income distributions and it is easy to estimate the parameters given only a truncated sample (in our case S_+). Its cumulative distribution function of the tail is given by

$$F(x;\theta;x_m) = 1 - \left(\frac{x_m}{x}\right)^{\theta}$$

for $x \geq x_m$, where x_m is the value of the cut-off point in the tail. In our case it is the value of the 100th or 250th highest disposable income. The parameter θ is a positive parameter of the shape. The smaller it is the skewer the distribution. The expected value of an observation in the tail (above x_m) is $x_m*\theta/(\theta-1)$ and its variance is $x_m*\theta/((\theta-1)(\theta-2)^2)$. When $\theta < 2$ the distribution has no variance and if $\theta < 1$ it has no expected value either.

Resampling once, eight or ten times introduce an error. This should be unnecessary since it is possible to compute the mean and variances exactly given the θ-values. We did so numerically by selecting 100 equidistant quantiles from the Pareto distribution for the 100 top incomes. (The quantiles, which correspond to the proportions 0.005, 0.015, 0.025, …, 0.995). Thus we avoided introducing extra variance by drawing random samples. The numerical integration was done since some indicators are non-linear functions.

We have estimated parameter θ of the Pareto tail distribution by using:

- **Maximum likelihood estimation** — where the maximum likelihood estimator for θ is

$$\widehat{\theta} = \frac{n}{\sum_i \left(\ln x_i - \ln \widehat{x}_m \right)}.$$

- **Optimal B-robust estimator (OBRE):** the MLE is not a robust method, which means that the parameters estimates may depends strongly on the occurrence of an extreme value in the sample. OBRE is a robust method defined in Hampel *et* al. (1986) and belongs to the class of M-estimators (Huber 1972, 1981). The algorithm is described in Victoria-Feser & Rochetti (1994) and used by van Kerm (2006).

In the OBRE method one must choose a parameter c that can be interpreted as a regulator between robustness and efficiency. The sensitivity of the OBRE to contamination depends on the choice of the bound c. The lower the bound c, the less sensitive the OBRE is to contamination. However, lowering the bound c leads to an efficiency loss at the model.

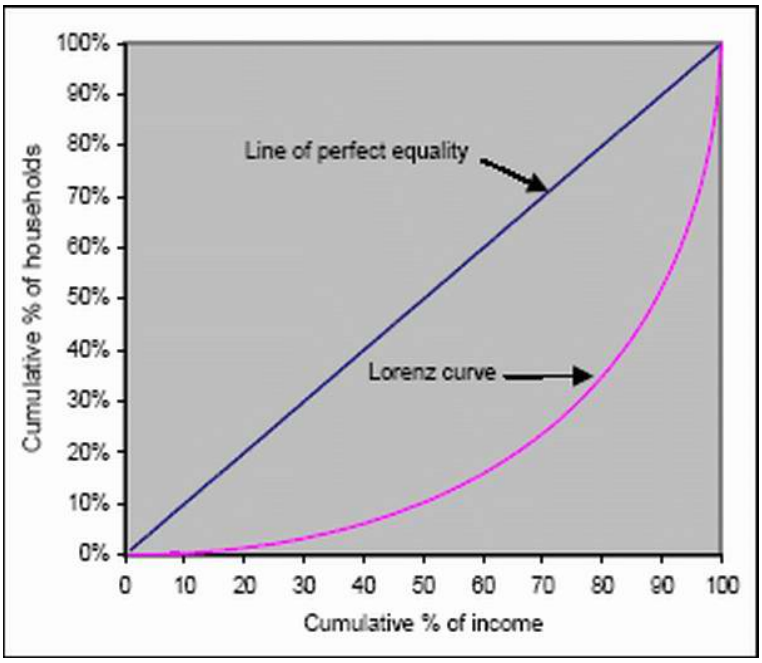
4. Social indicators

We will study several indicators one central tendency indicator and three inequality indicators:

Mean disposable income It is expected that the mean will be affected by the adjustments. Furthermore, it can be used to determine poverty line and therefore its sensitivity gives us indication about the susceptibility of the determination of the poverty line. The median is often used but uninteresting for our purpose.

Income share ratios (S80/S20): the ratio of total income received by the 20% of the population with the highest income to that received by the 20% of the population with the lowest income.

The Gini coefficient: is a measure of income inequality. 0 corresponds to perfect income equality (i.e. everyone has the same income) and 1 corresponds to extreme income inequality (i.e. one person has all the income, while everyone else has zero income). The Gini index is the Gini coefficient expressed in percent (i. e. multiplied by 100). It corresponds to the area between the Lorenz curve of the distribution and the uniform (perfect) distribution line divided by total area of the triangle. (The Gini coefficient is usually defined only for positive distributions. It can be defined for populations with negative values. In that case it may become larger than 1).



The variance coefficient. The standard deviation is the root mean square deviation of the values from their arithmetic mean. It is the most common measure of statistical dispersion. To make it dimensionless we have used the variation coefficient instead.

5. Description of the data

Data of relevance for this study are from an administrative register namely the income register at individual level, from the income year 2004 in Sweden. It is based on the self-assessments to the tax authorities (Jansson, 1999). The income register is the frame for the survey on the household finances that Statistic Sweden performs yearly. It consists of income information for all 9 011 000 individuals occurring in Swedish tax records. We will use the variable “disposable incomes” from the register. It is a very skew variable since it includes also interests and capital gains from dividends and selling of stocks and other physical capital like real estates. On the other hand taxes, social security contributions, paid interests and capital losses are withdrawn. Since our work deals with the problems at the upper tail we have excluded all persons younger then 18 years due to their insignificant influence on the overall incomes and all persons with negative incomes as well. Negative disposable incomes can occur when people suffer from large capital losses or have large debts with heavy interests. Most people under 18 have a disposable income of 0 in the registers. The exclusion of the lower tail was necessary for computing some of the inequality indicators. After this 6 973 854 observations remain in our corrected population.

When one talks of inequality indicators one may want to use different figures. For some purposes capital gains and losses should be periodized so that changes between adjacent years do not fluctuate too wildly due to the development of the economic cycles. For other purposes the data in the register should be considered as having measurement errors due to for instance the tax authorities' methods to determine the taxable income or people not reporting black incomes. We will, however, assume that we are really interested in the register figures and we further will assume that these are not subject to any measurement errors. In some cases it may be a good way to estimate the proportion of large outliers from previous years (cf. Thorburn 1991) but this is not a solution in our situation since the proportion of outliers vary much over the economic cycle. Years with a rapid increase in the prices of stock and real estates will have more extreme outliers.

The true values of our measures on the income distribution are given in Table 1. The SAS software has been used for all computations and estimations here and subsequently.

Table 1. True values on the distribution of disposable income in the Swedish population 2004. All persons who are 18 years or more and who have positive incomes

Mean	Std Dev	Variation Coeff	Gini coeff	S80/S20
165131	415182	2,514	0,301	4,994

6. Application to real data

a) An illustrative example

We first describe the methods using only one simple random sample of 10 000 observations. Subsequently we have estimated the five measures of interests in this sample *(for results see Table 2, row Sample values)*. The trimmed and winsorised estimates are also given in Table 2.

Table 2. Numerical comparison of trimmed and winsorised estimates.

		Mean	Std Dev	Var. Coeff	Gini coeff	S80/S20
True population values		**165131**	**415182**	**2,514**	**0,3010**	**4,994**
Sample values		**167906**	**251632**	**1,499**	**0,3105**	**5,174**
Trimming	**k=50**	159915	88961	0.556	0.279	4.552
	k=10	162828	101040	0.621	0.290	4.764
	k=5	163582	104362	0.638	0.293	4.831
Winsorising	**k=50**	163067	99255	0.609	0.290	4.792
	k=10	163853	106059	0.647	0.293	4.855
	k=5	164760	119214	0.724	0.297	4.927

In Table 3, you can find the comparison of the OBRE and the MLE method. The parameter θ is estimated by 2.61 by the OBRE-estimator and by 2.49 using the ML-method. After finding the parameter θ, we have drawn 10 new resamples of 250 random observations from Pareto distribution with parameter θ=2.61 and 250 with θ=2.49 above the cut-off-level. For each total sample all five indicators were estimated (from the 9750 smallest and the new resample)

The Pareto distribution is not yet common in many programming languages. But one can easily generate a random sample by using inverse transform sampling. Given a random variate U drawn from the uniform distribution on the unit interval (0;1), the variate $T = \frac{x_m}{U^{\frac{1}{\theta}}}$ is Pareto-distributed.

In Table 3 we first give the results for two of the resamples of size 250, then the means of the indicators from the ten resamples, then the mean of the original sample and the ten resamples. The last line for each method contain the estimates when we only estimate the θ-values from the 100 highest values and replace them by 100 values uniformly spread over the whole distribution.

Table 3. Results of the OBRE and the MLE for the Pareto distribution model.

		Mean	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
True population Values		**165131**	**415182**	**2,514**	**0,3010**	**4,994**
Sample values		**167906**	**251632**	**1,499**	**0,3105**	**5,174**
K=250 **θ=2,61**	**OBRE c=2**	**Mean**	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
	First resample	164754	116689	0,708	0,297	4,92
	Second resample	166442	208537	1,253	0,304	5,06
	Mean of 10 res.	164571	122831	0,746	0,297	4,91
	Mean res + obs	164874	139546	0,846	0,298	4,93
k = 100 **θ = 2,10**	***Expected value***	165691	138090	0,833	0,301	5,00
K=250 **θ=2,49**	**MLE**	**Mean**	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
	First resample	165208	121661	0,736	0,299	4,96
	Second resample	167234	240058	1,435	0,308	5,12
	Mean of 10 res	165025	130070	0,787	0,298	4,94
	Mean res + obs	165287	145386	0,880	0,300	4,96
k = 100 **θ = 2,17**	**Expected value**	165451	133485	0,807	0,300	4,98

This is only one sample and one should not draw any far-reaching conclusions from it. But one may note some features. The winsorised and trimmed estimates gave underestimates compared to the parametric method. It is probably not enough to base the estimates on only a few resamples, since they vary considerably. The θ–estimates are smaller when only the 100 highest values are used. This was a general tendency also from other computer runs. We decided to use k = 100 in the simulations, but now having seen the results of the simulation, we believe that it might have been better with even smaller k-values. For example to estimate θ from the 25 or 50 highest values but only replace the 10 highest values. One may also see that the difference between the robust OBRE-method and the ML-estimate is not very large. This is probably due to the fact that even the ML-estimate is fairly robust for the Pareto distribution. Finally one may note that it may be a good idea to keep the observed samples and include them in the estimate with a smaller weight.

b) Simulations

The question of the replaced proportion remains. Van Kerm (2006) used a complicated formula, which always ends up with replacing between 2 and 3% of the sample using a Pareto tail. We made some preliminary runs replacing the 3, 2.5, 1.5, 1 and 0.5 % largest incomes. Those preliminary runs indicated that the best results were obtained when the 100 top values were replaced. We used this value in our small simulation. Of course, different Pareto tail sizes are appropriate in different populations.

To see how the methods for dealing with extreme values perform we drew 50 independent simple random samples from our population. The 100 top incomes were replaced with the observations from the Pareto distribution by using the MLE and the OBRE θ estimates and the ten top incomes were trimmed or winsorised. We have estimated the mean, standard deviation, variance coefficient, Gini coefficient and income share ratio in each of the 50 samples for each of the five methods raw data, trimmed, winsorised, Pareto-OBRE and Pareto-MLE. When we looked at the 50 samples, it was obvious that even when really bad samples were drawn, both Pareto tail methods yielded convincing results. The estimates were similar to the population figures in almost all samples. Summary statistics of all 50 random samples are reported in Table 5 for Pareto tail (OBRE and MLE) estimates and trimmed or winsorised estimates.

Table 5. Mean, standard deviation and root mean square error averaged over 50 random samples.

True values	**Mean**	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
	165131	415182	2,514	0,301	4,99
SRS					
	Mean	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
Mean	165215	212936	1,245	0,301	5,00
Variance	18739939	120064321977	3,289	0,0003	0,136
Std Dev	4328	346502	1,814	0,0167	0,369
Bias	-84	202246	1,270	0,0002	-0,0052
MSE	18747009	160967952075	4,901	0,0003	0,137
RMSE	4329	401208	2,214	0,0167	0,369
OBRE					
	Mean	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
Mean	163711	115780	0,707	0,295	4,88
Variance	1422613	71207311	0,0024	0,0000	0,011
Std Dev	1192	8438	0,049	0,0040	0,103
Bias	1419	299402	1,807	0,0062	0,116
MSE	3437644	89713228749	3,268	0,0001	0,024
RMSE	1854	299521	1,808	0,0074	0,155
MLE					
	Mean	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
Mean	163828	118322	0,722	0,295	4,89
Variance	1747701	190011359	0,0064	0,0000	0,014
Std Dev	1322	13784	0,080	0,0049	0,117
Bias	1302	296860	1,792	0,0057	0,107
MSE	3445222	88316308769	3,219	0,0001	0,025
RMSE	1856	297180	1,794	0,0076	0,158
Trimming, k=10					
	Mean	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
Mean	161685	108327	0,669	0,286	4,71
Variance	1063566	85427178	0,0031	0,00002	0,011
Std Dev	1031	9242	0,056	0,0040	0,103
Bias	3446	306855	1,844	0,0145	0,284
MSE	12940441	94245839910	3,403	0,0002	0,091
RMSE	3597	306994	1,845	0,0151	0,302
Winsorising, k=10					
	Mean	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
Mean	162718	114490	0,704	0,292	4,83
Variance	1238886	259664686	0,0096	0,00002	0,014
Std Dev	1113	16114	0,098	0,0049	0,119
Bias	2413	300692	1,810	0,0092	0,169
MSE	7062597	90675747337	3,287	0,0001	0,043
RMSE	2657	301124	1,813	0,0104	0,206

From Table 5 it is seen that both the Pareto-estimators (OBRE and MLE) have about the same bias and mean square error. They are much less biased than the trimmed and winsorised estimators but have a much higher variance. Their mean square error is also smaller. It seems that no method works well for the estimation of the variation coefficient. It might be better to use a more skew distribution, i.e. the lognormal, or replace a smaller proportion. But on the other hand the estimator would become unstable. A good method may be to estimate θ from the 50 top values and then replace only the ten highest values with a Pareto tail with that θ value and the left hand limit, x_m equal to the 11th highest value.

c) The sensitivity to changes in one large observation

Previous sections confirm that the estimates based on the Pareto tail distribution are robust. The OBRE estimates do not change at all when a small fraction of the largest values are changed. Even when the 20 top values were multiplied by 10 the estimates did not change. In this respect the estimates were influenced too little from the extreme values. The MLE estimates are more sensitive to changes in the extreme values but still robust. We will focus on the MLE estimation in the next illustration.

We have taken a random sample and multiplied the top observation by different constants. A multiplication by a small number is equivalent to removing the largest value and a multiplication by a large number corresponds to the situation when the top sample value is unexpectedly high. The results are shown in Table 6. When the top value is removed, the estimates of the mean, the Gini coefficient and the S80/S20 ratio based on the Parameter tail method changed only very little compared to the changes of the raw estimates based on the sample directly. These indicators are also quite robust to one extreme observation. In the same way the estimates of the standard deviation and the variance coefficient behave much more sensible. However we noted after Table 5 that those estimates had an unacceptable bias. They seem to have so here too.

Table 6. Welfare measures in a data set with the top value changed by a factor (CONT). (First line is raw estimate from changed sample. Third line is MLE Pareto tail estimate).

Indicators		**Mean**	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
Original sample		174051	767069	4,407	0,344	5,91
OBRE	**1,90**	165273	157901	0,955	0,309	5,19
MLE	**1,84**	165592	165324	0,998	0,310	5,22
CONT	**MLE θ**	**Mean**	**Std Dev**	**Var. Coeff**	**Gini coeff**	**S80/S20**
		166871	254608	1,526	0,316	5,32
0,05	**1,95**	165002	151861	0,920	0,308	5,17
		167524	264225	1,577	0,318	5,37
0,1	**1,93**	165134	154766	0,937	0,309	5,18
		168975	333785	1,975	0,324	5,49
0,3	**1,89**	165349	159624	0,965	0,309	5,20
		170425	441691	2,592	0,330	5,61
0,5	**1,87**	165450,98	161993	0,979	0,310	5,205
		171875,70	566438	3,296	0,336	5,731
0,7	**1,86**	165519,28	163594	0,988	0,310	5,211
		176227,32	974999	5,533	0,352	6,087
1,3	**1,84**	165646,83	166622	1,006	0,311	5,221
		178403,14	1186395	6,650	0,360	6,265
1,6	**1,83**	165690,17	167663	1,012	0,311	5,225
		181304,22	1471065	8,114	0,370	6,502
2	**1,82**	165737,06	168795	1,018	0,311	5,229
		188556,94	2188996	11,609	0,394	7,095
3	**1,8**	165823,12	170892	1,031	0,311	5,236
		203062,36	3633616	17,894	0,438	8,281
5	**1,79**	165933,1	173605	1,046	0,312	5,245
		239325,93	7255512	30,316	0,523	11,247
10	**1,77**	166085,30	177419	1,068	0,313	5,257

7. Conclusions

Classical adjustments are easy to implement, but especially trimming is trading-off too much valid data information in order to get robust estimates. Winsorizing tends to be less sensitive and it brings markedly more stable estimates. The approach based on modelling a parametric-tail distribution performs quite well and it is promising method for outlier treatment, especially in highly skewed population.

The OBRE estimator provides truly robust estimates of parameter θ in the Pareto distribution tail. One should certainly use the OBRE estimator for θ

estimation in significantly skewed and severely contaminated population where this method surpasses the MLE estimator. With contaminated we mean the situations when the skewness depends on large measurement errors. However, in our application the MLE-method for θ estimation provides slightly better estimates, since the outliers are not incorrect values.

