STATISTICS IN TRANSITION

new series

An International Journal of the Polish Statistical Association

CONTENTS

Comparative surveys

Sampling methods and estimation

Other articles

Book review

Conference reports

Volume 11, Number 3, December 2010

Editorial Office

Marek Cierpiał-Wolan, Ph.D.: Scientific Secretary
m.wolan@stat.gov.pl
Roman Popiński, Ph.D.: Secretary
r.popinski@stat.gov.pl; Phone number 00 48 22 — 608 33 66,
Waldemar Orlik: Technical Assistant

ISSN 1234-7655

Address for correspondence

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

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 427—430

EDITOR'S NOTE AND ACKNOWLEDGEMENTS

This volume, the last issue of the *Statistics in Transition* for year 2010, provides us with occasion to express our sincere gratitude to all the journal's collaborators during the past year. We especially warmly thanks to 57 authors of 36 original articles published in the journal, and to 34 internationally recognized experts who served as peer-reviewers – their names are listed below in the 'acknowledgement' section.

The issue starts with new section, devoted to different aspects of surveys conducted on several populations, entitled simply 'Comparative Surveys'. It is comprised of two articles. First paper, *Surveying Child Labour Through Households: Sampling Issues and Strategies,* written by **Vijay Verma, Francesca Gagliardi**, addresses sampling issues arising in the context of household-based child labour surveys. It presents some of the sampling strategies elaborated in the ILO book *Sampling for Household-based Surveys of Child Labour* (Verma, 2008) and offers a typology of surveys of child labour. The fundamental distinction between two types with very different objectives – termed 'child labour surveys' and 'labouring children surveys', respectively – is clarified and emphasised. And linkages between different types of surveys, as well as some specific sampling techniques are being explained, based on a broad survey of national practices in conducting surveys of child labour.

The second paper, by **Bernhard von Rosenbladt,** *Adult education and training in comparative perspective – indicators of participation and country profiles,* while recognizing insufficiency of data for comparative research on adult learning until recently, explores new opportunities based on the European Adult Education Survey (AES). It argues that general indicators of participation in adult education and training must be complemented by more specific indicators revealing sectoral structures and behavioural patterns within the adult learning system. Country variations are explored across a set of 16 European countries.

The nesxt section, on sampling and estimation methods, begins with paper by **Hukum Chandra, HVL Bathla and U C Sud**, *Small Area Estimation Under a Mixture Model.* Since such models (SAE) may not be efficient when data contain substantial proportion of zeros, the SAE for zero-inflated data under a mixture model are specified that account for excess zeros in the data (Fletcher *et al*., 2005 and Karlberg, 2000). Results from simulation studies show that mixture model based approach for SAE works well and produces an efficient set of small area estimates. Also, an application to real survey data from the National Sample Survey Organisation of India demonstrates the satisfactory performance of the approach.

In paper *Fitting General Linear Model for Longitudinal Survey Data under Informative Sampling* **Abdulhakeem A.H. Eideh** discusses the problem of fitting superpopulation model for multivariate observations – in particular, multivariate normal distribution for longitudinal survey data. The proposed approach aims to extract the model holding for the sample data as a function of the model in the population and the first order inclusion probabilities; and then fit the sample model using maximum likelihood, pseudo maximum likelihood and estimating equations methods. An application of the results is illustrated by the general linear model for longitudinal survey data under informative sampling using different covariance structures: the exponential correlation model, the uniform correlation model, and the random effect model, and using different conditional expectations of first order inclusion probabilities given the study variable.

Carl-Erik Särndal's article, *Models in Survey Sampling*, focuses on the two types of approaches in modeling relationships between study variables and auxiliary variables that have influenced survey sampling theory and practice over the last four decades: the design-based and the model-based. Since, in their pure forms, these models offer two fundamentally different outlooks and approaches to inference in sample surveys, a complete reconciliation and agreement cannot be achieved. But the tendency today is that each of the two approaches recognizes and profits from important elements in the other. We see an often fruitful interaction, as discussed in this article.

In paper *On Efficient Difference Type Estimators* **A.K.P.C. Swain** search for a more efficient difference type estimator in a finite population set-up in the presence of auxiliary information. Ratio type and regression type estimators are derived as special cases. Further efficiencies of these estimators are compared with classical ratio and regression estimators and numerical illustrations are provided to compare efficiencies of different competitive estimators.

The section containing 'other articles' is opened with paper by **Sabina Denkowska and Monika Papież** *The Analysis of Mortality Changes In Selected European Countries in the Period 1960–2006,* which discusses some consequences of demographic changes that took place during the 20th century – such as the progressive ageing of European societies – for calculation of risk by insurance companies and pension funds. Since mortality is considered one of the most important factors in such calculations, the changes in male and female populations in selected countries of Central Europe (the Czech Republic, Hungary, Poland and Slovakia) and of Western Europe (France, Italy, Spain and Sweden), are of object of the analysis of data for the period 1960–2006. The analysis of the mortality changes has been carried out with the use of variables proposed J. P. Morgan (2007), using data available from www.mortality.org and employing the van Broekhoven algorithm for smoothing crude mortality rates across different ages.

A very important for economic policy issue of identifying change points in economic and financial series is discussed in **Reza Habibi's** paper *Distribution Approximations for Cusum and Cusumsq Statistics.* Using the cumulative sum (cusum) as statistic in testing for a change point, this paper considers the distribution approximations to the cusum statistic under the null and alternative hypotheses. Also, distribution approximations for the cumulative sum of squares (cusumsq) test statistic are under considerations, and some comparisons are made in a discussion section.

Some alternative approaches to decomposition of inequality are discussed by **Maurro Mussini** in paper *On the Link between Silber and Dagum Decomposition of the Gini Index.* The presented, combined decomposition can be used for overlapping as well as for no overlapping population subgroups, taking into account within-group and between-group, and overlapping inequalities. For this, all the information on income distribution contained in one matrix are being exploited, accounting for pairwise disparities between per capita income shares. The proposed matrix approach provides also insight into a more complex analysis of overlapping. Application of the new methodology to data from Italian employee income in 2000 and 2008 illustrates its usefulness.

In paper *A Typology of Polish Farms Using Probabilistic d–clustering,* **Andrzej Młodak** and **Jan Kubacki** search for an effective typology of Polish farms based on data collected from administrative sources during the preliminary agricultural census conducted in autumn 2009. A universal form of typology is proposed using fuzzy clustering method (that has been developed to this aim), with probabilistic d–clustering for interval data. The relevant criteria are arbitrarily established, but also, as an alternative way, are generated endogenically using an original optimization algorithm. For a comparison, relevant classification for data collected "from nature" is also provided.

Włodzimierz OKRASA
Editor-in-Chief

ACKNOWLEDGMENT OF REVIEWERS

The Editor and Editorial Board wish to thank the following persons who served since 31 December 2009 as reviewers of manuscripts for the ***Statistics in Transition new series*** – **Volume 11, Numbers 1–3;** the authors' work has benefited from their feedback.

(Note: In some cases reviewers' institutional affiliations may have changed).

Aleksandra Baszczyńska, University of Łódź, Poland
Jacek Białek, University of Łódź, Poland
Katarzyna Bolonek-Lasoń, University of Łódź, Poland

Denis Conniffe, University of Ireland – Maynooth, Ireland
Czesław Domański, University of Łódź, Poland
Nicolas Farmakis, Aristotle University of Thessaloniki, Greece
Elżbieta Getka-Wilczyńska, Warsaw School of Economics, Poland
Alina Jędrzejczak, University of Łódź, Poland
Cem Kedilar, Hacettepe University of Ankara, Turkey
Jan Kordos, Warsaw School of Economics, Poland
Jerzy Korzeniowski, University of Łódź, Poland
Barbara Kowalczyk, Warsaw School of Economics, Poland
Jerzy T.Kowaleski, University of Łódź, Poland
Nicholas T.Longford, Pompeu Fabra University, Spain.
Krzysztof Marczewski, Faculty of Physiotherapy and Pedagogy of Zamość University
George Menexes, Aristotle University of Thessaloniki, Greece
Andrzej Młodak, Statistical Office Poznań.
Amjad D. Al-Nasser, Yarmouk University of Irbid, Jordan
Włodzimierz Okrasa, Cardinal Stefan Wyszynski University in Warsaw and Central Statistical Office, Poland
Walenty Ostasiewicz, Wrocław University of Economics, Poland.
Iannis Papadimitriou, Aristotle University of Thessaloniki, Greece
Dorota Pekasiewicz, University of Łódź, Poland.
Mariusz Pilch, University of Łódź, Poland
Waldemar Popiński, Central Statistical Office, Poland.
Agnieszka Rossa, University of Łódź, Poland.
Mauro Scanu, ISTAT, Roma, Italy
Divakar Shukla, Dr.H.S.Gaur University of Sagar, India
Meenakshi Srivastava, Dr.B.R. Ambedkar University (formerly Agra University),India
Grażyna Trzpiot, Academy of Economics, Katowice, Poland.
Vijay Verma, University of Siena, Italy
Jacek Wesołowski, Warsaw University of Technology, Warsaw, Poland
Feliks Wysocki, Poznań Poznań University of Life Sciences, Poland
Janusz Wywiał, Academy of Economics, Katowice, Poland
Benhuai Xie, Takeda Global Research and Development, Minnesota, USA
Janusz Żądło, Academy of Economics, Katowice, Poland