There still remains a bias when estimating the mean and standard deviation. It seems as if the method yields too small estimates. This may depend on the fact that the tails do not follow a Pareto distribution exactly but have longer tails. This can be seen from the fact that θ–values estimated from a sample of 100 are smaller than those based on the 250 top values. One may remedy this either by using another distribution e.g. the lognormal one or base the θ–estimate on even fewer values, say 50.

One may also see that it does not seem to be necessary to replace all the 100 or 50 top values by quantiles from the Pareto tail distribution. It seems to be more sensible to replace a smaller number, say 10 values and to use the 11th value as the lower limit of the distribution. The suggested improvements will certainly decrease the bias further but the standard error estimation will certainly increase more. This should be explored further.

If one has reason to believe that the form of the distributions stays the same over time one should of course use previous years to get a better and more stable θ–estimate. But for this income distribution the shape of the tail is probably too dependent on the economic cycle and price changes on the stock and real estate markets.

Acknowledgements

We want to express our deep gratitude to Statistics Sweden who kindly provided us with the data and helped us in other ways. In particular we want to mention Susanne Svartengren, Michael Carlson and Dan Hedlin.

REFERENCES

CHAMBERS, R. & KOKIC, P.N. (1986) Outlier robust sample survey inference, *Bulletin of the International Statistical Institute* **55**, pp 55—72, Firenze.

DALÉN, J, (1987), Practical estimators of a population total, which reduce the impact of large observations, R&D report U/STM-32, Statistics Sweden

HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. & STAHEL, W. A. (1986), *Robust statistics: The approach based on influence functions*, Wiley, New York.

HUBER, P. J. (1972), Robust statistics: A review, *Annals of Mathematical Statistics* **43** pp 1041—67.

HUBER, P. J. (1981), *Robust statistics*, Wiley, New York.

JANSSON, K, (1999), Income Distribution Data for Sweden, Robustness Assessment Report, Internal Memo, Statistics Sweden, Örebro, Sweden

KARLBERG, F. (1999), *Survey estimation for highly skewed data*, PhD-thesis, Stockholm University, Stockholm

KARLBERG, F. (2000), Survey estimation for highly skewed populations in the presence of zeroes, *Journal of Official Statistics,* **16,** pp 229—242

VAN KERM, P, (2006), Extreme incomes and the estimation of poverty and inequality indicators from EU-SILC, Paper presented at Eurostat and Statistics Finland International Conference on "Comparative EU Statistics on Income and Living Conditions: Issues and Challenges" Helsinki, 6—7 November

LEE, H. (1995), Outliers in Business Surveys, in *Business Survey Methods,* ed Cox, Binder, Chinnappa, Christianson, Colledge, Kott, pp 503—526, Wiley, New York

THORBURN, D. (1991), Model-based estimation in survey sampling of lognormal distribution, in *A spectrum of statistical thought, Essays in statistical theory, economics and population genetics in Honour of Johan Fellman,* pp 228—43, Swedish school of economics, Helsinki.

VICTORIA-FESER, M-P. & RONCHETTI, E., (1994), Robust methods for personal income models, *The Canadian Journal of Statistics* **22,** pp 247—58

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 155—172

SPATIAL DIVISION OF LABOUR: AN EMPIRICAL EVIDENCE FROM POLAND

Henryk Gurgul[1] and Paweł Majdosz[2]

ABSTRACT

The paper is aimed at examining spatial division of labour in Poland. Defining a region's ratio of actual employment to a predicted one and using a non-survey method to estimate the latter, we provide evidence that some regions tend to specialize in jobs requiring low educated workers whereas other are more likely to specialize in the high-skilled jobs associated with a high educational level of employees. Furthermore, it turned out that there exists a statistically significant relationship between the rural/urban-type of a region and the direction of its specialization; while more rural regions specialize in low-skilled jobs, urban regions specialize in high-skilled jobs.

1. Introduction

The division of labour is an intrinsic element of classical economic thought and it may be surprising that spatial-differential of division of labour is still one of the contemporary economists' concerns in the economic literature. Common-sense intuition prompts us to believe that rural areas are characterized by a higher portion of jobs requiring low skills and a lower portion of high-skill jobs. In urban areas, on the other hand, the same proportions are expected to be exactly the reverse; such areas should be relatively more specialized in high-skill jobs than in jobs requiring low skills. Examining the trend of rural-urban shifts in jobs requiring some-college or higher educational attainment, McGranahan and Ghelfi (1998) found evidence supporting, at least partially, the hypothesis of a spatial differential in the division of labour. More recently, Wojan (2001) documented that specialization trends observed in rural and urban areas coincide with what should be expected when assuming that the division of labour is spatially

[1] Department of Economics and Econometrics, University of Science and Technology, Poland. h.gurgul@neostrada.pl

[2] Department of Economics and Econometrics, University of Science and Technology, Poland. pmajdosz@go2.pl

differentiated. In contrast to this, Loveridge and Christiadi (2000) claim that it might be due to the methodology used in the previous work, which assumes a fixed input structure of occupation per sector at both national and local level, that a spatial-differential in the division of labour emerged. A further investigation carried out by the above-mentioned authors, which was based on survey data provided some evidence contradicting the hypothesis that rural areas are more specialized in low-skill jobs and urban areas in high-skill jobs.

The Polish economy can be seen as providing a unique opportunity for a re-examination of the hypothesis of a spatial differential in the division of labour for several reasons.. Firstly, Poland is one of the countries which at the beginning of the 1990s were forced to start the transition process of their economies from a centrally planned to a market model. When it was centrally planned what, how and for whom goods will be produced, it is not surprising that also the decisions concerning the location of individual enterprises supplying commodities to intermediate and final users were mostly determined by political issues, not pure market stimulus. Consequently, in the course of years, urban areas have became more ‘urban’ as centrally controlled governmental investments proceeded, which were deprived of such investments, rural areas have became even more ‘rural’. One of the observable effects of the transition process is, obviously, a decrease in the differences between urban and rural areas when decisions about the location of production are based on market mechanisms and not political relationships. However, the existence of a spatial division of labour is still more likely in Poland in the light of all these facts than would otherwise be possible.

Secondly, it is a well-known fact that the absence of barriers in access to education after the second world war has resulted in a relatively higher level of qualification of employees in Poland as compared with other countries (see e.g. Balcerowicz, 1995; Kołodko, 1999). In addition to this, the great emphasis which was at that time placed on the issue of equalizing educational attainment independently of which areas (i.e. urban or rural) people came from has probably contributed to an even distribution of educational attainment among different kinds of areas. This means that the problem of the spatial differential in the division of labour in Poland gains in importance and indicates a burning need scrutinize the level of regional specialization in Poland after this country started its reforms over sixteen years ago.

It is perhaps useful to specify more clearly right at the beginning of this paper what is meant by regional specialization and how it is measured under this study. If a certain region has a distribution of educational requirements associated with a occupations requiring differential skills which is close to the national average, the region’s specialization is implausible. In contradistinction to this case, the existence of substantial differences between a region’s educational requirements on the one hand and the national average, on the other, indicates that the region is specialized in some jobs at the expanse of others. For example, with more employees having at least a Master’s degree and fewer of those with only elementary education in the region when compared with the economy as a whole, taking into account both the level as well as structure of the regional production, it can be said that the region is characterized by a specialization in high-skill jobs at the expanse of those requiring low skills. There are two reasons why we decided to base our examination on educational requirements rather than directly on occupations. One of them is that the regional statistics available in Poland provide information concerning the number of employees by level of educational attainment but not by occupation. The latter data are accessible only at the national level. Another reason stems from the fact that the names of occupations in themselves do not necessarily reflect skill differentials.

In order to measure a regional level of specialization, we followed Wojan (2001) and defined it as the ratio between a region's real employment and its predicted employment for a given level of educational attainment. It also deserves to be emphasized that Wojan originally used local employment by occupation. Expected regional employment is obtained by means of the non-survey method proposed by Goetz and Debertin (1993) which imposes the assumption of a fixed distribution of the occupations demanded by each individual sector and of a fixed distribution of educational attainment demanded by each individual occupation.

Our analysis reveals that some regions tend to specialize in jobs requiring the lowest educated workers whereas other are more likely to specialize in high-skilled jobs associated with a high education level of employees. As regards the second main question of this study, that is, whether or nor urban regions tend to specialize in high-skilled jobs while low-skilled job specialization characterizes mostly rural regions, it turned out that there exists a statistically significant relationship between the rural/urban-type of a region and the direction of its specialization and that its sign coincides with our initial presumptions.

The remainder of this paper proceeds as follows. Section 0 is devoted to a brief description of the data set used under this study and also provides an outline of the methodology employed to assess an individual region's level of specialization and statistically test its significance. In Section 0 we present empirical results concerning the subject under consideration and the last section concludes the paper and discuses fields left for further investigation in the future.

2. Research design

2.1.Identification of a region's specialization degree

As mentioned above, a certain region's level of specialization is measured as a ratio of the actual employment within a given type of educational attainment to that of what is predicted using the non-survey method put forward by Goetz and Debertin (1993). Out of the assumptions underlying this method two are of striking importance in identifying the degree of specialization in regions. Of these the first is that the distribution of occupations across sectors in each individual region is fixed and coincides with that of the distribution derived at the national level. The second assumption, on the other hand, imposes a fixed distribution of educational attainment across occupations in each individual region which is identical to the corresponding distribution for the economy as a whole (at the national level). Taking this into account, if the values of our measure of regional specialization fluctuate close to unity for any type of educational attainment then it implies there is no specialization in a region. By contrast, if the values differ substantially from one then it indicates that a region specializes in some jobs for which the required educational attainment is accompanied by regional employment in excess of the predicted one.

Employing Goetz-Debertin's method requires data of two types. Both of them are easy available in Poland from the published statistics. At the national level we need the number of employees for each occupation by industry and for each level of educational attainment by occupation. As regards regional data, only the employment by industry is indeed needed.

Let *k*, *m*, *n* denote the number of selected occupations, of levels of educational attainment and of industries, respectively. Assuming $\mathbf{Z}_{(k\times n)}$ and $\mathbf{W}_{(k\times m)}$ to be matrices which contain the total national employment presented in an occupation-by-industry and occupation-by-educational attainment form, the national distribution of occupations within each individual sector ($\mathbf{\Omega}$) and the corresponding distribution of educational attainment for each individual occupation ($\mathbf{Q}$) can be expressed as follows:

$$\mathbf{\Omega} = [\omega_{ij}] = \mathbf{Z}\left(\widehat{\mathbf{1}^T\mathbf{Z}}\right)^{-1}, \tag{1}$$

$$\mathbf{Q} = [q_{ip}] = \left(\widehat{\mathbf{W1}}\right)^{-1}\mathbf{W}, \tag{2}$$

where **1** is a summation column-vector of appropriate dimension and the circumflex denotes the diagonal matrix formed from the vector.

In the next step, combining the first two matrices (1) and (2), a block-matrix ($\mathbf{D}$) is derived:

$$\mathbf{D}_{(kxmn)} = \left[\begin{array}{ccc:ccc:c:ccc} \omega_{11}q_{11} & \cdots & \omega_{11}q_{1m} & \omega_{12}q_{11} & \cdots & \omega_{12}q_{1m} & \cdots & \omega_{1n}q_{11} & \cdots & \omega_{1n}q_{1m} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots \\ \omega_{k1}q_{k1} & \cdots & \omega_{k1}q_{km} & \omega_{k2}q_{k1} & \cdots & \omega_{k2}q_{km} & \cdots & \omega_{kn}q_{k1} & \cdots & \omega_{kn}q_{km} \end{array}\right],$$

where each of *n* blocks represents an individual industry and for *j*th industry the corresponding block of it ($\mathbf{D}^{<j>}$) can be obtained by multiplying the diagonal matrix formed from *j*th column of matrix $\mathbf{\Omega}$ by $\mathbf{Q}$:

$$\mathbf{D}^{<j>} = \widehat{\mathbf{\Omega}}^{j}\,\mathbf{Q}. \tag{3}$$

Let $\mathbf{E}^l = [e^l_j]$ be a column-vector of employment by industry in region *l*, $l = 1,...,r$, then the regional counterparts of the matrices **Z** and **W** can be derived as follows:

$$\mathbf{Z}^{lj}=e_j^l\mathbf{D}^{<j>}\mathbf{1}, \tag{4}$$

$$\mathbf{W}^l=\sum_{j=1}^{n}e_j^l\mathbf{D}^{<j>}. \tag{5}$$

Some authors who have used the same method (Christiadi, 2004; Loveridge and Christiadi, 2000) suggest an adjustment of the elements within the blocks of **D** before calculating the local equivalents of **Z** and **W** to ensure that

(i) the blocks of **D** sum up to unity for each individual industry.

However, as will be shown, such an adjustment is completely redundant because it is impossible that the sum of elements located within a given block of **D** deviate from unity when using the method described above. In addition, employing the method yields results which are consistent with the respective national totals, that is,

(ii) summing z^l_{ij} (for any i and j) across regions yields z_{ij}, and

(iii) the sum of w^l_{ip} (for any i and p) for all regions is equal to w_{ip}.

To show this, notice first that:

$$\sum_{i=1}^{k}\omega_{ij}=1,\ \forall j \text{ or equivalently } \mathbf{1}^T\boldsymbol{\Omega}=\mathbf{1}^T, \tag{6}$$

$$\sum_{p=1}^{m}q_{ip}=1,\ \forall i \text{ or in an equivalent form } \mathbf{Q1}=\mathbf{1}, \tag{7}$$

$$\sum_{j=1}^{n}z_{ij}=\sum_{p=1}^{m}w_{ip},\ \forall i \text{ or } \mathbf{Z1}_{(n\text{x}1)}=\mathbf{W1}_{(m\text{x}1)}, \tag{8}$$

$$\sum_{l=1}^{r}e_j^l=\sum_{i=1}^{k}z_{ij},\ \forall j\,. \tag{9}$$

Expressing the condition (i) as

$$\mathbf{1}^T\mathbf{D}^{<j>}\mathbf{1}=\mathbf{1}^T\ \square\boldsymbol{\Omega}^j\mathbf{Q1},$$

it becomes immediately obvious (using (6) and (7)) that this condition must always hold true.

The second condition (ii) can be written as

$$\omega_{ij}\sum_{l=1}^{r}e_j^l\sum_{p=1}^{m}q_{ip}=z_{ij},\ \forall i,j\,.$$

Applying (7) and (9), we obtain $\omega_{ij}\sum_{i=1}^{k}z_{ij}=z_{ij}$, because $\omega_{ij}=z_{ij}/\sum_{i=1}^{k}z_{ij}$.

And finally, there is the last condition (iii), which can be given by

$$q_{ip}\sum_{j=1}^{n}\omega_{ij}\sum_{l=1}^{r}e_j^l=w_{ip},\ \forall i,p\,.$$

As has been shown above, however, $\omega_{ij}\sum_{l=1}^{r} e_j^l = z_{ij}$, and thus the left-hand side of (iii) can now be rewritten as $q_{ip}\sum_{j=1}^{n} z_{ij}$. Applying (8) ultimately yields $q_{ip}\sum_{p=1}^{m} q_{ip} = w_{ip}$ because what stems from the way the coefficient q_{ip} is calculated is that $q_{ip} = w_{ip} / \sum_{p=1}^{m} w_{ip}$.

2.2.Data

The survey of economical activities of people living within the boundaries of Poland which was carried out by the Central Statistical Office (CSO) in the middle of 2002 (the resultant publication followed at the end of 2003) forms the basis of our investigation in terms of data needed at the national level. This source provides the total employment registered for the economy as a whole in a form which is suitable for the purpose of this study, that is, arranged on an occupation-by-industry as well as occupation-by-educational attainment basis. The applied division of the economy into sectors followed the sections of NACE (see Appendix B). Occupations were grouped according to the Classification of Occupations and Specialities (COS) introduced at the end of 2002. The survey distinguished eight different levels of education from those having at least a doctoral and master's degree to those of primary and incomplete primary education.

Poland was divided into sixteen regions (corresponding to the two-digits Geopolitical Entity Classification) differentiated with respect to area, population, share of national output and of national employment as well as what is considered as the dominant activity within a region. Among them are such regions as Lubelskie and Podkarpackie which are typically rural and metropolises such as Śląskie and Małopolskie. In Appendix A a map and some useful quantitative characteristics offering a more comprehensive insight into the Polish regions are presented. Regional employment by industry and by level of educational attainment for the corresponding year come directly from the Bank of Regional Data (BRD).

Minor adjustments to the original data had to be made at the very beginning of the analysis to ensure the complete consistency of the data at the national level with those concerning the regions. After finding that the number of sectors specified by the CSO's survey of economical activities exceeds by two that used within the regional employment statistics published by the BRD, an appropriate aggregation of the data has been employed which left us with 15 sectors. Then, the initial number of selected occupations has been limited to 9 by omitting the last two entries called “Employed in military” and “Undetermined occupation”. Finally, due to the fact that the number of levels of educational attainment used within the CSO’s publication differs from that applied by the BRD statistics, the respective data had to be aggregated to reconcile both data sources with one another.

The aggregated data, in the way described above, provide, however, controlling sums of employment by sector and of employment by levels of educational attainment which slightly deviate from the national values. Therefore, the bi-proportional method (the RAS) was used here so that summing an industry’s employment across occupations as well as across levels of educational attainment yields values which are equal to the national totals.

3. Empirical Results

We start the empirical investigation of the spatial division of labour in Poland by calculating separately for each of the regions the predicted distribution of employment of each occupation category by educational attainment by means of the non-survey method of Goetz and Debertin described previously. Summing each individual entry indexed by occupation and educational attainment of a given level within an individual industry across the regions and then rearranging industries in such a manner as to obtain only four major sectors yields the figures presented in Table **1**. We decided not to tabulate the results in all possible divisions, that is, into regions, industries, occupations and educational attainment so as to keep the paper short, however, they all are available from authors on request.