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 431

SUBMISSION INFORMATION FOR AUTHORS

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

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

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

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

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

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

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

For other aspects of editorial policies and procedures see the SiT Guidelines on its Web site: http://www.stat.gov.pl/pts/15_ENG_HTML.htm

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 433—464

SURVEYING CHILD LABOUR THROUGH HOUSEHOLDS: SAMPLING ISSUES AND STRATEGIES

Vijay Verma[1], Francesca Gagliardi[2]

ABSTRACT

This paper addresses sampling issues arising in the context of household-based child labour surveys. It presents some of the sampling strategies elaborated in the ILO book *Sampling for Household-based Surveys of Child Labour* (Verma, 2008). A typology of surveys of child labour is identified, and the fundamental distinction between two types with very different objectives – termed 'child labour surveys' and 'labouring children surveys', respectively – is clarified and emphasised. Following a broad survey of national practices in conducting surveys of child labour, linkages between different types of surveys and some specific sampling techniques are explained.

Key words: child labour, household surveys, sampling, ILO.

1. Introduction

Child labour is an important global issue. Detailed and up-to-date statistics on working children are needed to determine the magnitude and nature of the problem, identify the factors behind child labour, reveal its consequences, generate public awareness of the related constellation of issues, and to formulate policies and projects to combat it (International Labour Organisation, 2004).

Data on child labour may be obtained from diverse sources, often used in combination. Although few *national population censuses* provide data on the prevalence of child labour, information from censuses serves as an essential basis for the interpretation and analysis of data on child labour from other sources. The population census is also the basic source of sampling frames for child labour and similar surveys. Countries collect socio-economic and demographic data through *general household-based sample surveys*, such as surveys on the labour force, living conditions, household income and expenditure, demography and health. Such surveys normally do not produce detailed data on child labour, but they can yield information that is useful for analysis of the situation concerning child

[1] University of Siena, verma@unisi.it.
[2] University of Siena, gagliardi10@unisi.it.

labour. Moreover, attaching child labour modules to such household-based surveys is also a potential source of information. In addition, a wide range of *secondary and administrative sources*, while not primarily concerned with child labour, can provide useful information pertaining to it.

Nevertheless, more comprehensive and pertinent information on child labour requires special studies and surveys focussed on the subject. Apart from *household-based child labour surveys*, with which we are concerned here, different types of instruments include rapid assessments, establishment surveys, school-based surveys, community-level enquiries, street children surveys, and baseline studies undertaken in the context of specific intervention projects. The various sources are generally complementary. The practical implication of this from the point of survey and sampling design is that the household-based instrument need not be developed to meet all the information needs: in fact some of these needs are better (or even, can only be) met from other types of instruments. Experience demonstrates that collecting comprehensive data on child labour is a challenging task, and no single survey method may in itself satisfy all data needs. “Children are found working in a vast array of circumstances, and no single technique can be devised to survey all of these situations. Furthermore, policy analysis and targeted project intervention require information from a variety of potential respondents who may influence the life and development path of the child. These include the children themselves, parents or guardians, employers, school teachers, community leaders, child peers, and siblings. Circumstances in the home, school, workplace, and larger community to which the child belongs all bear on child labour outcomes and characteristics. To collect all relevant data from all relevant parties by means of a single survey or on a single occasion is impossible.” (ILO, 2004). This applies with particular force in surveying what are called the *worst forms of child labour*. It is practically impossible to make contact with children engaged in such forms of labour to collect the necessary information. Worst forms of child labour usually remain hidden and the necessary sampling frames do not exist for their enumeration. Nor can the required samples be designed and selected without prior information on the location, characteristics and circumstances of the children engaged in it. Consequently, regular household-based surveys are largely ineffective for this purpose; special sampling and enumeration procedures must be employed.

This paper is concerned with sampling issues arising in the context of household-based child labour surveys. It presents, albeit briefly and selectively, the sampling strategies elaborated in the ILO book *Sampling for Household-based Surveys of Child Labour* (Verma, 2008). This book has been reviewed in *Statistics in Transition* by Kordos (2008).

As noted, for some purposes and in certain circumstances child labour surveys may also involve non-household based data collection, and may even force some departures from the principles of probability sampling. This paper does not aim to address special considerations involved in the design of surveys aimed at estimating the prevalence and nature of child labour confined to particular sectors or activities. Various statistical techniques used for sampling non-standard units need separate treatment, though many of the techniques discussed below can be useful in the design of such surveys as well.