Table 1. Predicted distribution of employment across occupations and by educational attainment in four main sectors of the Polish economy in 2002

	Occupation category								
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Agriculture, hunting, forestry and fishing									
A	0.35	0.43	0.37	0.08	0.01	1.41	0.02	0.03	0.01
B	0.34	0.07	1.26	0.33	0.10	15.83	0.42	0.62	0.39
C	0.08	0.02	0.24	0.13	0.03	2.85	0.05	0.10	0.11
D	0.08	0.00	0.09	0.08	0.08	32.54	0.93	1.20	0.92
E	0.01	0.00	0.02	0.02	0.02	37.18	0.17	0.33	0.64
Industry									
A	2.64	4.77	1.96	0.82	0.08	0.00	0.67	0.25	0.04
B	2.54	0.81	6.68	3.39	0.58	0.01	11.64	5.09	1.15
C	0.61	0.25	1.28	1.38	0.19	0.00	1.43	0.81	0.34
D	0.62	0.01	0.45	0.81	0.50	0.01	25.67	9.86	2.74
E	0.09	0.00	0.08	0.19	0.12	0.01	4.81	2.73	1.91
Construction									
A	3.72	3.97	1.53	0.41	0.02	0.00	0.84	0.10	0.08
B	3.57	0.68	5.21	1.72	0.14	0.00	14.71	2.06	1.98
C	0.86	0.21	1.00	0.70	0.05	0.00	1.80	0.33	0.59
D	0.87	0.01	0.35	0.41	0.12	0.00	32.45	3.99	4.74
E	0.12	0.00	0.06	0.10	0.03	0.00	6.09	1.10	3.30
Services									
A	3.15	17.75	3.31	1.48	1.10	0.00	0.09	0.09	0.05
B	3.02	3.03	11.29	6.14	8.04	0.01	1.56	1.88	1.42
C	0.72	0.92	2.16	2.49	2.66	0.00	0.19	0.30	0.42
D	0.74	0.05	0.77	1.47	6.97	0.02	3.44	3.65	3.39
E	0.10	0.00	0.14	0.34	1.61	0.02	0.64	1.01	2.36

Occupation categories: Legislators, senior officials and managers – [1], Professionals – [2], Technicians and associate professionals – [3], Clarks – [4], Service workers and shop and market sales workers – [5], Skilled agricultural and fishery workers – [6], Craft and related trades workers – [7], Plant and machine operators and assemblers – [8], Elementary occupations – [9]. Levels of educational attainment: Higher education – A, Post-secondary and Vocational secondary – B, General secondary – C, Basic vocational – D, Lower secondary, Primary and incomplete primary – E. Note that all presented values are expressed in percentages.

Before we go into details, several general trends observable in the figures should be pointed out. It is a well-documented fact that Services has attracted, mostly due to relatively higher wages offered by this sector, better educated people at the expense of other sectors of the economy (see e.g. Lin and Christiadi, 2002). This finding is likely to hold as well in Poland when comparing the ratios of people with higher education, post-secondary and vocational secondary (levels A and B) across the sectors. One can find that for Services the above-mentioned ratio equals 63.4%, whereas Agriculture, Industry and Construction are characterized by a substantially smaller portion of highly educated people (respectively 22.6, 43.1 and 40.7). In addition to this, at the end of 2004 an average monthly wage was indeed slightly higher in Services (2,449 zł compared with 2,216 zł, the average, in the others). Certainly, a scrupulous investigation of this phenomenon should provide further depth in understanding the occupational mobility emerging in the Polish economy but these issues are beyond the scope of this paper.

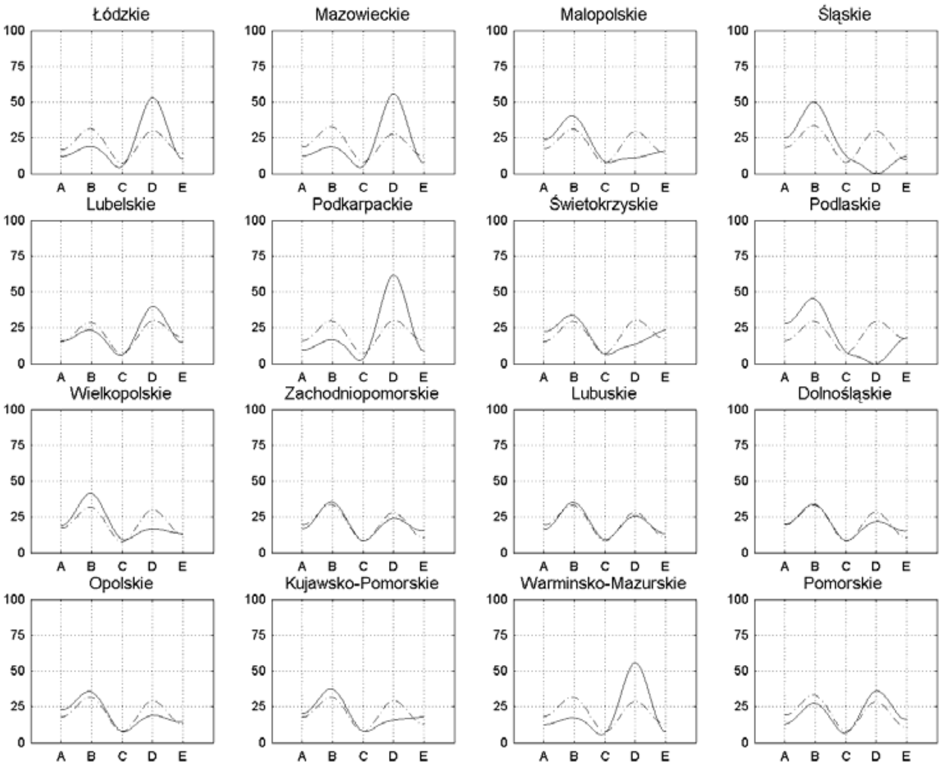
Another important indication which is given particular attention by policy makers is the ratio of the lowest-skilled workers, that is, those with primary and incomplete primary education (level E). The static nature of this study does not allow for examining trends of the respective data which may appear to be particularly revealing and, therefore, should be the subject of further investigation. Instead, the number of workers having the lowest level of educational attainment can, however, be compared in the overall employment for each of the selected sectors. With the exception of Agriculture, the ratio of such workers oscillated around 10% in the Industry and Construction sectors and was the lowest in Services, slightly exceeding 6%. In Agriculture the portion of the least educated workers was much higher and amounted to almost 40%. A gradual decline in this ratio should, however, be expected, also in Agriculture, as an older generation of workers will be successively replaced with new ones much better educated than their predecessors.

So far the attention has been focused on educational attainment without any reference to its occupational context. Now, attention will be extended to the predicted occupational structure of workers with different levels of education. As might be expected, in the Agriculture sector the portion of workers employed as agricultural and fishery workers dominated (almost 90% of the sector's total employment) of which up to over 40% have the lowest educational level (Primary and incomplete primary) and only 1.5% the highest (Higher education). The occupational structure of Services appeared to be much more differentiated but it can be found that the first three top-reward jobs, namely legislators and managers, professionals, technicians and associated professionals (occupation category from 1 to 3) account for almost a half of the total employment in this sector. At the same time, over a half of the persons employed in the above-mentioned sectors had the highest level of educational attainment.

The remaining two sectors, namely Industry and Construction one somewhere in the middle between the Agriculture sector on the one hand and Services on the other. In both sectors approximately 70% of the sector's total employment are attributable to the last three lowest-reward jobs (craft and related trades workers, plant and machine operators and assemblers as well as elementary occupations). As regards the educational attainment structure experienced by these sectors, it deserves to be noticed that again in both of them neither the highest nor the lowest educational level was superior to the others. On the contrary, both higher education as well as primary and incomplete primary education accounted only for slightly above 10% of the total employment in the respective sector. Of prime importance in these sectors, in terms of shares in the overall employment, are post-secondary and vocational secondary as well as low secondary education levels (using a notation introduced in Table **1**, this refers to B and D levels), explaining together almost 60% of the sectors' total employment.

Summing for a region the predicted employment across occupations gives the predicted distribution of educational attainment. This estimated distribution is then confronted with the real distribution of the corresponding region obtained directly from the BRD. Figure **1** depicts the results. As was mentioned in the previous section, the BRD statistics do not contain data on regional employment by occupation, making it impossible to compare the real and expected employment for each of the individual occupation categories across regions.

Figure 1. Comparison of real and predicated distribution of educational attainment across the regions



Levels of educational attainment: Higher education – A, Post-secondary and Vocational secondary – B, General secondary – C, Basic vocational – D, Lower secondary, Primary and incomplete primary – E. Note that a solid line represents the real distribution of educational attainment whereas a dashed line that of predicted.

It is worth recalling here that according to the definition introduced in Section 0, any deviation of the real distribution of educational attainment from the predicted one should be recognized as evidence of regional specialization. Keeping this in mind, the Polish regions can be divided into three main categories/groups. The first of them is composed of those regions where both distributions (i.e. predicted and real) coincide with each other, for example, Zachodniopomorskie, Lubuskie and to some extent also Dolnośląskie, thereby indicating that a region does not specialize in any occupation. Another group consists of regions in which the portion of high-skilled jobs requiring educational attainment of levels A and B is higher than what is suggested by the non-survey method and, at the same time, the portion of low-skilled jobs associated with educational attainment of levels D and E is below the predicted one. Such regions as Małopolskie, Śląskie, Świętokrzyskie, Podlaskie, Wielkopolskie, Opolskie, and Kujawsko-Pomorskie belong to this group which will be referred to as "High-skilled job specialization – (HSJ)". And finally, the last category, called "Low-skilled job specialization – (LSJ)", groups such regions as Łódzkie, Mazowieckie, Lubelskie, Podkarpacie, Warmińsko-Mazurskie, and Pomorskie; in these cases the relations of the observed high-skilled and low-skilled jobs to the predicted counterparts are exactly reversed compared to the High-skilled job specialization group, that is, the portion of high-skilled jobs is below and of low-skilled jobs is above what is derived from the estimation.

It is worth noticing that the middle level of educational attainment (i.e. general secondary education (C)) surprisingly appears to be well predicted by means of the non-survey method in all the regions under consideration with only slight errors represented by distances between a solid line (for the real distribution) and a dashed line (for the predicted one) at this point.

To facilitate a better insight into the relative significance of educational attainment of different types for a given region in terms of its degree of specialization, a region's ratios of the real to the predicted employment for each level of educational attainment were calculated (hereafter it will be referred to as "specialization ratio"). Then, the resultant ratios were ranked in decreasing order across the regions, for each of categories separately, so that a number one assigned to a certain level of educational attainment means the region having the highest relation of the number of employees with such a education level to that of the expected one. The results are reported in Table **2**.

Table 2. Comparison of regions' degree of specialization

	Level of educational attainment				
	A	B	C	D	E
Łódzkie	0.715 [12]	0.612 [13]	0.669 [14]	1.777 [4]	0.737 [13]
Mazowieckie	0.654 [15]	0.577 [14]	0.626 [15]	2.000 [2]	0.636 [14]
Małopolskie	1.369 [3]	1.293 [4]	1.181 [4]	0.374 [14]	1.097 [9]
Śląskie	1.356 [4]	1.492 [2]	1.599 [1]	0.000 [15]	1.208 [7]
Lubelskie	1.029 [9]	0.812 [12]	0.881 [11]	1.336 [5]	0.778 [12]
Podkarpackie	0.576 [16]	0.557 [15]	0.468 [16]	2.048 [1]	0.531 [16]
Świętokrzyskie	1.476 [2]	1.133 [6]	0.957 [10]	0.444 [13]	1.329 [5]
Podlaskie	1.765 [1]	1.523 [1]	1.206 [3]	0.000 [15]	1.007 [11]
Wielkopolskie	1.095 [7]	1.299 [3]	1.254 [2]	0.553 [11]	1.026 [10]
Zachodniopomorskie	0.852 [10]	1.062 [9]	1.016 [9]	0.854 [8]	1.458 [2]
Lubuskie	0.835 [11]	1.064 [8]	1.144 [5]	0.907 [7]	1.237 [6]
Dolnośląskie	1.036 [8]	1.018 [10]	1.071 [6]	0.767 [9]	1.455 [3]
Opolskie	1.276 [5]	1.115 [7]	1.024 [8]	0.652 [10]	1.098 [8]
Kujawsko-Pomorskie	1.143 [6]	1.174 [5]	1.059 [7]	0.534 [12]	1.389 [4]
Warmińsko-Mazurskie	0.677 [13]	0.547 [16]	0.832 [12]	1.942 [3]	0.592 [15]
Pomorskie	0.672 [14]	0.831 [11]	0.820 [13]	1.262 [6]	1.564 [1]

Levels of educational attainment: Higher education – A, Post-secondary and Vocational secondary – B, General secondary – C, Basic vocational – D, Lower secondary, Primary and incomplete primary – E. The ranks in a decreasing order are shown the square brackets.

As should be expected, the figures presented in Table **2** strictly coincide with what has been said previously. All the regions belonging to the HSJ have a specialization ratio greater than one in the group of best educated employees. Of the highest specialization degree among these regions is Podlaskie which employed 1.8 times as many people with higher education as would be in the case of an absence of specialization. The following three positions on this list are taken by such regions as Świętokrzyskie, Małopolskie, and Śląskie with their specialization ratio being respectively equal to 1.48, 1.37, and 1.36. Taking into account that as far as the two sequential education categories labelled by B and C (post- and vocational secondary and general secondary level) are considered, one can find the same regions to be ranked first to fourth, albeit with a slight permutation in their order, the recognition of them as regions specializing in high-skilled jobs seems to be particularly valid.

But what can be further said about the regions having the highest specialization ratios within top and mid level educated employees? Podlaskie is located in the north-eastern part of Poland and takes in an area of over 20 thousand square kilometre. Its attractiveness with natural resources favouring the development of manufacture of food and beverages as well as the tourism sector, and its strategic location in regards to the Lithuanian and Belorussian markets combine to make it a desirable place to live and work. The key sectors of this region include the manufacture of machinery, textiles and tobacco products. The manufacture of food and beverages, especially milk and spirits, is also of crucial importance for the region's economic background. What distinguishes the three remaining regions is that they all are located in the southern part of Poland and border on each other. Their geographical proximity might contribute to a better transfer of new technology between the neighbouring regions and thereby to maintaining their fast development. Not accidentally, Śląskie is recognized as the most industrialized and urbanized region in Poland with coal mining and metallurgy – steel, tin and lead – being key industries. In addition to these traditional branches, such sectors as electricity, motor vehicle, textiles and chemicals are significantly represented in the region's total output. As regards Małopolskie and Świętokrzyskie, while both regions specialize in metallurgy and food manufacture, the former is considered as a banking and high-tech technology centre whereas the latter's powered sectors are machine, textiles and precision sectors.

With regard to low-skilled job specialization, one can find that Podkarpackie is ranked first with respect to a specialization ratio value in the group of workers with basic vocational education. This region is followed by Mazowieckie and Warmińsko-Mazurskie. Each of these regions employed approximately twice as many workers having only basic vocational education as what would be the case assuming the regions not to be specialized. In the case of workers with the lowest level of education (E) the top two positions of the ranking are taken by the following regions: Pomorskie and Zachodniopomorskie whose specialization ratios amount respectively to 1.56 and 1.46. As might be expected, all the above-mentioned regions were previously grouped into the LSJ.

What features are characteristic of the regions specializing in jobs which require low-educated workers? Podkarpackie, taking in an area of almost 18 thousand square kilometre in the South-eastern Poland, is a typical rural region specializing in the production of meat, corn, fruit and vegetables as well as milk and sugar. Although in the course of last decade the aviation sector concentrated in such cites as Rzeszów, Krosno and Mielec, has gained in significance, over 60% of the population of this region sill lives outside its main cities. Of the same rural-type is Warmińsko-Mazurskie, located in the north of Poland, whose greatest economical potential lies in agriculture and food manufacture. Complementary industries include the manufacture of wood and textiles. Warmińsko-Mazurskie also stands out as one of the regions having the biggest potential for tourism development in Poland. On the face of it, it might be a bit surprising that Mazowieckie was grouped into the LSJ in that in this region is located the capital of Poland, Warsaw, with its concentration of so important for proper functioning the economy as a whole institutions as the Warsaw Stock Exchange and with the highest expenditures on R&D. However, it should be recalled that Mazowieckie, located in the centre of Poland, is simultaneously one of the largest regions in Poland with an area equal to over 35 thousand square kilometre. Almost 70% of the region's area, that is approximately 23 thousand square kilometre (the largest out of all the regions), is utilized as a cropland. It is the key to explain why despite the high level of economic social, and cultural development of some parts of Mazowieckie, the region as a whole was classified as specializing in jobs which require low-educated workers.

The two last regions having the highest specialization ratio within the group of the lowest educated workers, namely Pomorskie and Zachodniopomorskie are located in north-western Poland. Both regions specialize mainly in the shipbuilding sector and fishing. Having direct access to the sea, they also are characterized by the great potential to develop their tourism industries.

Before we move to a more formal testing of the hypothesis that typically rural regions specialize in low-skilled jobs whereas urban regions tend to specialize in high-skilled jobs, note that the regions with the highest values of specialization ratio in the case of best educated employees have, in general, the lowest values for workers with a low level of education and vice versa. Such a finding is, however, not surprising as they are two sides of the same coin.