2. Household-based surveys of child labour: a typology

2.1.Household-based surveys

Regular child labour surveys are household-based national sample surveys whose target are children, and also their parents or guardians living in the same household. Such surveys may be conducted as stand-alone surveys, or as separate but linked operations, or simply as modules attached to other national household-based surveys such as a labour force survey (LFS). The statistics generated by these surveys include economic activities and non-economic activities (such as household chores) of children, working hours, nature of the tasks performed, health and safety issues including injuries at work, and also background variables such as demographic and social characteristics of household members and other basic characteristics of the household.

Household-based sampling provide an efficient approach for estimating the prevalence and characteristics of predominant forms of child labour for children living in private households, irrespective of whether the work is performed at home or outside. In so far the survey is based on a scientifically designed probability sample, it permits generalization of the study results to the whole population that was sampled.

In practice, the household survey content may be detailed and specialised, providing information on the dynamics of child labour or gross flows between different child labour categories; or it may be confined to a few basic characteristics of working children. The choice depends on the data needs, available resources, and the arrangements and circumstances under which the survey is conducted. The key respondents in a child labour survey are the working and potentially working children and their parents or guardians.

Household surveys may apply a variety of designs and organizational structures. The main factors determining the design are, of course, substantive objectives concerning the content, complexity and periodicity of the information sought. These substantive requirements determine features of the survey structure such as its timing, frequency, reference period and sampling arrangements. For instance, the survey may be a continuing survey designed to obtain regular time-series data or, as has been more often the case in national surveys, it may be an occasional survey primarily for obtaining benchmark and structural information.

In the designs followed most commonly for the labour force and similar population-based surveys, especially in developing countries, the sample is selected in two (or more) stages: the selection of area units; followed by the selection of addresses, households or individuals in each area as the ultimate sampling units. For convenience and concreteness of the exposition, we will assume such a ‘typical’ design throughout in the following technical discussion of sampling issues.

2.2. Child labour survey (CLS)

As a generic term we use ‘a (regular household-based) child labour survey’ to indicate a household sample survey the main objective of which is to provide information on the phenomenon of child labour – its prevalence, distribution, forms, economic sectors etc., as well as its conditions, characteristics and consequences. The prefix ‘regular’ is used to emphasise that the context is that of a broad household-based survey, as distinct from other types of studies concerning children not residing in – or at least not identified for enumeration through – private households.

While retaining the above more general, descriptive use of the term ‘child labour surveys’, it is very useful to keep in mind two quite different types of such surveys. These differ in their objectives, or at least in the emphasis given to different types of objectives.

The first type refers to surveys where the primary objective is to measure the *prevalence* of child labour. The surveys may also study variations in this prevalence by geographical location, type of place (urban-rural), household type and characteristics, the household’s employment and income situation, children’s age and gender, and similar factors. The target population of a survey with this type objectives is *the total population of children exposed to the risk of child labour*. This base population is defined essentially in terms of age limits, and therefore tends to be well-distributed in the general population. The size and structure of the sample is determined largely by the size and distribution of the population of all children, or more commonly by its approximation – the size and distribution of the general population.

We propose to reserve a more strict use of the term *Child Labour Survey* (CLS) for such surveys with the primary objective of measuring the prevalence of child labour, as distinguished from a *Labouring Children Survey* (LCS) described below. The defining factor in this distinction is the relevant base population for which the survey estimates are generated – essentially, all children within specified age limits for the CLS, and only those considered to be in child labour for the LCS.

2.3. Labouring children survey (LCS)

We have a different type of survey when the primary objective is to investigate *circumstances, characteristics and consequences of child labour*: what types of children are engaged in work-related activities, what types of work children do, the circumstances and conditions under which children work, the effect of work on their education, health, physical and moral development, and so on. The objectives may also include investigating the immediate causes and consequences of children falling into labour. We refer to this type of survey as a *Labouring Children Survey* (LCS). The relevant base population in the LCS is *the population of working children*. What is meant by the LCS concept is that, when the objective is to determine the conditions and consequences of child labour, as distinct from its prevalence among all children, then it is appropriate that the size and structure of the sample is determined primarily by the size and distribution of the population of working children.

At the same time, it is important to clarify that the concept of a 'labouring children survey' does not imply that the ultimate units enumerated in the survey be only labouring children. On the contrary, it will normally be necessary in such a survey to enumerate comparable groups of children not engaged in labour, so as to provide a control group for comparison with the characteristics and circumstances of those subject to child labour. Nevertheless, the sample size and design of a LCS is determined primarily by the need to represent the population of labouring children; any sample of non-labouring children is supplementary, selected and added to the main sample as necessary for analytical purposes.