So far, it has been shown that some of the regions specialize in high-skilled jobs requiring the best educated employees whereas others tend to specialize in jobs performed by workers with a low education level. In this part the main question is whether or not the rural regions tend to specialize in low-skilled jobs whereas high-skilled jobs are the domain of the urban regions. In order to ascertain this, we first had to decide what should be used as the measure of a region's degree of rurality/urbanisation. With a region's area of cropland serving as a proxy measure of rurality, the Spearman rank correlation R coefficient between this measure and a specialization ratio for each educational attainment category was calculated. This coefficient can be thought of as the well-known Pearson correlation coefficient, that is, in terms of the proportion of variability accounted for, except that R-Spearman is computed from ranks. Table **3** provides the R-Spearman values across educational attainment categories supplemented by test statistics and corresponding p-values.

Table 3. Results for testing significance of rank correlation

	R-Spearman	*t*-stat	*p*-value
Higher education – A	–0.488	–2.093	0.055
Post-secondary and Vocational secondary – B	–0.553	–2.483*	0.026
General secondary – C	–0.650	–3.200**	0.006
Basic vocational – D	0.511	2.222*	0.043
Primary and incomplete primary – E.	0.168	0.636	0.535
A + B	–0.526	–2.317*	0.036
D + E	0.665	3.329**	0.005

* significance at the 5% level

** significance at the 1% level

From the figures, one can find how far the regions characterized by low degrees of urbanisation are also sources of high ratios of specialization in jobs which require a certain educational attainment category. The signs of the *R*-Spearman coefficient confirm our initial presumption that the rural (urban) regions specialize in low-skilled (high-skilled) jobs. For example, the negative value of the coefficient in the highest educated employee category (A) means that the regions with the highest specialization ratios (placed at the top of the list) are, in general, listed at the end in terms of a region's area of cropland, and vice versa. Similarly, the positive value of the *R*-Spearman indicates that the regions listed at the top in terms of a region's specialization ratio are simultaneously at the top of the list based on a region's area of cropland, and vice versa.

Excluding of two extreme educational attainment categories (A and E), all the test statistic values are significant at least at the 5% level. Also, when jointing together the two highest and tow lowest education levels, the statistical significance of the test statistics can be observed. Concluding, it can be stated that the hypothesis that rural regions tend to specialize in low-skilled jobs whereas urban regions are more likely to specialize in high-skilled jobs finds support from our results.

4. Conclusions

Under this study we were mainly concerned with the issue of the spatial division of labour. Based on the latest available data referring to the sixteen regions in Poland, we started by ascertaining the differences between regional degrees of specialization using the ratio of the actual employment in a given region to the expected employment obtained by means of the non-survey method exploiting national proportions with respect to industries, occupations and education levels. It has been found that only in the case of three regions is there no observable divergence of the predicted employment from that existing in reality. Therefore, it has been concluded that regions mostly tend to specialize in jobs of a certain kind.

Although the profiles of the respective regions seemed to correspond with the initial presumption that rural regions specialize in low-skilled jobs as opposed to urban regions where a high-skilled job specialization is most plausible, formal statistical testing needed to be carried out to confirm or reject this hypothesis. Applying the Spearman rank correlation *R* coefficient has provided evidence supporting the hypothesis of a region's specialization depending on its rural/urban type.

The presented empirical results should, however, be handled with caution. This stems from the two crucial assumptions underlying the non-survey method used to estimate the predicted level of employment in the regions. As demonstrated by Loveridge and Christiadi (2000), it is possible that assuming a fixed distribution of occupations demanded by each individual sector and a fixed distribution of educational attainment demanded by each individual occupation, the method tends to overestimate the degree of specialization in regions. A relaxation of these assumptions of the original method and recalculating the results by means of an appropriately modified method are tasks which are left for future research.

Although Loveridge and Christiadi (2000) point to an urgent need for modifying the non-survey method originally proposed by Goetz and Debertin (1993) to make it more suitable for application under different distributions of occupations across sectors and different distributions of educational attainment across occupations, there are still few guidelines as to how such a modification might be carried out in practice. Furthermore, taking into account the data dearth, especially in Poland, a relaxation of the assumptions underlying the method seems indeed to be a very difficult task. Nevertheless, further research will be aimed at overcoming the above-mentioned shortcomings of the present study.

REFERENCES

BALCEROWICZ, L. (1995) *Wolność i rozwój. Ekonomia wolnego rynku* (Kraków, Wydawnictwo Znak).

CHRISTIADI, C. (2004) *Estimating Employer Demand for County Level Educational Attainment: A Case Study of Taylor County*, Research Paper No. 2004, West Virginia University.

GOETZ, S. J., DEBERTIN, D. L. (1993) *Estimating Country-Level Demand for Educational Attainment*, Socio-Economic Planning Science, 27 (1), 25—34.

KOŁODKO, G. W. (1999) *From Shock to Therapy. The Political of Postsocialist Transformations* (Warszawa, Poltext).

LIN, G. AND CHRISTIADI, C. (2002) *Interregion-Occupational Persistence and Dispersion: A Model of Geographic and Occupational Mobility*, Research Paper No. 2002-18, West Virginia University.

LOVERIDGE, S. AND CHRISTIADI, C. (2000) *A Comparison of Survey and Non-Survey Methods for Estimating Country-level Demand for Educational Attainment*, Research Paper No. 2021, West Virginia University.

MC GRANAHAN, D. A., GHELFI, L. M. (1998) *Rural Education and Training in the New Economy: The Myth of the Rural Skills Gap*. In R. M. Gibbs, P. L. Swaim, R. Teixeira (Eds.), Rural Education and Training in the New Economy: The Myth of the Rural Skills Gap (Ames, IA: Iowa State University Press).

WOJAN, T. R. (2001) *The Composition of Rural Employment Growth in the 'New Economy'*, American Journal of Agricultural Economics, 82, 594—605.

Appendix A. REGIONS IN POLAND

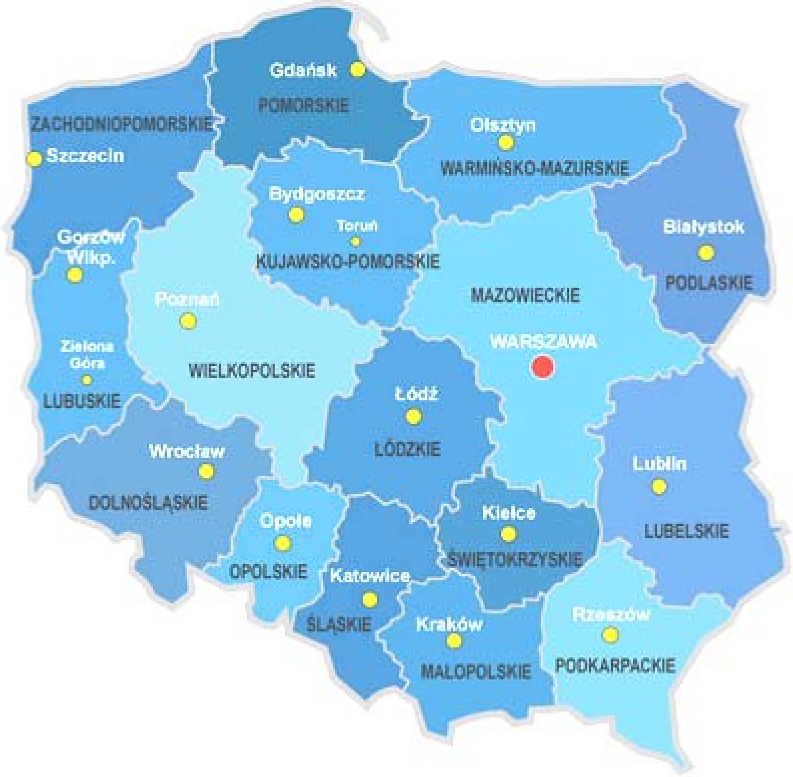


Table A. 1. Some quantitative characteristics of the regions in Poland

	Total output (in million Polish zloty)	Area (in thousand square kilometre)	Population (in thousand)	Employment in enterprises (in thousand)	Average wage in industry (in Polish zloty)
Łódzkie	100 702.5	18 219.0	2 577.5	273.6	2 173.0
Mazowieckie	359 426.0	35 579.0	5 157.7	1 051.0	3 129.0
Małopolskie	118 906.5	15 144.0	3 266.2	361.0	2 464.0
Śląskie	228 715.7	12 294.0	4 685.8	688.6	3 068.0
Lubelskie	63 290.1	25 114.0	2 179.6	150.1	2 123.0
Podkarpackie	64 245.2	17 926.0	2 098.3	205.2	2 126.0
Świętokrzyskie	38 771.7	20 186.0	199.7	92.7	2 185.0
Podlaskie	42 731.0	11 691.0	1 285.0	467.0	2 287.0
Wielkopolskie	37 482.8	13 984.0	1 009.2	102.6	2 119.0
Zachodniopomorskie	155 442.7	29 826.0	3 372.4	509.6	2 322.0
Lubuskie	65 547.4	22 902.0	1 694.2	154.7	2 258.0
Dolnośląskie	134 646.6	19 948.0	2 888.2	362.7	2 642.0
Opolskie	37 002.5	9 412.0	1 047.4	94.8	2 368.0
Kujawsko-Pomorskie	82 198.7	17 970.0	2 068.3	208.4	2 220.0
Warmińsko-Mazurskie	95 259.5	18 293.0	2 199.0	252.4	2 574.0
Pomorskie	47 377.8	24 203.0	1 428.6	133.6	2 065.0

All figures presented in the table refer to the first quarter of 2006 excepting for total outputs which were counted at the end of 2003 (the latest available data in the BRD).

Appendix B. Classification of economic activities – Sections of NACE

Code	Description
A	Agriculture, hunting and forestry
B	Fishing
C	Mining and quarrying
D	Manufacturing
E	Electricity, gas and water supply
F	Construction
G	Wholesale and retail trade; repair of motor vehicles, motorcycles and personal and household goods
H	Hotels and restaurants
I	Transport, storage and communication
J	Financial intermediation
K	Real estate, renting and business activities
L	Public administration and defense; compulsory social security
M	Education
N	Health and social work
O	Other community, social, personal service activities

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 173—189

THE GRAVITY MODEL — THE STUDY OF POLAND'S TRADE INTEGRATION WITH THE EUROPEAN UNION: TRADE CREATION AND TRADE DIVERSION EFFECT

Maria Majewska[1], Jolanta Grala-Michalak[2]

ABSTRACT

In this study, we use a version of the gravity model to analyze the effect of both trade creation and trade diversion on Polish external trade with European Union (EU) for the years: 1990, 1995, 1997, 1999, 2001 and 2003. The results show that both trade creation and trade diversion for Poland bilateral trade with EU have positive and statistically significant coefficients, which means that Poland have traded with outer-region countries as well as with intraregion countries above the hypothetical level and Poland trade integration with EU has not caused a negative trade diversion effect. It is also observed that the trade creation effects in the Poland exporting and importing activities are proved to be generally stronger in the case of the old EU (15 countries) than in the case of bilateral trade with the new EU.

Key words: trade creation, trade diversion, gravity model

1. Introduction

The European Union, previously European Economic Community (EEC) exemplifies the customs union trading nowadays on principles of free trade with the rest of the world. Its genesis was numerous post-war initiatives heading towards co-ordination of economic politics and towards the development of the countries of Western Europe.

The aim of European Economic Community was creating common market via limiting the trade barriers. EEC never was to become a protectionist area of free trade, though. Countries that suffered more due to war operations created it, e.g. France, West Germany, Italy, and that weakened their protectionist tendencies in

[1] Adam Mickiewicz University of Poznan, Poland; e-mail: majewska@amu.edu.pl.
[2] Adam Mickiewicz University of Poznan, Poland; e-mail: grala@amu.edu.pl.

the economic politics. Next to EEC in Western Europe there appeared in 1960 European Free Trade Association (EFTA), which was a free trade zone including initially Austria, Denmark, Norway, Portugal, Sweden, Switzerland and Great Britain. For many countries the membership in this regional trade group was mainly a transitional stage before joining the EEC. Today the members of EFTA are Island, Liechtenstein, Norway and Switzerland (see, for example: Magee and Lee, 2001).

The aim of this paper is studying the influence of Poland's integration with the European Union on trends of Poland's trade exchange. It will be studied via checking with the gravity model if the trade creation and trade diversion effects occur in terms of Poland. The first part of the paper deals with the integration effects in form of customs union and presents the EU situation in the very context. The second part of the paper describes the circumstances and the methodological assumptions of the paper concerning the gravity model, and the third part includes the material and the research results.

2. The customs union effects

Customs union is defined as a trade agreement reached in order to maximise the aggregated prosperity of all its members. According to this agreement all the member countries eliminate duty and quantitative limitations in common trade. At the same time these countries determine common duty barrier and uniform quantitative limitations distinguishing them from the rest of the world. At this point common external tariff is appointed on behalf of all the member countries by a central organ of union, which is equal with executing common trade politics towards other countries. The area thought to be even more protectionist than customs union is free trade zone, the perfect example of which is North America Free Trade Association (NAFTA), the strongest next to EU regional trade group. Distinct from customs union free trade area members by abolishing duty and quantitative limitations in common trade, determine their own external tariffs and quantitative limitations when being in contact with the rest of the world, maintaining control over their own trade politics towards the outer-region countries.

Therefore, in case of free trade area there usually is a lower level of economic and political integration comparing to custom union. It can easily be observed in a considerable number of regional trade groups arising in form of free trade area rather than in customs union (e.g. at the beginning of the 90-ties of the XX century according to GATT data there were 8 custom unions and 33 free trade areas). In both cases we deal with the mixture of free trade and protectionism (Yi, 1996, p. 154; Jantoń-Drozdowka, 2004, p.171, 184—186; Magee and Lee, 2001, p. 496—499).

Let's have a look now into the mechanism of creating the statistic effects of customs union such as the trade creation and trade diversion effect and the dynamic effects. The latter include the intensification of competition, modernising the factors of production as a result of introducing the technological progress, the increase of the efficiency of the business activity.

In the theory of international economic integration, creating customs union can lead to both positive and negative effects. Positive effects are identified as the effect of trade creation and improving the allocation of resources, and negative ones are connected with introducing common external duty tariff, which contributes to trade diversion. The traditional attitude towards the effects of integration says that trade integration in the form of customs union can worsen the allocation of resources within the union, which can be transferred to decline in efficiency within the whole group. Such situation appears when the import from a much more efficient country that is not a member of the group, becomes replaced by trading with a less efficient partner (country) that belongs to the customs union. Such point of view over customs union, analysing not only its positive effects was firstly introduced and expanded by J. Viner (1950).

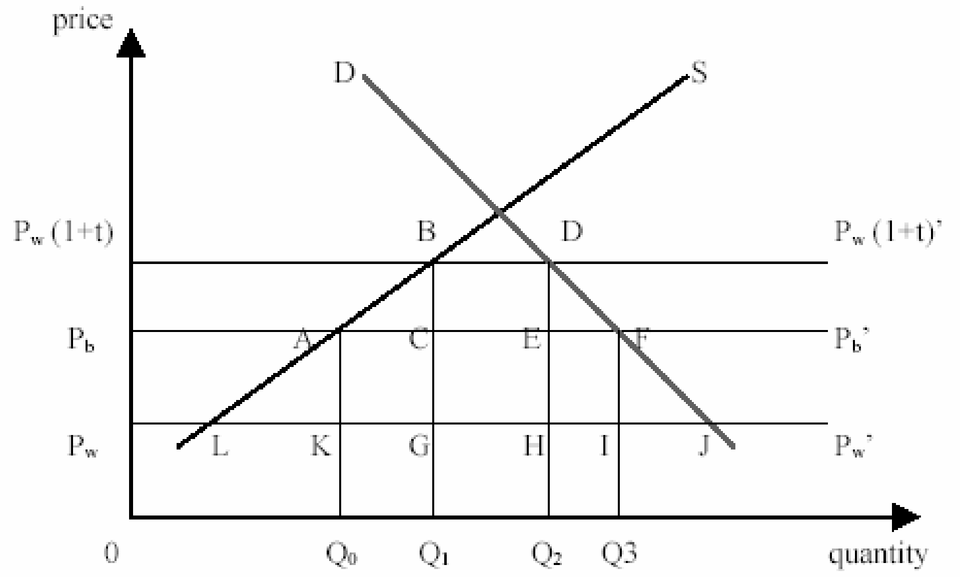
J. Viner examined the case of three countries where the producer with the highest price of the item X (the least efficient) – country A, has a chance to create customs union with another producer with the highest price of the item X, but more efficient than country A – country B or with the most efficient producer – country C, known as the rest of the world. If customs union is created by countries A and B that agree on common external tariff towards country C there will be the diversion effect, that is replacing lover-cost importing activities from other countries with relatively more expensive products from the customs union area. Country A will import goods from country B for higher price comparing to the price from before creating the union, assuming that previously there were no other duty tariff towards country C as high as after the union's formation. Within the group then, there will appear the effect of trade creation i.e. there will be so far non-existing trade connections created.

Before joining the union country A covered some of its demand inefficiently producing item X and applying duty protection of it. When country A eliminates duty barriers towards country B (a more efficient producer of item X) and all required market readjustments appear, inefficient industry producing item X in country A will partially be destroyed by the competition together with the increase of importing from B to A. Therefore the trade creation will follow. Country B will get some extra market and the amount of item X produced by country A will decrease. The rate of decline of item X production depends on the level of the industry's strength from before creating the union was conditioned on duty protection**.** In case of both effects the country with highest production cost (least efficient producer) will experience the increase of importing from the partner's country and the decrease of its own production. We shouldn't forget here about the initial guidelines concerning the countries' efficiency assumed by J. Viner (Caves, Frankel and Jones, 1998, p. 349-350; Bohara, Gawande and

Sanguinetti, 2004, p. 65-66, 69; Jantoń-Drozdowska, 2004, p. 171, 177-179; Nicholls, 1998, p. 324-325).

Figure 1 illustrates the above static view over the integration effects according to J. Viner, which is based on the model of a partial equilibrium for a perfect competition market, assuming a linear form of demand and supply curves. In the picture curve D represents the demand curve for item X, and curve S represents the supply of the very product within the customs union. When the price is $0P_w$ (1+t), i.e. the sum of free trade price (P_w), and external tariff designated on importing item X, $0Q_1$ units of item is produced while $0Q_2$ is the demand for this item reported by the consumers. The difference between $0Q_1$ and $0Q_2$ is imported from country C. Having removed the tariffs between the countries A and B, due to creating the customs union, $0Q_0$ units of item X is produced, and up to $0Q_3$ increases the demand reported by the customs union with the price $0P_b$, which is the price of importing the very item from the partner country.

Figure 1. The creation and diversion effect — partial equilibrium diagram of custom union formation.



Source: Nicholls, 1998, p. 325.

The area DEF represents the consumer's gains (consumer surplus), and ABC represents the producer's gains (producer surplus). Except benefits, creating customs union causes also the loss in production - area ALK and in consumption - area FIJ. These are the consumption or production triangles based on A. Marshall's point of view. Moreover, there will be a decrease in profits from economic tariffs of country A due to the decline of importing from the rest of the world. In the partial equilibrium model the benefits coming from the trade creation effect are the area of the triangle ABC and DEF, and the loss caused by the trade division – area CEHG. The area CEHG, which is a part of country A consumer's expenditure on item X produced at a higher price in country B than in country C, is a pure social loss. The sum of these two effects gives the net welfare effect resulting from the union formation. First, however, the areas ABC, DEF and CEHG should be measured using the methods of classical geometry. The procedure of measuring can be simplified using the method of mapping[1] (Nicholls, 1998, p. 324—325; Caves, Frankel and Jones, 1998, p. 349—352; Bohara, Gawande and Sanguinetti, 2004, p. 70—71).

On the basis of the critics of J. Viner's view there were many other concepts created. They took into consideration other aspects than price changes, as variables influencing the allocation of the resources, different sizes of the countries, the forms of the market and the characteristics of the demand and the supply curve, etc. Most of these were indeed the widening or the modification of the above view, in order to adjust this view to the changes of the trade reality. This paper, however, does not include those views.

The works of many authors and their research show that creating a customs union betters the whole group's welfare. The chances of substitution and transformation of the goods are much better due to abolishing internal trade barriers, which decreases other distortions of the market. Nowadays the advantages of the trade integration are much more mentioned in the subject's literature than they used to be, and the disadvantages are not emphasised. Some of those papers indicate as well that the customs union is a better choice than free trade zone.

Within the customs union and later within the common market which very often is a natural effect of the customs union's development, the movement towards the liberalisation of the trade exchange and the trade integration evokes. That forms a more competitive market where the country's consumers and producers have a better access to foreign goods and knowledge. The flow of information about the goods markets and prices increases. In other words, the integration within the customs union generates a bigger and more competitive market where the companies have access to a wider scope of the flow of the knowledge and the technologies that support a faster economic growth. Bigger sizes of the market stimulate new infrastructure investments favourable to the general economic growth and rising the efficiency of the whole economy as increased investment demand stimulate the economic growth via the multiplier effects**.** The bigger effect of growth the less competitive economy from before the integration, while it has little importance for highly developed economies.

The growth of market competition due to the elimination of the trade barriers causes that the country's producers are protected to a lesser extent from the other

[1] The ways of measuring using the method of mapping are described thoroughly for instance in: Greenaway, 1983; Marques-Mendes, 1987; Nicholls, 1998.

countries' competition. This may result in a decline of the global number of firms and the number of the variety of goods produced in the country. Such situation, from the consumer's point of view, is compensated via the introduction of other producers from partner countries to the domestic market and the drop of prices, as the competition forces others to reduce profit margin and to increase the financial profitability together with the companies' effectiveness. It causes the growth of the consumption and gradual stabilising of the prices within the union. The scale of this phenomenon depends on the level of elasticity of the demand in the partner countries, though. The producers then have the access to the larger number of consumers and larger scope of the knowledge and technology flow and the consumers to the wider variety of less expensive goods.

The companies that survive may achieve bigger sizes via both interior and exterior (e.g. mergers, accessions, strategic alliances) growth, which increases their possibilities to use the statistic and dynamic integration effects, including the economies of scale and specialisation. Since the companies on a larger, more competitive market invest more in their development which can lead to the increase in the expenditure on the research and development activity, and the improvement of the quality of the production factors caused by the technological development. It is seen in the increase of the productivity and in the improvement of the resources' allocation. There appears the rationalisation effect that leads to the growth of the trade group's welfare. It also contributes to the changes within the national markets' structures and within the industry structure of the customs union market that is connected with the growth of the specialisation of the business activity based on the comparative or competitive advantages. Gradual elimination of tariffs also leads to a larger economic concentration, which can increase the level of controlling the market by the bigger firms or their groups (Panagariya and Krishna, 2002, p. 354-358; Peretto, 2003, p. 177-188, 198-200; Jantoń-Drozdowska, 2004, p. 172-184).

Within European Union, as research conveyed by various authors confirm, the creation effect always surpasses the diversion effect.

After the EEC was created, the intraregion tariffs reduction caused the creation of additional trade streams, which resulted in the decline of national producers' participation in the consumption of the processed goods. M. E. Kreinin (1974) estimated that in the years 1969 and 1970 the existence of EEC caused trade diversion worth 1.1 billion USD, at simultaneous creation of the trade creation worth 8.4 billion USD. The latest research concerning the last decade of XX century show that the diversion effect within EU at that time did not exist, which results to a large degree from the group's character and EU specific trade policy towards other countries. Since the diversion effect may be eliminated via reducing the external tariffs, which actually took place within EU. It is so as within the customs union the profits stemming from its formation start to be present. These are the increase of productivity and welfare of the partner countries. Then the pressure for maintaining a higher exterior duty tariff, especially for the factors of production declines. Among the union members appears also a bigger industry specialisation, which correlates with the growth of the exporting of the industries that have comparative advantages.

The research concerning the trade creation effect also shows that the trade creation in the EU took place according to the comparative advantage theory. The country that before the reduction of internal tariff was the most efficient producer of some item among other partner countries usually became a supplier of the major part of the importing the very product to other countries of the Union.

The directions of intraregion trade exchange within EU reflect the competitiveness of the union companies and the changes in the comparative advantage. Moreover, the EU research confirm the appearance of the dynamic gains resulting from the creation of the customs union, so mainly the growth of the market's competitiveness. The experiences of other countries that became the members of the group show that it is likely that in the future, after Poland has joined EU, there will be a regional and local adjustment of the industry structure. Its results may be the weakening of some lines and the development of some others, which may increase the degree of the specialisation of Poland's trade exchange within the Union. Therefore the dynamic and static trade integration effects appear (Caves, Frankel and Jones, 1998, p. 356; Endoch, 1999, p. 215; Kennan and Riezman, 1990, p. 70-83; European Commission, 2002, p. 10; An and Iyigun, 2004, p. 465-483; Ulosoy, 2001, p. 133-145; Berenton, 2001, p. 599; Richardson, 1993, p. 309-311; Yi, 1996, p. 154).

3. The gravity model

One of the methods of quantification and analysing the economic integration effects in the field of the international trade size and directions is the gravity model based on Newton's principle whose main and highly valued advantage is simplicity. Therefore it is used in many fields of social science. In the subject's literature diverse formulas of the gravity equation can be found. There is no one form accepted of estimating the integration effects caused by the common growth of trade exchange between the countries by means of the gravity model. The gravity model assumes, according to the Newton's principle, that the flow of people, capital and goods between two countries (localisations) are positively correlated to the size of the economies of these countries and negatively correlated to the distance between them. In other words, bilateral trade should increase together with the increase of the Gross National Product (GNP) (measures of the economy sizes in the very model) and should decrease together with the increase of the physical distance between the countries.

J. Tinbergen (1962) and P. Pöyhönen (1963) used this model assuming the formula of the gravity equation:

$$trade_{ij} = A\frac{\left(GNP_{i}GNP_{j}\right)^{b_1}}{\left(\text{distance}_{ij}\right)^{b_2}}$$

or having used the natural logarithms

$$\ln(trade_{ij}) = A + b_1 \ln(GNP_i * GNP_j) + b_2 \ln(\text{distance}_{ij}) + \varepsilon_{ij},$$

where A, b_1 and b_2 are coefficients that are to be estimated, $trade_{ij}$ is the value of bilateral trade between the countries i and j, GNP_i i GNP_j are the measures of economy sizes i and j, $distance_{ij}$ is the measure of the distance between the two countries (usually between the capitals). The lognormally distributed error ε_{ij} includes other phenomena that might influence the bilateral trade between two countries that were not mentioned in the first part of the equation.

The primary form of the gravity equation described in this work was tested by the researchers in various conditions and for the countries of a different level of economic development (see, for example: Hewett, 1976; Brada and Méndez, 1985; Helpman and. Krugman, 1985; Bergstrand, 1989; Deardorff, 1998). They affirmed that it is consistent with the trade theories based on the imperfect competition models and with the Hecksher-Ohlin model.

Then, the prime form of the gravity model equation was modified and developed by individual researchers in different ways depending on their research. They transformed its formula and added variables such as the price and foreign exchange rate level for instance (nowadays these variables are preferably substituted by the countries' membership in the monetary union), the size of the population, GNP *per capita*, common language and the border of the countries studied, the measures of institutional adjustments of all kinds. What is more, together with the growth of the tendencies towards regionalisation of the international trade exchange via creating varied trade groups in the gravity model, the variables allowing measuring the trade creation and diversion effect started to be taken into consideration. The research that used the very model may be carried out in the region, country, industry line or individual items.

What joins all the models studying the effects of trade integration using the gravity model is that most of their authors based on the basic formula of gravity equation :

$$\ln trade_{ij} = \ln a_0 + a_1 \ln Yi + a_2 \ln Y_j + a_3 \ln X_i + a_4 \ln X_j + a_5 \ln D_{ij} + a_6 \ln CDE_{ij} + \ln e_{ij},$$

where $trade_{ij}$ is the flow of goods from country i to country j; Yi i Yj are the sizes of the country economics measured by e.g. GNP or by the size of the population of that country or by GNP *per capita*; X_i and Xj are the additional variables (e.g. common language or the level of the prices), Dij is the measure of the distance between country i and j; CDE_{ij} are the variables of the measures of the trade creation or diversion effect; ε_{ij} is the normally distributed error with zero expectation. Variables Y and X in the model may be taken into consideration for both countries simultaneously or separately. The form choice depends on the assumed research concept .

It is worth to emphasise the fact that diversity of the variables taken into consideration in these models for different trade groups and the groups of countries and the distinctness of the results concerning the significance of the studied variables , makes, that it is not possible to state unambiguously which of those additional variables should be taken into consideration in the formula of the gravity equation to measure the integration effects (Endoch, 1999, p. 207-209; Ghosh and Yamaric, 2004, p. 370-371).

If the above formula of the gravity model is the basis in order to measure the creation and diversion effect the following equation has been used:

$$\ln IT_{ij} = \ln a_0 + a_1 \ln(Yi * Y_j) + a_2 \ln(POP_i * POP_j) + a_3 \ln D_{ij} + a_4 \ln B_{ij} + a_5 \ln DEIMP_{ij} +$$

$$+ a_6 \ln DEEXP_{ij} + a_7 \ln CEIMP_{ij} + a_8 \ln CEEXP_{ij} + a_9 \ln CEIT_{ij} + \ln e_{ij},$$

where:

IT_{ij} is the bilateral trade value (imports plus exports) of Poland (country *i*) in USD with country *j;*

ln a_0 is a constant in the model;

Y_i and Y_j are nominal values GNP *per capita* of the country *i* (Poland) and *j* in USD;

POP_i *and* POP_j the size of the population of the country *i* and *j*;

lnD_{ij} is the measure of the distance between the capital of Poland and the capital of country *j*;

lnB_{ij} is the dummy variable equals to 1, when Poland borders on country *j*, and value 0 while the opposite;

$DEIMP_{ij}$ is the value of Poland's imports in USD from country *j*, that is not a member or did not join the EU in 2004;

$DEEXP_{ij}$ is the value of Poland's exports in USD to country *j*, that is not a member or did not join the EU in 2004;

$CEIMP_{ij}$ is the value of Poland's imports in USD from country *j*, that is a member or joined the EU in 2004;

$CEEXP_{ij}$ is the value of Poland's exports in USD to country *j*, that is a member or joined the EU in 2004;

$CEIT_{ij}$ is the value of Poland's bilateral trade in USD with country *j*, that is a member or joined the EU in 2004;

$ln\varepsilon_{ij}$ is the lognormally distributed error.

Variables $DEIMP_{ij}$ and $DEEXP_{ij}$ reflect the diversion effect appearing accordingly in Poland's imports and exports. If the coefficients of the two variables are negative and statistically significant we can affirm that Poland's trade exchange with the outer-region countries is smaller and in favour of the EU member countries. The depiction of the diversion effect in importing activities appeals to the B. Balassa (1967, p. 5) or H. G. Johnson (1962, p. 53) definition and differs from the definition of J. Viner (1950, p. 43). However, M. Endoch (1999, p. 210) first introduced the depiction of the diversion effect in exporting.

$CEIMP_{ij}$, $CEEXP_{ij}$ and $CEIT_{ij}$ are the variables that representing the creation effect occurring in case of Poland in importing, exporting and international trade on the whole, that results from the economic integration within the EU. The authors decided to divide the trade creation effect into importing and exporting trade creation effect, similarly to M. Endoch's actions in case of the diversion effect. In the situation when a coefficient is positive and statistically significant we can affirm that Poland's trade exchange with the countries of the EU is bigger than the hypothetical level, so the trade creation effect occurs.

The above depiction of the creation and diversion effect makes it possible to differentiate these two effects in a simple way. Thanks to that there is no need to use a more complicated approach that had been used more often to study these phenomena in other works which measures the net creation effect obtained out of the subtraction from the gross creation effect the diversion effect (see, for example: Frankel, 1993; Frankel and Wei, 1995, Endoch, 1999). It was also chosen because of the research aim which is not an accurate measuring how big the effects are for Poland, but checking whether they actually occur.

4. The results

Due to the fact that European Union treated the development of the trade exchange with the candidate countries during the pre-access time as a way to approach each other and increasing the common economic relationships between its present and future members actually creating the free trade zone between the candidate countries and EU before they access on the 1st of May, 2004, it was checked whether the trade creation and diversion effect appeared in Poland in the pre access time. For the reasons the research population was divided into three groups. However, it was the percentage of the countries' participation in Poland's foreign trade at all that decided of the choice of the countries of the rest of the world:

- the countries of so called old European Union (EU15): Austria, Belgium, Denmark, Finland, France, Greece, Spain, Holland, Ireland, Luxembourg, Germany, Portugal, Sweden, Great Britain, Italy;
- the countries of so called new European Union (EU25): EU15 plus Cyprus, The Czech Republic, Estonia, Lithuania, Latvia, Malta, Slovakia, Slovenia, Hungary;
- the rest of the world (RW): Argentina, Australia, Belarus, Brazil, Bulgaria, China, India, Japan, Canada, Mexico, Norway, Southern Korea, Russia, Romania, the United States of America, Switzerland, Turkey, Ukraine.

The equation of the gravity model has been estimated for bilateral trade flows between Poland and 42 other countries divided into three groups, in the years 1990, 1995, 1997, 1999, 2001 and 2003. While choosing the years to be studied the stages of Poland's trade integration with EU were taken into consideration, within the confines of the Transitional Agreement signed by Poland and which was put into effect on the 1st of March, 1992. It was a part of European Treaty signed in December 1991. According to this agreement during the transitional time there was to be a free trade area for industrial items gradually created between Poland and EU, but the Union was to finish off the elimination of trade barriers for industry items after six years (till the end of 1996), and Poland after ten years since the trade part of European Treaty has been put into effect. As far as the agricultural products and food are concerned a partial reduction of duty and compensatory charges and gradual elimination of quantitative limitations were provided.[1]

The year 1997 gives to Polish producers of industrial products the access to the EU market on the free trade principles. That is why beginning with this year the creation and diversion effects for the countries of the new EU started to be measured. The significance of the year 1997 for Poland's foreign trade with EU confirm also other research of the authors concerning the influence of the integration with the Union into the changes in the structure of Poland's foreign country, especially in exporting. The researches were conveyed with the method of the pattern of the economic growth and they show that from 1997 on the stronger convergence of the structure of Poland and EU foreign trade can be observed (Majewska and Grala-Michalak, 2005, p. 127—137).

The subject of the research were the data concerning: Poland's trade exchange from the Foreign Trade Statistical Yearbooks of Central Statistical Office (CSO); GNP *per capita* announced by UNCTAD and the size of the population from Poland's Statistical Yearbook of CSO as a source. Moreover, the program for calculating the distance between the capital cities was used.

Table 1 shows the results of the research obtained from estimating the gravity equation using the method of stepwise progressive regression. In the stepwise regression procedure some of the considered variables may not be taken into consideration in the model which has then a simpler form. The variables are added to the model if the value of the F statistic exceeds 1 for them, in no more than 13 steps. Using the *t-Student* test for the regression coefficients we can state whether they are significant. In the table the regression coefficients statistically significant on the significance level α = 0.05 have been written in bold. The standard error of estimation for the regression coefficient stands in the table in the brackets under the coefficient.

[1] The description of the stages of Poland's integration with EU and the schedule of liberalisation of the access of Polish industry goods to the EU market can be found for instance in the work by Jantoń-Drozdowska, 1998, p. 41-56.