2.4.CLS vs. LCS: sampling implications

There are some important differences in terms of sampling aspects between Child Labour Surveys and Labouring Children Surveys in the above sense.

The target population of the CLS, being all children in a certain age range, tends to be distributed in a way very similar to the general population. Hence the required structure and distribution of the CLS sample is likely to be quite similar to that of a survey of the general population, in particular to that of the Labour Force Survey (LFS) to which the CLS is very similar in concepts, definitions and even survey content.

The target population of the LCS – whether the population of children engaged in any work-related activity, or defined more narrowly as these engaged in specific forms of child labour – is, by comparison, smaller and more unevenly distributed, often in areas of heavy concentration. Consequently, the sample design required is generally also different from that of LFS or CLS.

The two types of surveys differ in their size and complexity. The CLS is normally less intensive (i.e., involves a simpler and shorter survey interview), and requires larger sample sizes. The primary statistical consideration dictating its sample size is the precision with which the proportions of children engaged in child labour is to be estimated and the reporting domains requiring separate estimation.

By contrast, for investigating the detailed conditions and consequences of child labour, the LCS is more intensive in data collection, often involving interviewing the guardians as well as the children concerned separately, collecting attitudinal and other qualitative data, and carrying out associated enquiries such as at the school or place of work of children in the sample. Consequently, the appropriate sample size for a LCS is likely to be much smaller than that for a CLS in similar circumstances. For an intensive survey such as the LCS, large sample sizes are often unnecessary from the statistical point of view, and in any case are precluded by practical and cost considerations. Having too large a sample in an intensive survey can in fact damage the quality and value of the information collected, in so far as it hinders close control over the survey operation.

2.5.Labouring children vs. child activity vs. children's surveys

In practice, the target populations of interest can be more diverse than the all children versus labouring children distinction of CLS and LCS.

The focus of all LCS's is on the details of child labour, or more generally, on children's work-related activities. However, some surveys collect a broader range of information on children. The following three types of situations may be distinguished.

Option (1). A majority of the LCS surveys have as their main focus the study of conditions and consequences of child labour (see Section 5 for a review of country practices).

Option (2). In a number of countries, the scope is broader, and covers all types of activities of children, including economic and non-economic activities, education, leisure, and even non-activity. Many such national surveys are actually named a *Child Activity Survey* (CAS).

Option (3). Occasionally, the scope is even broader to include more general information about children beyond their economic and non-economic activity – such as information on children's health, housing conditions, etc. These are termed a *Children's Survey*.

There is also much debate as to the definition of what constitutes 'child labour', and in particular whether 'substantial' domestic chores should be included (ILO, 2004, Chapter 2).

These variations have important sampling implications. For surveys focused on working children, option (1), the LCS samples should reflect closely the patterns of concentration of child labour. With their broader and more defused scope children's surveys, option (3), would require a design similar to that of the CLS, or may even incorporate the latter. However, even for this type of surveys, and especially for the child activity surveys, option (2), the measurement of conditions of child labour as such is likely to remain a special objective. A compromise design is therefore desirable – which covers both working and non-working children but with greater weight given to the former. Such a compromise design would of course require a CLS-type operation preceding it, so as to identify – even if with limited precision – the level of child labour in survey areas.

2.6. An example

The diverse objectives of a survey on child labour, and the diverse population groups to which the results from the survey apply, are well-illustrated by a survey from Portugal Ministry of Labour and Solidarity (1998). The survey identifies seven different target groups of questions. For three of these groups, namely

(1) families and the work of children,

(2) children and their activities in general, and

(3) aspects of children's life

the target population is all children (and their families). The questionnaire modules pertaining to these constitute the CLS component of the survey. Group (1) is the basis for estimating the proportion of working children. For another three groups, namely

(4) characteristics of children with economic activity,

(5) characterisation of those responsible for children with economic activity, and

(6) attitudes and perceptions towards child labour of the child, and of adults responsible for the child

the target population is labouring children, defined here as children engaged in any economic activity. The questionnaire modules pertaining to these constitute the LCS component of the survey. Group (4) – children engaged in economic activity – is the basis; the population of persons responsible for children in groups (5) and (6) is defined through association with the base population in group (4). The group

(7) children who carry out domestic chores

covers a somewhat different group of children. This group is often considered less important than group (4) in determining the LCS sample size requirements.

The final sample size is normally a compromise between the requirements of the CLS and LCS components. It is of course possible, in principle, to introduce sub-sampling from group (1) to group (4) (i.e., to follow-up only a sub-sample of the labouring children identified from the former); or to introduce sub-sampling from group (4) to groups (5)–(6) (i.e., to follow-up only a sub-sample of adults responsible for the labouring children). There is greater flexibility in this respect when the various components involved are operationally separated from each other.

3. Linkage of CLS to a base survey

Surveys may be needed to collect different types of information on child labour. As noted, different types of data differ in their content, mode of collection, and – with particular relevance to sampling – in the base population to which they relate and the required sample size. *Different types of data involved may be viewed as constituting different components of the survey.* The components may, for instance, be combined into a single integrated whole; or they may remain distinct but linked in various ways; or they may form more or less separate stand-alone operations, each like a survey in itself. Similarly, the child labour survey or its components may be related in various ways to other existing surveys, such as a labour force survey. In designing the sample for a household-based child labour survey, the first step is therefore to define the survey structure, i.e., choose the manner in which different components of the survey are to be arranged in relation to each other.

In this section we consider the linkage of the CLS with the operation preceding it. There are three dimensions: (i) the preceding operation may be a household listing operation, or it may be a large-scale survey such as the LFS; (ii) the CLS may be combined with that operation, or be conducted subsequently as a separate operation; or (iii) the CLS may be conducted on the same sample, or on a sub-sample of the preceding operation. The combinations of these are shown below. Of course some of these combinations are more likely (more meaningful, practical) than others.

Link of a CLS with the operation preceding it

Whether the two operations are integrated or separate				
Preceding operation	Integrated		Separate	
Household listing	CLS involving brief screening questions	same (full) sample	Stand-alone CLS	using the full listing sample (unlikely option)
		sub-sampling (unlikely option)		sub-sampling
LFS or similar	Modular or 'combined' CLS	same (full) sample (used frequently)	Linked CLS	same (full) sample (used sometimes)
		sub-sampling (used sometimes)		sub-sampling (used frequently)

3.1.Modular or 'combined' CLS

The collection of information on child labour in conjunction with a broad-based survey of the general population such as the labour force survey (LFS) may typically take the form of child labour questions attached to the LFS as a *module.* In this case, the essential CLS information, namely estimates of proportions of children in various categories who are engaged in child labour, may be obtained by extending downwards the lower age limit for the standard LFS questions on economic activity. This is a possibility when child labour is defined in terms of the standard LFS concept of economic activity (on the latter, see Hussmanns *et al*, 1990). Different and generally more elaborate questioning will be required with a different interpretation of what is meant by 'child labour'.

The major attraction of a modular child labour survey is that it provides an economical and convenient arrangement for obtaining essential information on child labour. Furthermore, items enumerated in the base survey are available for use as explanatory and classification variables in the analysis of child labour data. Modular surveys present some potential problems, however. The number and detail of child labour items that may reasonably be inserted into an operation primarily concerned with other topics is quite limited. Secondly, in order to ensure high-quality data the various survey topics must be compatible in terms of concepts, definitions, survey methods, reference periods, coverage, and design requirements. This compatibility requirement may enforce compromises that limit the usefulness of the resulting data.

When a set of child labour questions is attached as a module to an existing base survey, it is generally understood that the number of additional questions involved is small enough not to significantly affect the base survey sample design or data collection. We use the term *combined survey* to refer to a more comprehensive version of the modular survey. It indicates the situation when the child labour questions constitute a substantial addition to the base survey, influencing the survey design, sample size and data collection operations of the latter.

Any of the above arrangements involves operational integration of the CLS with the base survey. This normally implies that the CLS module is applied on the same sample of household as the base survey. Sub-sampling at the area level (i.e. introducing the CLS as a module only in a subsample of LFS areas) can however be possible. A particularly convenient form of this is to include the CLS as a module during only some of the rounds of a continuing LFS.

3.2.Linked CLS and its sampling aspects

A *linked CLS* means a survey dependent on the base survey (such as a LFS) for its sample and possibly also for other information fed forward, but otherwise operationally separated from the latter for the purpose of data collection. A linked survey permits more detailed measurement of child labour than is normally possible in a modular survey. Also, more elaborate sub-sampling from the base survey, both at the area and the household levels, is possible. But of course there is the extra cost of separate operations.

As to the sampling aspects of such linkages, one extreme option is to draw the CLS sample as a *sub-sample of the individual children enumerated* in the base survey. At the other extreme, the two surveys may be based on *independent samples*. There are a number of intermediate possibilities.

Even when the two surveys are based on independent samples it is desirable and efficient to draw them from a *common frame or 'master sample' of area units*. This permits the sharing of costs of preparation and maintenance of the area frame.