Table 1. Part one: The results of the regression analysis

Year	1990	1995	1997	
	EU15	EU15	EU15	EU25
Constant	0.444 (0.230)	**0.754** (0.346)	**1.366** (0.550)	-4.078 (2.627)
Regression coefficients				
(Y_i*Y_j)	**0.029** (0.011)	-	-	-
(POP_i*POP_j)	-	-	0.004 (0.035)	**0.793** (0.124)
B_{ij}	0.046 (0.046)	-0.148 (0.118)	-	-
D_{ij}	-	-	-	-
$DEIMP_{ij}$	**0.314** (0.018)	**0.448** (0.036)	**0.938** (0.036)	-
$DEEXP_{ij}$	**0.667** (0.024)	**0.563** (0.049)	-	-
$CEIMP_{ij}$	**0.171** (0.036)	**0.542** (0.143)	-	**0.170** (0.035)
$CEEXP_{ij}$	-	**0.455** (0.145)	**0.995** (0.033)	-
$CEIT_{ij}$	**0.772** (0.041)	-	-	-
R^2	0.999	0.984	0.975	0.548
Standard Error of Estimation	0.063	0.221	0.315	1.328
F	3895.838	443.145	499.798	23.680

Source: own study.

In all these models, except for one only, the determination coefficients R^2 reached the value of over 97%, which means that the regression equation at least in 97% explains the changes of the endogenous variable by the changes of the exogenous variables. It just goes to show that these models were perfectly adjusted to the data.

The values of the obtained regression coefficients make it possible to estimate the values of the trade creation and diversion for Poland at the given year, and the changes of those indicators in time show, if the given effect has weakened or strengthened. The positive value of this indicator for the trade creation effect lets us conclude that Poland's trade at the given year with EU member countries was higher from the theoretical level. There occurs then the trade creation effect. If the value of this indicator for the trade creation effect falls when it is positive and statistically significant means weakening of that effect and approaching the theoretical (hypothetical) level by the trade value. Similarly we interpret the regression effects obtained for the diversion effect, bearing in mind its sign.

Table 1. Part two: The results of the regression analysis

Year	1999		2001		2003	
	EU15	EU25	EU15	EU25	EU15	EU25
Constant	**1.588** (0.394)	0.705 (0.478)	**0.886** (0.234)	**1.368** (0.487)	**0.914** (0.238)	0.769 (0.424)
Regression coefficients						
$(Y_i{*}Y_j)$	-	-0.010 (0.026)	-	-0.005 (0.030)	-	-0.018 (0.023)
$(POP_i{*}POP_j)$	-	-0.005 0.027	-	0.012 (0.032)	-	-0.008 (0.025)
B_{ij}	-	0.020 (0.052)	-	-0.004 (0.060)	-	-0.016 (0.045)
D_{ij}	-	-0.007 (0.043)	-	-0.055 (0.047)	-	0.004 (0.041)
$DEIMP_{ij}$	**0.504** (0.032)	**0.758** (0.027)	**0.618** (0.026)	**0.705** (0.032)	**0.590** (0.027)	**0.721** (0.025)
$DEEXP_{ij}$	**0.438** (0.046)	**0.266** (0.031)	**0.374** (0.028)	**0.265** (0.031)	**0.402** (0.028)	**0.310** (0.027)
$CEIMP_{ij}$	**0.495** (0.150)	0.071 (0.053)	**0.572** (0.086)	**0.450** (0.037)	**0.554** (0.092)	0.135 (0.077)
$CEEXP_{ij}$	**0.442** (0.147)	0.071 (0.089)	**0.415** (0.086)	**0.521** (0.037)	**0.431** (0.090)	0.163 (0.098)
$CEIT_{ij}$	-	**0.838** (0.132)	-	-	-	**0.696** (0.168)
R^2	0.976	0.997	0.991	0.996	0.990	0.998
Standard Error of Estimation	0.236	0.090	0.144	0.098	0.152	0.079
F	382.622	1206.395	969.610	1040.990	921.313	1524.624

Source: own study.

We should not forget that it may be misleading to compare the size of the creation and diversion effect estimated for individual years in order to check whether its strength grows or weakens as the size of these effects depends on the real value of the trade at the given year. There may be a situation that in spite of the decline of the trade creation effect coefficient's value, the estimated size of the effect in the nominal quantity will increase. Therefore we will limit ourselves to the interpretation of the changes of the obtained regression coefficients' value.

Poland's bilateral trade with the countries of EU15 and EU25 (for 24 countries all together) during the pre access time was characterised by a lack of the trade diversion effect towards the member countries both for importing and exporting. The positive sign of the regression coefficients for the variables $DEIMP_{ij}$ and $DEEXP_{ij}$ means that Poland did not lessen the trade exchange with the rest of the world in a statistically significant way during the analysed period. To the contrary, Poland's bilateral trade with the outer-region countries was on a higher level than it would result from the hypothetical trade level. The values of these coefficients were changing in such a way that it is difficult to state whether the very effect had growing or falling tendencies. We can only state that the positive value of the exporting diversion effect for so called old and new EU after 1997 was lower than the importing diversion effect.

The value of the regression coefficients for the variables representing the creation effect is positive which just goes to show that the effect occurs for Poland in the analysed period. As to EU15 the exporting creation effect was not statistically significant on 5% level only in 1990 and in importing in 1997 when the strongest exporting creation effect in the analysed time occurred for Poland. Analysing the changes of these coefficients' values we can conclude that the strength of the very effects for EU15 after 1997 has been constant in some interval. Comparing the value of these indicators for old and new Union we can observe that Poland's exporting and importing creation effect during 1999—2003 was stronger for the old Union than for EU25.

5. Conclusions

To conclude it is characteristic for Poland that in the studied time the trade creation effect appeared and the trade diversion effect did not. The results show that Poland's exports and imports with other countries did not suffer due to Poland's integration with EU. Using B. Balassa terminology we can affirm on the basis of the research that in case of Poland a complete trade creation appeared which includes the effects of the appearance of new trade streams between the EU member countries and the exterior trade creation.

The exporting creation effect was the strongest in 1997 when EU accomplished the creation of the free trade zone as far as Poland's industrial goods are concerned. The creation effect occurred to a larger extent in Poland's trade exchange with the countries of so called old EU because, among other things, these countries participate much more in the production and world trade than the new members of EU. There are many world wide known producers in their areas, which increases the possibility that a stronger trade creation occurs. On the basis of the research results we can affirm that in case of the trade integration the positive static effects of this integration dominate which is supported by the results of some other research conveyed by the authors of this work.

REFERENCES

AN, G. IYIGUN M. F. (2004), The export technology content, learning by doing and specialization in foreign trade, Journal of International Economics, 64, p. 465– 83.

BALASSA B. (1967), Trade creation and trade diversion in the European Common Market, Economic Journal, 77, p. 1—21.

BERENTON P. (2001), Anti-dumping policies in the EU and trade diversion, European Journal of Political Economy, 17, p. 593—607.

BERGSTRAND J.H. (1989), The generalized gravity equation, monopolistic competition, and the factor-proportions theory in international trade, Review of Economics and Statistics, 71, p. 549—56.

BOHARA A. K., GAWANDE K., SANGUINETTI P. (2004), Trade diversion and declining tariffs: evidence from Mercosur, Journal of International Economics, 64, p. 65—88.

BRADA J. C., MÉNDEZ J. A. (1985), Economic integration among developed, developing and centrally planned economies: a comparative analysis, Review of Economics and Statistics, 67, p. 549—56.

CAVES R. E., FRANKEL J. A., JONES R. W. (1998), World Trade and Payments (in Polish), Warsaw, PWE.

DEARDORFF A. V. (1998), Determinants of bilateral trade: does gravity work in a classical world, [in:] Frankel J. (ed.), The Regionalization of the World Economy, Chicago, University of Chicago, p. 7—22.

ENDOCH M. (1999), Trade creation and trade diversion in the EEC, the LAFTA and the CMEA: 1960-1994, Applied Economics, 1999, 31, p. 207—16.

EUROPEAN COMMISSION (December 2002), Making globalization work for everyone, The European Union and world trade, Manuscript for information brochure.

FRANKEL J. A. (1993), Is Japan creating yen block in East Asia and the Pacific?, [in:] Frankel J. A., Kahler M. (ed.), Regionalism and Rivalry — Japan and the United States in Pacific Asia, Chicago, The University of Chicago Press, p. 53-85.

FRANKEL J. A., Wei S. (1995), Emerging currency blocks, [in:] Genberg H. (ed.), The International Monetary System — its Institutions and its Future, Berlin, Springer, p. 111—170.

GHOSH S., YAMARIC S. (2004), Are regional trading arrangements trade creating? An application of extreme bound analysis, Journal of International Economics, 63, p. 369—95.

GREENAWAY D. (1983), Trade Policy and the New Protectionism, New York, St. Martin's Press.

HELPMAN E., KRUGMAN P. (1985), Market Structure and Foreign Trade, Cambridge, MIT Press.

HEWETT E. A. (1976), A gravity model of CMEA, [in:] Brada, J. C. (ed.), Quantitative and Analytical Studies in East-West Economic Relations, Bloomington, International Development Research Center, p. 1—15.

JANTOŃ-DROZDOWSKA E. (1998), Regional Economic Integration (in Polish), Warsaw-Poznan, PWN.

JANTOŃ-DROZDOWSKA E. (2004), International Business Economics (in Polish), Poznan, Ars boni et aequi.

JOHNSON H. G. (1962), Money, Trade and Economic Growth- Survey Lectures in Economic Theory, London, George Allen and Unwin.

KENNAN J., RIEZMAN R. (1990), Optimal tariff equilibria with custom unions, Canadian Journal of Economics, 23, p. 70—83.

MAGEE S. P., LEE H.-L. (2001), Endogenous tariff creation and tariff diversion in a customs union, European Economic Review, 45, p. 495—518.

MAJEWSKA M., GRALA-MICHALAK J. (2005), Convergence between Polish and European Union foreign trade structures and its implications for economic development of Poland, [in:] Dudka M. (ed.), European Agricultural Policy and Cohesion Policy, Selected Problems, Zielona Góra, University of Zielona Góra Printing house, p. 127—137.

MARQUES-MENDES A. (1987), Economic Integration and Growth in Europe, London, Croomheim.

NICHOLLS S. M. A. (1998), Measuring Trade Creation and Trade Diversion in the Central American Common Market: A Hicksian Alternative, World Development, 26, p. 323—35.

PANAGARIYA A., KRISHNA P. (2002), On necessarily welfare-enhancing free trade areas, Journal of International Economics, 57, p. 353—67.

PERETTO P. F. (2003), Endogenous market structure and the growth and welfare effects of economic integration, Journal of International Economics, 60, p. 177—201.

PÖYHÖNEN P. (1963), A tentative model for the volume of trade between countries, Weltwirtschaftliches Archiv, 90, p. 93—99.

RICHARDSON M. (1993), Endogenous protection and trade diversion, Journal of International Economics, 34, p. 309– 11.

TINBERGEN J. (1962), Shaping the World Economy, New York, The Twentieth Century Fund.

ULOSOY V. (2001), Trade and Convergence: A Dynamic Panel Data Approach, Southern Economic Journal, 68, p. 133—44.

VINER J. (1950), The Customs Union Issue, New York, Carnegie Endowment for International Peace.

YI S.-S. (1996), Endogenous formation of customs union under imperfect competition: open regionalism is good, Journal of International Economics, 41, p. 153—77.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 191—209

THE QUALITY POLICY IN THE MANAGEMENT OF THE POLISH LABOUR COSTS SURVEY

Jolanta Szutkowska[1]

ABSTRACT

This article presents the quality policy in the labour costs survey which is European survey conducted every 4-years strictly according to Eurostat recommendations. This survey is well advanced in the quality policy because earnings and labour costs statistics was one of the first domains covered by quality requirements of European Statistical System. The quality policy in labour costs survey is illustrated on the basis of different quality approaches, namely: (i) the labour costs survey as the coherent and comparable part of the integrated system on earnings and labour costs statistics, (ii) quality reporting in labour costs survey, (iii) the application of DESAP as the self-assessment quality tool for the survey manager, (iv) the quality audit in labour costs survey.

***Key words*:** quality dimensions, quality audit the self assessment checklist,

1. Introduction

The quality policy in the management of statistical survey may be defined as the way of thinking which is based on following fundaments:

- the state of mind which creates consciousness and initiates efforts of statistical staff oriented on quality improvements;
- the framework of rules aimed at constantly fulfilment of quality improvements;
- the implementation of tools for the aim of quality policy such as quality reports, DESAP — the form for the self-assessment of quality;
- the identification and monitoring of quality aspects of statistical process by quality auditors (identification of weaknesses, strengths of which the best practices and improvement ideas).

[1] Central Statistical Office , Al. Niepodleglosci 208, 00-925 Warsaw, Poland,
Email: J.Szutkowska@stat.gov.pl

The approach to the quality policy in labour costs survey was created by the activities of Eurostat within the Working Group on wages and labour costs statistics (at the present this group is the part of Eurostat Working Group LAMAS on Labour Market) as well as by actions taken by the Working Group on Quality Assessments in Statistics when adequate rules such as the Leadership Group on Quality (LEG) recommendations and quality tools were created. The Polish section on labour costs and earnings statistics gradually implements all rules regarding quality definitions, quality reporting and quality evaluation defined by topic European Commission regulations on data quality in earnings and labour costs statistics as well as by LEG recommendations. Experiences of Statistics Sweden in quality policy including the implementation of TQM rules are also taken into account.

Below, the implementation of selected elements of quality policy in the Polish labour costs survey is presented in details.

2. The labour costs survey as the coherent and comparable part of the integrated system on earnings and labour costs statistics

One of the elements of quality policy is the creation of the integrated system on earnings and labour costs statistics.

According to Eurostat recommendations the integrated system is created to obtain a coherent set of earnings and labour costs statistics in terms of concepts and definitions on short-term, annual and four yearly basis. Hence, the important element of the integrated system is the improvement of the overall coverage of the economic activities, improving clarity and completeness, harmonisation of definitions and coverage especially in the field of the two structural surveys on structure of earnings by occupations and labour costs as well as the implementation of the coherence with the other business statistics and national accounts.

Earnings and labour costs statistics conducted by enterprises in Poland is based on full surveys and sample surveys.

The labour costs survey plays the important role in the evaluation of situation on labour market. It gives the picture on basic indicators on labour costs by different breakdowns what have the significance for monitoring labour market policy within the Revised Lisbon strategy and the European employment strategy. Labour costs are defined as the costs borne by the employer on acquiring, using, maintaining and upgrading labour resources.

The place of labour costs survey in the integrated system on earnings and labour costs is presented on the graph.

Labour costs survey is conducted every 4-year on the sample basis (20% of the random frame) in units of whole national economy (type of activity NACE: A-O) employing 10 and more employed persons. The GUS makes efforts to enlarge coverage of units also by small units employing up to 9 employed persons according to Eurostat recommendations.

Breakdown of data includes: type of activity by NACE, size of units, ownership sectors, NUTS.

Objective scope refers to number of employed persons, average employment, hours of work by full-time employees, part-time-employees and apprentices, components of labour costs (wage components, non-wage components), selected components of labour costs for apprentices.

The basic measures of labour costs are following:

- labour cost per 1 employee calculated as monthly average,
- labour cost per 1 hour paid for,
- labour cost per 1 hour worked.

In addition, the information on structure of labour costs components, structure of earnings and structure of working time are available.

The sampling scheme and grossing-up procedures applied in the labour costs survey for 2004 is described in details in Polish—English publication „Labour costs in the national economy in 2004"

Response rate in labour costs survey is about 80%.

As for the dissemination policy, data are presented in the publication on labour costs 11 months after the reference period. Data are transmitted timely to Eurostat according to required technical format and up to the deadline given in the EC regulation e.g. data for LCS 2000 were transmitted on the basis EC regulation No 1726/1999; data for LCS 2004 are transmitted on the basis the EC regulation No 1737/2005. Quality rules are fulfilled according to EC regulation No 452/2000 and its updated version No 698/2006.

As the important supplement to labour costs statistics are annual estimations on labour costs by type of activity and ownership sectors (average monthly labour costs per 1 employee; hourly labour costs) and estimations on quarterly labour costs index. Annual estimations on labour costs are based mainly on annual employment, earnings and working hours statistics conducted on the form Z—06. Estimations on quarterly labour costs index are prepared on the basis on monthly, quarterly and annual earnings statistics as well as the last labour costs survey. The quarterly labour costs index is compiled in the form of labour costs index total, wage costs index and other costs index by sections C_K and transmitted to Eurostat up to 70 days after the reference quarter. The work on the enlarged LCI by sections L, M, N, O and by labour costs index excluding bonuses as well as the preparation of seasonal adjusted time series is developed. Quality rules for the LCI are fulfilled according to EC regulations No 450/2003 and No 1216/2003. As we see the links between labour costs survey with other type of earnings statistics conducted by enterprises is strong. Hence, it is necessary to present in brief the essential characteristics of these statistics.

Short-term statistics in Poland is conducted *on the monthly basis* in units on own accounts of the enterprise sector employing 10 and more employed persons and conducting their activities mainly in sections C—K, O. Sections NACE such as J, L, M, N are not included.

This survey is a complete enumeration for big units employing 50 and more employed persons and a sample survey for medium units employing from 10 to 49 employed persons- sample amounts to more than 10% of the frame.

Breakdown of data includes: type of activity by NACE, size of units, ownership sectors, NUTS.

Objective scope within the framework of labour statistics refers to number of employed persons, average employment, hours worked, earnings with specifications following earnings components: bonuses from profit, income taxes, social security contributions paid by employees.

Response rate is about 95%.