A closer link between the surveys is obtained by basing them on a *common sample of areas*. In principle, all the areas in the base survey may be included for the CLS sample. However, sub-sampling of the LFS areas is often desirable and appropriate: the sample sizes required for a CLS tend to be much smaller than those for major surveys such as the LFS.

Within the common sample areas, various possibilities exist in terms of the relationship between the ultimate units (e.g. households) in the two samples – from entirely independent samples from household lists in the common areas, to confining the CLS to a sample of children actually identified during the LFS information on household composition.

(1) At the one extreme, the common LFS-CLS sample areas may be 're-listed' to obtain a more up-to-date frame of households for the CLS, and an entirely new sample of the units selected. However, creation or updating of household lists can be an expensive operation, and is justified only if the two surveys are separated by, say, one year or longer.

(2) When the same lists are used, the two samples may still be selected independently or without overlap. This is desirable when respondent fatigue is a concern, or when the first sample is subject to high rates of non-response. Independent or at least additional sampling is also required when the first survey is not able to yield a sufficiently large sample for the CLS.

An alternative is to base the CLS on all or a sub-sample of ultimate units included in the first sample. The outcome depends on the type and characteristics of the units involved.

(3) Selecting a sub-sample of *addresses* in the base sample is often the simplest.

(4) The use of *households* as units for sub-sampling comes next. It is simpler if households from the first survey are subject to sub-sampling without reference to any particular characteristics of the households involved.

(5) However, sometimes information on various *household characteristics* is evoked for the purpose of stratification or for applying different sampling rates. This involves the collection of such information in the base survey, and its preservation and transfer to the CLS. This can be expensive and cumbersome, and in many cases not very effective in improving the efficiency of the resulting sample. It is also common to exclude certain types of households from the selection, such as households not found to contain any child relevant to the CLS. This can improve control over the CLS sample size, and also efficiency of its fieldwork. However, the drawback is the assumption that the situation of the households with respect to the exclusion criteria has not changed during the interval between the two surveys.

(6) Using *children* identified during the first survey as the units for sampling for the CLS is also an option. However, this is a demanding choice in that lists of children identified have to be prepared, transferred to the CLS

operations for sampling, and then the selected children identified during fieldwork. Misidentification of individual children can easily occur. Such an option should be followed only if the interval between the two surveys is very short.

(7) At the extreme is the procedure where information on various *characteristics of children* is used for the purpose of stratification or for applying different sampling rates, such as the child's educational and/or activity status. Such a procedure may appear attractive when the CLS sample size is very small and its structure needs to be tightly controlled. However, generally this is a demanding and expensive procedure, prone to implementation errors. It should be used only when the CLS approximates the conditions of being a 'module' of the first survey – close in timing the first survey, and drawing significantly on the substantive information collected in it.

3.3.CLS based on a household listing operation

A child labour survey may be based on a household listing operation required to create or update the sampling frame. (In an area-based sample, such a listing operation is normally confined to sample areas selected at the preceding stages.) At least two options are possible. The child labour survey may be a separate *stand-alone* survey, conducted subsequently to a household listing operation, and not linked to another survey such as the LFS. Sub-sampling of households or similar units from the lists will be normally required within each sample area. However, a stand-alone CLS is not always a feasible or even a desirable option.

An alternative is *integration* of the CLS with household listing in the form of a single operation. This necessitates that the child labour component involves no more than a few brief questions which can be incorporated into the listing form. In an area-based sample, the listing operation normally involves exhaustive coverage of each sample area, i.e. the listing of all households or similar units in it without involving any sub-sampling. The same applies in the case of the CLS where it is operationally integrated with household listing. The resulting data may have the advantage of being based on a large sample, but the measurement of child labour is likely to be approximate. This type of CLS is essentially no more than a *screening operation* for a more detailed survey of labouring children.

4. CLS-LCS linkage

A labouring children survey (LCS) requires a prior operation which identifies, explicitly or implicitly, a sample of working children. Normally this is the function of the CLS component. Two forms of the relationship between CLS and LCS may be identified: (i) an integrated CLS-LCS operation; and (ii) linked CLS and LCS operations.

4.1. Integrated CLS-LCS operation

The distinction, discussed in Section 2, between the two types of surveys, the CLS and the LCS, by no means implies that they must (or sometimes even can) be organised as two separate operations. In fact, in a majority of the national surveys conducted so far they have been completely integrated into a single survey operation – the same survey covering the two different types of objectives. The LCS questions form additions to the CLS questionnaire, which become applicable if the child concerned is found to be engaged in work-related activity.