As for the dissemination policy, the first data on employment and earnings are available 2 weeks (10 working days) after the reference period in the form of announcement of President of the GUS. About 3 weeks after the reference period the press conference is organized where data from earnings statistics are presented. Monthly publications (such as Statistical Bulletin, Information on socio-economic situation in the country) are available 3 weeks after the reference period. Monthly earnings statistics is the source on employment and earnings data transmitted to Eurostat in GESMES on the basis of EC regulation No 1165/98 and its updated version No 1158/2005.

Quarterly earnings statistics is based on a complete enumeration conducted in units of the whole economy (i.e. units for sections J, L, M, N and budgetary entities from the rest of the national economy) employing 10 and more employed persons. Breakdown of data includes: type of activity by NACE, size of units, ownership sectors, NUTS.

Objective scope refers to number of employed persons, average employment, hours of work, earnings with specifications following earnings components: bonuses from profit, income taxes, social security contributions paid by employees.

Response rate is about 95%.

As for the dissemination policy, data on quarterly basis are available in quarterly publication on employment and earnings 3 months after the reference quarter.

Annual earnings statistics is based on a complete enumeration conducted in units on own accounts and units financed from the budget of the whole economy employing 10 and more employed persons in sections NACE A—O. Breakdown of data includes: type of activity by NACE, size of units, ownership sectors, manual and non-manual posts, gender only for number of employed persons, NUTS.

Objective scope refers mainly to number of employed persons, average employment, hours of work, earnings with specifications following earnings components: bonuses from profit, income taxes, social security contributions paid by employees, agents' earnings, homeworkers' earnings, apprentices' earnings.

Response rate is about 95%.

As for the dissemination policy, annual data are available in annual publications up to 6 months after the reference period.

Annual earnings statistics in small units of the whole economy employing up to 9 persons is conducted on the basis on sample survey on economic activity — the size of sample is 4% of the random frame.

Breakdown of data includes: type of activity by NACE, ownership sectors, NUTS.

Objective scope of this survey in the field of labour statistics refers to number of employed persons, average employment, earnings.

Response rate is up to 50 %.

As for the dissemination policy, annual data on small units are used in annual publications up to 6 months after the reference period.

With reference to statistics conducted on the basis different than annual it is important to mention *the structure of earnings survey by occupations* (the SES). The SES gives the detailed picture on earnings statistics by different socio-demographic features. This survey is also used as the source for the structural indicator -gender pay gap giving the information for the aim of analysis of equal treatment of men and women on labour market. Eurostat shows strong interest to harmonise the coverage and definitions of variables from this survey with the adequate variables from the labour costs survey.

Structure of earnings survey by occupations is conducted every 2-year on the sample basis (20% of the random frame) in units of the whole national economy (type of activity NACE: A—O) employing 10 and more employed persons. The GUS makes efforts to recognize the possibility to enlarge the coverage of units by including also units up to 9 employed persons.

Breakdown of data includes: socio-demographic characteristics of employees (full-time employees up to 2004) namely: age, sex, occupations by ISCO'88, level of education by ISCED'97, work seniority and the other characteristics by type of activity by NACE, size of units, ownership sectors, NUTS.

Objective scope refers mainly to number of employed persons at the end of October, number of employees, hours paid, selected components of earnings (personal earnings, bonuses from profit, 13-th payments).

Up to 2002, annual information on earnings was not available from the SES and was replaced by estimations. In the SES 2004 all necessary improvements were introduced according to Eurostat recommendations.

Response rate in this survey is about 74 %. At the present the GUS is under preparation of the SES 2006.

As for the dissemination policy, data are presented in the publication on structure of earnings by occupations up to 11 months after the reference period, in Labour Yearbooks edited every 2—3 year, annual yearbooks. Up to this time data on SES were transmitted to Eurostat according to required technical format and up to the deadline given in the EC regulation No 1916/2000 for the SES 2002. Quality rules were fulfilled according to EC regulation No 72/2002.

It is worth to stress that Polish SES is strong relevant to users needs (orders of governmental bodies, individual customers research institutes, Eurostat, ILO are fulfilled). Accuracy of data from this survey is recognized by investigation of sampling errors (calculation of standard errors for main variables) and non-sampling errors. This component of quality is improved by:

- effective cooperation between staff of regional statistical office and respondents;
- different method of data collection: IT program used in the responding unit, paper form sent by post, diskette, ;
- strict arithmetical and logical control of data.

SES is the source of data that is well described and well documented (publications include methodological notes, tables and graphs which are presented in clear, friendly way for users). Data from this survey are prepared timely (when it is possible all delays are reduced). All changes in definitions and methodology of this survey are recognized and described what is very important for the comparability over time and geographical comparability. For the analytical aims all data are given in comparable conditions. As for coherence SES data with other sources, the data on earnings by occupations and other socio-demographic characteristics (sex, age, level of education, work seniority) may be taken also from the labour force survey but this survey is conducted by households.

To sum up, the development of integrated system on earnings and labour costs statistics from the point of view of quality policy it is necessary to notice that earnings and labour costs statistics is well developed but further efforts for improvements are needed. Polish earnings and labour costs statistics is strong oriented on users needs because the aim of statistical activity is to obtain statistical product that satisfies users’ needs. The priority is to get fresh, timely data with adequate accuracy. Besides, earnings and labour costs statistics should be flexible and include changes in internal and external environment (demands for breakdown by sex, changes in socio-economic classifications, new, attractive way of dissemination of data — Internet, new, attractive way of data collection — the development of electronic forms).

3. Quality reporting in labour costs survey

The next important element of the quality policy in labour costs survey is connected with the development of reporting on data quality in a harmonised and coherent manner with European Statistical System. *The concept of quality reporting* in the labour costs survey is defined by thematic regulations of the European Commission concerning quality evaluation on the labour costs statistics. In the survey on the last labour costs 2004 is used the regulation of the European Commission Regulation No. 698/2006. This regulation replaced two quality regulations which existed before separately for labour costs and structure of earnings surveys. It was caused by the necessity of harmonization quality requirements within the integrated system on earnings and labour costs statistics.

Quality evaluation of data from the LCS is carried out *on the basis of 6 quality components in accordance with European Statistical System requirements*, namely: relevance, accuracy, timeliness and punctuality, accessibility and clarity, comparability, coherence.

Data relevance is the most important component of quality that is connected to the other quality components, such as: accuracy, timeliness and punctuality, accessibility and clarity, comparability and coherence. The relevance analysis takes under consideration categories of users, their needs and the level of these needs satisfaction.

The main domestic users in the survey on the labour costs are:

- ministries and central offices, i.a.: the Ministry of Economy, Ministry of Labour and Social Policy; Ministry of Finances;
- investors, employers;
- scientific institutions;
- trade unions;
- media.

Among the foreign users of the survey data are, i.a.:

- The Statistical Office of the European Commission Eurostat;
- International Labour Office;
- European Central Bank;
- OECD.

The data of the survey are used in carrying out economic and social analyses that comprise basis for developing the policies concerning economic growth, employment and prevention of unemployment. The survey results are mainly available in a form of tabulations, tables, and charts (graphs) in statistical yearbooks and publications.

Data accuracy of the sample survey, that is the survey on labour costs, is assessed on the basis of the analysis of sampling and non-sampling errors. Therefore, minimalisation of these errors significantly influences improvement of data quality, proper interpretation of the results.

Sampling errors are connected with the sample size, sampling frame and sample design. Their nature derives from the fact that lack of the complete information on a phenomenon influences lack of confidence concerning accuracy of the estimates obtained from a sample survey.

Therefore, the results of an incomplete survey should be treated only as the approximated estimate on the value of the unknown parameter from the population. On the one hand, we should *be aware of the imperfect reliability of the results* (i.e. existing differences between the values obtained from a sample and the actual value observed in the population, which is possible to determine only after carrying out the complete enumeration). On the other hand, we should proceed in such a way as to *maximize data reliability through adequate enlargement of a sample.*

The mentioned conditions were included in the labour costs survey. The sample drawn to the survey reflects the structure of the whole population and it has representative character. Stratified proportional sampling scheme was applied.

The estimation on sample errors in the survey is carried out on the basis of relative standard error.

MSE is a measure of data accuracy. The lower is MSE the higher is accuracy, and vice versa. In the survey on the labour costs *the relative standard errors of the estimators for the selected results of the labour costs survey were calculated (*e.g. the relative standard errors for labour costs total amounted to 0.92 %, for average monthly labour costs per 1 employee — 0.45%, for labour costs per 1 hour paid — 0.51%, for labour costs per 1 hour worked — 0.52%).

Non- sampling errors are divided into:

- coverage errors;
- errors of measurement;
- data processing errors;
- non-response errors;

Coverage errors include over-coverage errors and under-coverage errors.

Over-coverage errors in labour costs survey concern mainly inactive entities, liquidated entities and entities with the lack of contact. This kind of error in the labour costs survey for 2004, amounted to about 7 %. *Under-coverage* refers to new units not included in the frame, either through real birth and wrongly classified units. Measurement of *under-coverage errors* is difficult, thus they were not analysed .

Errors of measurements are connected with the occurrence of the following categories of errors:

- errors that derive from the respondent — these errors are due to misunderstanding of the survey objective and subjective range of the questionnaire. They consist in giving by respondents incomplete answers, skipping some questions of the questionnaire;
- errors that derive from the questionnaire — these errors are connected with unclear questions included in the questionnaire, e.g. respondents had

difficulties with payments for vacations due to the lack of relevant records in the establishment;

- errors that result from the uncompleted answers given by statistical staff for the questions asked by the respondents — such errors were reduced due to the detailed instruction concerning the survey methodology and explanations of the variables included in the questionnaire;
- errors due to the method of data collection: in the survey on labour costs the mailing via post method was used. This method evoked the risk of some questionnaires missing in the process of mailing what might cause non-response.

Data processing errors result from the errors connected to feeding data from the questionnaires to the computer system, and incorrect coding of the data. This kind of errors was limited in the labour costs survey for 2004 due to the appliance of the OCR technique, i.e. optical data reading. The possibility to scan questionnaires significantly speeded the data processing phase, while the assumptions for logical control and control of calculation resulted in a thorough analysis on correctness of the input data.

Non-response errors were due to the occurrence of the units that were inactive or liquidated, lack of contact and refusals. *The non-response rate* in the survey on labour costs for 2004 comprised 20% of which refusals 13%.

Evaluation of the survey data timeliness and punctuality is completed throughout the analysis of the survey timetable regarding the preliminary phase, field work connected with data collection, processing of the results, and the dissemination phase. In the survey on labour costs for 2004, the preliminary phase covered the period between April 2004 and March 2005. It comprised, i.a. preparation of the questionnaire, assumption for the sample and sampling process, preparation of the instructions. Carrying out the field survey that involved mailing the questionnaires to the reporting units, and then sending completed questionnaires to the regional statistical offices took place between the mid March 2005 and mid April 2005. The processing of the survey results was carried out in the regional statistical offices and COIS — Computing Centre in Radom between May and July 2005. The phase of the survey results dissemination started in November 2005. Recapitulating the *data timeliness and punctuality* in labour costs survey, it should be notified that the period between the reference year (2004) and dissemination of the survey results was 11 months.

The analysis on data accessibility and clarity in the survey on labour costs concerns the means used for dissemination of these statistics and the methods of data presentation. The data on labour costs are disseminated mainly in a form of publications available on CD, Internet and paper forms, statistical yearbooks (i.a. Labour Yearbook). The data are presented in a form of tables and in a graphic form as charts. Publications include definitions of the variables and the survey methodology.

Data comparability in the survey on the labour costs is harmonized in accordance with the requirements of the European Statistical System. Definitions of the variables and the applied classifications are based on the thematic regulations of the European Commission. There is ensured data comparability in both aspects: over time, as well as spatial one.

Data coherence determining the ability of particular statistics for the secondary use plays a very important role in the labour costs survey, as the subjective range of this survey is wide. Therefore, the variables covered with the survey, e.g. the employed persons, average employment, wages and salaries, hours paid, hours worked are used also in other statistics, *inter alia* in:

- surveys on employment and wages and salaries;
- labour force survey;
- structural business statistics;
- national accounts.

From the point of view of the users' needs, the important role in the analysis on coherence plays explanation of the differences in variables of the same names coming from different sources. The classic example of such situation is labour costs in labour statistics which source of data is the labour costs survey and employment costs (compensations costs) in national accounts.

The *costs of labour comprise the sum of gross wages and salaries* (including deductions made towards income tax of natural persons and contributions towards compulsory retirement, pension, and sickness insurances (paid by the insured persons)), and *non-wage expenditures* (*inter alia* contributions towards retirement pay, disability pension, and in case of accidents paid by the employers, expenditures on occupational training and the staff re-qualification). *The employment costs (compensation costs) in the context of the national accounts* comprise the component of the accounts on income generating. They comprise wages and salaries (wages and salaries concern amounts included in the costs of activities of a given period, therefore they do not include prizes and bonuses due to participation in profits, and balance surplus in cooperatives), contributions towards compulsory social security (paid by the employers and insured employees) plus contributions towards Labour Fund and other employment related costs, e.g.: prizes, funded scholarships, and premiums not included in wages and salaries, also deductions towards the enterprise social fund, sustenance pay on business trips, the MPs and Senators allowances, in case of the incomes in the household sector the incomes defined as „other incomes related to paid, hired work".

The fundamental differences between the labour costs as the concept used in labour statistics and employment costs (compensation costs) for the use of national accounts, are due to the fact that the labour costs survey refers to units employing 10 and more employed persons and conducting economic activity within the range of NACE sections: A—O. While, the statistical observation in the national accounts covers enterprises (regardless of the number of the employed) and households in the sphere of registered employment as well as in the hidden economy.

When recapitulating the analysis on quality dimensions in the survey on labour costs, it should be underlined that data quality is sufficient from the point of view of the adequate satisfaction of users' needs. *Continuous improvements of the statistical process concerning the labour costs* (eliminating labour absorbing questions from the questionnaire, elaboration of the detailed instructions with a particular stress on the most often mentioned inquiries and doubts, appliance of the data imputation in case of non-response, the use of the results weighting methods with due consideration of non-responses, development of controlling procedures in the electronic system of data processing), *as well as amendments of the forms of data dissemination concerning this statistics significantly influences improvement of the data quality,* which, in turn, has a positive impact on cognitive value of the carried out analyses.

The strict links with described above quality dimensions are indicated also in the content of DESAP form as the self-assessment checklist for the survey manager that is another element of quality policy implemented in the labour costs survey.

4. The application of DESAP as the self-assessment quality tool in the labour costs survey

The application of DESAP form in the labour costs survey was introduced first time for the LCS 2000 within the pilot project PHARE 2002 on Quality in Statistics.

The checklist DESAP is defined according to recommendation No 15 of the Leadership Group on Quality as the self-assessment tool for survey managers in the European Statistical System. This tool allows to illustrate the quality profile of the survey in the form of the DESAP assessment diagram that may be presented for users of the survey.

The checklist DESAP consists of 7 topics that describe the statistical process from the starting point as the decision to undertake a survey up to improvement cycle.

Within each topic three categories of questions can be distinguished:

I. question with numerous responses i.e. improvement questions;
II. assessment questions with the scale of points (up to 5 points);
III. open questions.

Improvement questions cover a lot of possibilities of answer depending on the choice of the survey manager. The large choice of answers on the one hand diminishes the time consuming connected with the completion of the checklist and on the other hand it initiates and stimulates the creation of quality improvement ideas of statistical process. It is worth to stress that there is the special page in DESAP form prepared for the aim of collection of improvement ideas. This page contains topics of main sections of DESAP with the empty place to make notes of improvement ideas. The designers of DESAP give the rule that this page should be taken out from the DESAP and should be completed at the same time when survey manager works with the separate sections of the DESAP form. Such solution forces the human mind for the creativity and building of new innovations. This exercise of creativity makes from the checklist DESAP very special tool to investigate the quality aspects of statistical processes that can give the inspiration for significant development in the field of up-grading the quality in statistics.

The assessment questions in DESAP are based on 5 degree of assessment points scale for each of investigated quality aspect. The result of answers measured in points is put on the diagram that presents the quality profile for each investigated statistical process. The larger the quality area the better the quality profits.

For the aim of this article the survey manager prepared answers for the assessment questions of DESAP for the LCS 2004 are following:

- overcoverage (Section II/Q6) — 4 points. The un-weight unit non-response was 20% of which overcoverage 7%. For the aim of comparison, the un-weight unit non-response for the LCS 2000 was 21 % of which overcoverage 11%. The changes show improvements in updating of statistical business register by reducing biases caused by dead units or inactive units.
- undercoverage (Section II/Q7) — 4 points. The undercoverage was not recognized and it is assumed that this error is not significant in the labour costs survey because statistical business register is updated and amended including changes in the environment of units.
- misclassification (Section II/Q8) — 4 points. The misclassification error was connected mainly with the change of size of units from medium to big and otherwise. The size of units was established on the basis of statistical business frame and it was not changed during the reference period. The impact of misclassification was assessed as insignificant.
- the necessity of editing (Section IV/Q4) — 3 points. The editing of data was improved in comparison with the previous LCS 2000 thanks to introduction of the OCR technique i.e. optical data reading and constantly improvement of arithmetical and logical control but further development is needed.
- Coefficient of variances for key variables (Section V/Q6) — 3 points. Instead of the coefficient of variances the standard sampling errors were calculated for key indicators of labour costs. The basic information on standard sampling errors is presented in section 3 of this article. It is

conducted the steady improvement in updating the frame and in the sampling scheme to reduce the impact of sampling errors.

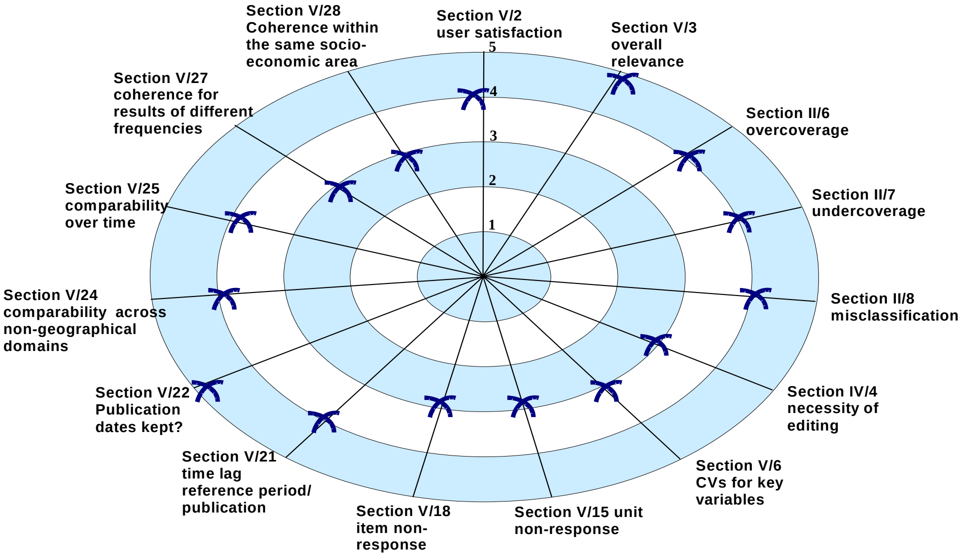
- unit non-response (Section V/Q15) — 3 points. The un-weighted unit non-response is 20%. This result is not bad but significant improvements are still needed.
- item non-response (Section V/Q18) — 3 points. Item non-response is within the brackets (5—15%). This phenomenon is overcome by using the data for the same variables from coherent annual survey on employment, earnings and working hours completed by units with the similar characteristics in strata or by using the average value from strata for the missing variable.
- Time lag between the reference period and the dissemination of publication (SectionV/Q21)— 4 points. The publication „Labour costs in the national economy in 2004" was disseminated 11 months after the reference period. In comparison with the other member countries the Polish timeliness and punctuality of labour costs data is competitive.
- Publication dates kept (Section V/22) — 5 points. Publication on labour costs was published timely according to deadline given in the editorial plan of the GUS.
- Comparability across non geographical domains (Section V/24) — 4 points. Comparability is kept. All changes in definitions of data and changes in classifications are well documented.
- Comparability over time (Section V/25) — 4 points. Comparability is kept. All changes in definitions of data and changes in classifications are well documented.
- Coherence for results of different frequencies (SectionV/27) — 3 points. The coherence is still developed within the work on the integrated system of earnings and labour costs statistics.
- Coherence within the same socio-economic area (Section V/28) — 3 points. The coherence is still developed within the work on the integrated system of labour statistics. Coherence as the one of quality dimensions is recognized systematically in quality reporting to Eurostat beginning from the LCS2000.
- User satisfaction (Section V/Q2) — 4 points. User satisfaction is recognized on the capacity of fulfilment of users' orders and the recognition of ways of data dissemination. It is strong interest to focus on users and to develop customer services in data dissemination policy of the GUS.
- Overall relevance (Section V/Q3) — 5 points. Overall relevance of labour costs statistics is based on the aims formulated by essential users of data, mainly by the European Commission, International Labour Organisation. All main national and international users and their needs are recognized. The data on labour costs have high priority in the creation on the

investment and employment policy at the national level and at the European level.

It is worth to stress that the answers presented above expressed only the survey manager point of view. This approach proposed by DESAP by some people is treated as disadvantage of this tool, because survey manager is emotionally linked with the statistical process and he has the tendency to overestimate some quality aspect. Nevertheless, it is worth to notice that work with DESAP is the special training for a survey manager that helps him to create the way of thinking oriented on quality improvements. This value of DESAP should be treated as the kind of valuable investment which will bring the effects in the development of statistical process.

On the basis of these answers strengths and weaknesses of the labour costs survey were identified which were also confirmed by the results of quality audit from this survey.

The DESAP assessment diagram - the LCS



Source: The own presentation on the basis of DESAP form.

The topics analysed on the basis of the DESAP form were used for the creation of the standard quality audit report presented in section 5.

5. The quality audit in labour costs survey

The statistical quality audit as the next element of quality policy applied in statistical surveys of which in labour costs survey is under preparation in the

GUS. The experiences in this field from Statistics Sweden are used. The pilot quality audit of the labour costs survey for 2004, which took place from 24.04.2006 to 11.05.2006, is one of the first step to initiate this procedure in Polish statistics. Before starting of the pilot quality audit the general training for auditors and managers of selected surveys was organised in the GUS. In addition the survey manager received the information on aims of audit and criteria of audit 2 weeks before the starting of quality audit. The aim of audit was to identify weaknesses, strengths (of which good practices) and to elaborate the improvement ideas. The criteria of the LCS 2004 quality audit included: evaluation of conformity with users needs, accuracy (sampling errors and non-sampling errors), timeliness and punctuality, accessibility and clarity of metadata. The audit team consisted of 3 persons: survey manager of other survey not connected with the investigated statistical process, IT expert, employee with the speciality in the conducting of statistical surveys from regional statistical office. This team worked during 10 working days in the special room destined for the Audit team. Auditors carefully studied papers, documents and publications on labour costs including especially: the extract from the statistical programme for public statistics, the timetable of the survey, the form on labour costs Z—02, instruction for regional statistical offices how to conduct this survey in the field, description of sampling scheme and grossing-up procedure, the assumptions for IT data processing, the completeness report, the set of final tables and publication tables, comments and opinions from the staff of regional statistical offices regarding the data collection method, the evaluation of the form Z—02, the evaluation of arithmetical and logical control. The auditors used the DESAP self-assessment form as the supplementary source of information for the aim of investigation of statistical process on labour costs.

During the third day from the beginning of the audit the kick-off meeting of auditors was organised with the survey manager. Before this meeting the survey manager received the list of questions from auditors that was next discussed during the meeting. In general, also organised three meetings were organized (including final meeting) with survey manager where was discussed issues for the aim of the content of quality audit report. The relation between audit team and the survey manager was based on mutual exchange of view on different quality aspects in the statistical process of labour costs survey. The discussion gave a lot of inspirations and benefits for both sides on the quality development of labour costs survey. This very good effect was achieved thanks to consciousness that the audit work brings advantages for the survey manager in the form of new solutions that exist in other statistics and that can be implemented in the labour costs survey. For auditors as the advantage can be recognized the familiarity with problems and solutions that occur in other statistical process in which they are not involved. The audit opens the way for unlimited exchange of valuable information that is able to increase the quality by creation of „added value”.

As the final effect of conducted audit the quality audit report was compiled which consisted of following three main parts: the administration information, description of the statistical process of labour costs, findings (good practices, weaknesses, recommendations). *The administrative part* included mainly: the aim of audit, the criteria of audit, the personal data of audit team, the personal data of survey manager and staff in Department of Organisation and Coordination of Statistical Surveys, the character of the survey, the frequency of the survey, the number of reporting units. The second part devoted to *the description of statistical process* was very large and it included: planning of the survey (the analysis of users needs, the aim of the survey), design of the survey (the concept of the survey, design of sampling scheme, compliance and testing of the form), the data collection (the source of data, work organisation and training, the reduction of missing answers, the work in the field), data processing (the recording and coding of data, the editing procedures, the imputation method), data analyses and the data quality evaluation (the conformity with the users needs, accuracy, timeliness and punctuality, comparability, coherence, data analysis, data confidentiality), data dissemination and documentations (metadata, the dissemination strategy, the data management), improvements activity (the capacity to satisfy new needs of users, potential of human capital, the quality management). All mentioned items described in this part included a description of situation and development plan. The last part of quality audit report included *findings in the sphere of good practices, weaknesses and improvement recommendations.* This part has the special meaning for the further elaboration of action plan for the implementation of quality audit recommendations. Remarks of the survey manager regarding the evaluation of work with audit team were attached as the supplement to the quality audit report.

As for good practices, the quality audit report indicated methodology part in the publication ”Labour costs in the national economy in 2004”. This part covered the chapter on data quality in labour costs survey 2004 that presented the evaluation of 6 quality dimensions according to European Statistical System. The aim of the chapter was to spread out the idea of data quality among users of publication. The content of this chapter is presented in section 3 of the article.

With reference to weaknesses, it was noted among others that objective scope of the form on labour costs is large and constituted the burden for respondents. To overcome this difficulty the survey manager eliminates some variables which from the view of users are not necessary to collect them. To facilitate the respondents work with the form the instruction for variables are placed under each section. The another weakness was non- response rate that amounted to 20%. The steady work on updating of the frame and improving of sampling scheme is recommended to reduce a bias due to missing answers.

As for improvement recommendations, auditors prepared following main conclusions:

- With reference of the survey timetable

It is recommended to prepare the survey schedule with the planned deadline for each task and the actual deadline for each task together with details on responsible persons. This schedule facilitates to identify fast millstones in separate phases of statistical process.

- With reference to documentations

It is recommended to prepare the whole documentations in the electronic form.

- With reference to work on publication

It is recommended to include additional persons for the technical works regarding the preparation of graphs, printing of tables and making corrections of publication. It is recommended also to include additional IT experts to prepare the computer programme for data processing. The final report of acceptation of IT system for data processing should be prepared after its testing. It should be considered the introduction of the link including the section Frequent Asked Questions (FAQ) to the publication on labour costs available in Internet.

- With reference to the design of the form and data collection

It is recommended to maintain one e-mail address of respondents on the form. Besides, the statement on the form should be inserted regarding the using of data for statistical aims and observing the confidential rules. This change in design of the form encourages respondents to participate in the survey.

As for data collection method, it should be considered the introduction of electronic form. When the paper collection method will be maintained the respondent should receive envelopers for sending the forms with the reduction of charge. The regional statistical staff should prepare own remarks after the end of work on labour costs survey what will be valuable source of information for improvements. What is more, it should be considered the training of statistical staff before the next edition of LCS for 2008.

- With reference to frame

The continuous work is recommended on the updating of the frame based on business statistical register BJS with the attention on the updating process especially in section D and G. This work should bring the results in the form of reduction of missing answers due to dead units, inactive units and lack of contact.

- With reference to automatisation of data confidentiality

It is useful to use the software to accomplish cell suppression for the aim of confidential data.

- With reference to data coherence

It is recommended to follow up changes in arithmetical and logical control assumptions in coherent surveys with the statistical process on labour costs for the aim of the complex assumptions control in next editions of the labour costs survey.

- The treatment of data in bookkeeping system

It should be considered to contact the certain responding units before starting next edition of labour costs survey to recognise their capacity to complete the form. It will be useful to initiate the preparation by responding units the computer programme that will help them to complete the form on labour costs.

Complex policy of response duties of reporting units is recommended. Such policy will impose duty for the GUS to provide responding units with the easy and fast access to the register of responding duties that would be available for responding units by their own access code on the official Internet side of the GUS.

To sum up of the results of quality audit in labour costs survey, it is worth to stress that the audit report is treated by the survey manager as valuable source of information which gives the new inspirations for further progress in the field of quality improvement of statistical process. What is more, it is predicted to monitor by high supervisors the implementations of recommendations formulated by auditors in statistical process of labour costs survey and in other statistical processes covered by statistical audit.

6. Conclusions

As we see, the quality policy in the labour costs survey includes several quality approaches ranged from the integrated system in the field on earnings and labour costs statistics up to implementation of statistical quality audit. All these efforts are concentrated on monitoring, evaluation and improvement of quality of statistical process. This focus brings a lot of benefits in the form of steady monitoring of changes in quality. The regular monitoring of quality reporting gives the observation on quality dimensions where the progress is noticed and on quality dimensions where improvements are needed. The fulfilment of DESAP form with the part on improvement ideas imposes the way of thinking on quality criteria and brings to find the mutual links between quality criteria. The other quality tool -the audit of statistical process conducted by experts not linked with investigated statistical process offers fresh look how to work better and more effectively. Such obstacles as routines and habits of the survey manager in the treatment of statistical process are overcome by auditors. For certain of us it is difficult to believe that the audit team within the short-term of work with statistical process is able to find solutions which are unable to find by the permanent staff working daily with the statistical process. The one of the explanation of this phenomenon is that the power of willingness, strong motivation and curiosity of audit team is not contaminated by daily routines. The permanent staff is burdened by checked ways of actions, habits, sometimes by lack of enthusiasm. It is important to notice that audit team is able to create „added value" in the friendly cooperation with the survey manager. The survey manager should treat quality audit not as the normal control but as the support in development of statistical process. To create successful factors on the way on making progress in quality we can go further by the implementation of job rotation, job enlargement, delegation of tasks and responsibilities, transparency offered by TQM.

To sum up the quality policy in the management of statistical surveys, it may be concluded that the solutions offered by the quality policy in statistical surveys in short-term period may be treated as the burden- additional duties. On the other hand in the longer perspective they give measurable effects in the form of the development of total quality management procedures and quality culture related to the continuous improvement of statistics production, the enlargement and enrichment of the content of statistical work what happens in the case of the work with labour costs survey and what is the interesting experience of the manager of this survey.

REFERENCES

EUROSTAT (2003). Definition of Quality in Statistics (item 4.2.). Eurostat Working Group on Assessment of Quality in Statistics, Luxembourg, 2—3 October.

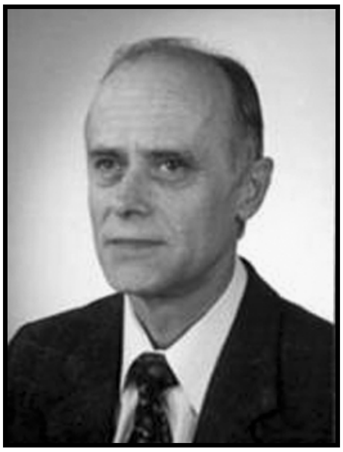
GUS (2005), Labour costs in the national economy in 2004, Information and statistical papers, the PCSO, Warsaw — see item 7 available on http://www.stat.gov.pl/english/dane_spol-gosp/praca_ludnosc/index.htm

LARS LYBERG (2003). Quality improvement in European National Statistical Institutes. Statistics Sweden. Stockholm, 23 November.

STATISTICS IN TRANSITION-new series, April 2007
Vol. 8, No. 1, pp. 211—212

OBITUARY

Professor Andrzej Balicki (1944—2007)



Professor Andrzej Balicki, the Head of the Department of Statistics, University of Gdansk, Poland, died in his house in Gdynia on 23 March, 2007, at the age of 63.

Professor Andrzej Balicki was graduated from the University of Economics in Sopot in 1967. On the basis of the thesis entitled: *"An analysis of the influence of selected variables on the stability level of employees"*, he received his PhD from the University of Gdansk in 1974. Two years later, he published his first book *"Stability of employment. A statistical study"* (in Polish, PWE, Warsaw 1976) in which he developed his ideas presented in PhD thesis. In those years and some years later his main research interests focused on methodological aspects of analysis of mobility of employees and stability of employment in enterprises. In several his papers, published by "*Studia Demograficzne*" (Demographic Studies) in the period 1986-1987, and by "*Wiadomości Statystyczne"* (Statistical News) in the years 1985-1986, Professor Balicki successfully promoted using cohort data for a non-parametric estimation of probability models describing mobility of employees. In his further articles, published after 1989, he attempted to apply a cohort analysis for investigating and modeling distributions of time lengths without jobs among the unemployed (see "Wiadomości Statystyczne" no. 7/1997).

In Poland, Professor Balicki has been well known as a researcher and expert in the field of environmental statistics. The Author of the book *"Statistics in environmental protection studies"* (published in Polish by the Central Statistical Office, Warsaw 1998), and many inspiring papers from this area published by "*Wiadomości Statystyczne"* (see e.g.: 12/1995, 5/1996) and "*Statistics in Transition*" (no. 4/1998), he was one of the most distinguished pioneers forming and developing this branch of statistics in Poland. His research efforts were supported by a prestigious ACE Fellowship Grant, which Professor Balicki received in 1997 in order to continue his research at the University of Nottingham (UK). Later, he several times acknowledged that the few months spent at Nottingham had been a unique opportunity for him to collaborate with Professor Vick Barnett and his colleagues. The book mentioned above was welcomed with great interest by those who dealt with statistics and environment protection. The book also received very good opinions of referees. Professor Vick Barnett in the conclusion of his book review wrote: *Author has on the one hand produced what I believe to be a valuable representation and characterisation of the pollution problems in Poland (...) and on the other hand has been able to show how skilfully applied and quite advanced statistical methods can aid the understanding of major national environmental issues".*

At the Faculty of Management, University of Gdansk, where Professor Balicki used to work for more than 30 years, initially as a lecturer and later as a Professor, he was a founder of a well developing teaching programme called "Actuarial Statistics". Professor Balicki was a supervisor of many Master thesis and several PhD thesis from the area of life insurance. Only few months before his death, Professor Balicki published his new book entitled: *"Survival analysis and mortality tables"* (in Polish, PWE, Warsaw 2006). When the review of the book was published by "*Wiadomości Statystyczne"*, the Author heavily suffering from cancer, was already too week to read it.

Professor Andrzej Balicki left all of us and his work in the most fruitful period of research activity. The most important books and research papers were published at the end of his short life. Some works have been left uncompleted, including another book entitled: *Multivariate statistical analysis and its socio-economic applications"*. It all reflects how hard-working person he was. I recall him also as a very reliable and conscientious colleague and good friend.

Professor Andrzej Balicki received the title of Professor from the President of the Republic of Poland in April 2005. In the period 1990-1996 he was an active member of the following national organisations: the Polish Statistical Society, Mathematical Committee of the Central Statistical Office, the Scientific Statistical Council. He used to hold several responsible positions in the University of Gdansk and in other educational institutions.

Mirosław Szreder, University of Gdanski, Poland