This is an *integrated design*, by which is meant that the information for the CLS and LCS components is collected as a single interview operation. Note that a ‘single interview operation’ is not meant to necessarily imply that only a single integrated questionnaire is used, or that all interviewing takes place during a single visit to the household, or even that all the information is obtained from a single respondent in the household. Multiple questionnaires, repeated interviewer visits, and different respondents (head of household, parents or guardians, children themselves, and sometimes even employers and teachers etc.) may indeed be involved. Rather, the term is meant to indicate that the information on the CLS and LCS components is collected at the same time or at least within a short time of each other, and that all the information collected during the CLS component is directly available to (and is generally not repeated in) the LCS component.

Simultaneous implementation of the CLS and LCS components implies that normally no sub-sampling from the one to the other can be introduced (i.e., all children identified as working during the CLS part are subject to the LCS part of the interview). In any case, it is desirable in practice that any sub-sampling involved is straightforward, such as applying the LCS part of the interview to only a pre-selected subset of CLS survey rounds or sample areas.

In an integrated survey, the sample size requirements of both the CLS and LCS components have to be met. Hence even in a single integrated survey, the distinction between the two types of survey components still remains a conceptually useful one: it reminds us that an integrated design has to be a compromise between different objectives. It comes to requiring that the CLS is large enough to provide estimates of prevalence of child labour with the required precision, and also to yield sufficient numbers of working children for the LCS.

Unfortunately in national practices, often the survey design has been determined one-sidedly – with over-emphasis on one type of objective at the expense of the other. For instance, some surveys have been too small to yield useful estimates of the extent and distribution of prevailing child labour (the CLS component was too small in sample size), while others have been too large in size to permit sufficiently in-depth investigation of the characteristics and consequences of child labour (LCS component was too large in sample size from the practical point of view). By contrast, there are also examples where the sample size, while adequate for the CLS, turned out to be inadequate in providing enough cases for the LCS component for the purpose of investigating child labour activities in detail.

It is worth emphasising that with an integrated design, the sample for the LCS component is determined entirely by the results of the screening provided by the CLS component: not only *sample size* but also – and more importantly – the *quality of coverage* of the population of working children in the LCS component is determined by the quality of the screening questions in the CLS part for identifying those children.

Advantages and disadvantages of an integrated design

There can be practical and cost advantages in integrating the CLS and LCS components into a single operation. Clearly, it is cheaper and convenient to collect all the required information in one go. This is a major advantage, and may explain why a large majority of child labour surveys to-date have chosen the integrated arrangement.

However, there are also disadvantages of an integrated (CLS+LCS) design.

(1) An integrated implementation tends to limit the information which can be collected during the LCS component without jeopardising the quality of the CLS component.

(2) The increased burden associated with the LCS part can have serious consequences for the quality of the CLS part in identifying the incidence of child labour.

(3) Apart from the obvious upper limit imposed by the number of eligible cases (working children) identified in the CLS component, there is no independent control over the size and distribution of the sample for which the LCS information is collected.

(4) The level of child labour may be very unevenly distributed over sample areas, thus greatly varying interview workload. This becomes all the more troublesome when the LCS part involves lengthy interviews with adults as well as with children individually.

When an integrated CLS-LCS operation may be suitable

An integrated CLS-LCS operation may well be the most suitable option under certain conditions such as the following.

(1) The CLS does not require a very large sample, which would be the case when it is not required to produce estimates of the prevalence of child labour for many different regions, population groups, sectors of activity, or other types of domains.

(2) The CLS is a stand-alone survey, so that its sample can be designed as a compromise for meeting both types of information needs – of estimating the prevalence of child labour with necessary precision on the one hand, and of investigating the conditions and consequences of child labour with necessary detail on the other.

(3) Child labour is not too heterogeneous or extremely unevenly distributed for it to be 'captured' in a reasonable way by a general purpose sample of the population of children.

(4) The resulting compromise sample size is not too large for the in-depth investigation which the LCS component typically requires.

(5) And in any case, the resulting integrated interview is not too heavy to have an adverse effect on the quality (particularly completeness) of measuring the prevalence of child labour which is the concern of the CLS component. This is a common problem which has been encountered in many other types of surveys with similarly dual objectives.

When one or more of the above conditions are violated, it is necessary to at least consider the possibility of operationally separating the CLS and LCS components. In the case of such a separation, it would be generally appropriate to consider basing the LCS component on a subsample of the CLS. The objectives of the sub-sampling would be both to reduce the sample size for the LCS, and also to make it more concentrated and targeted to reflect the uneven geographical distribution of child labour. Specific subsampling procedures for this purpose are discussed in Section 7 below.

4.2.LCS linked to the CLS

An alternative to the integrated design is to conduct the CLS and LCS components as separate operations. However, these two cannot be stand-alone (i.e. entirely separated) surveys, but must be *linked* to each other in some way. The LCS sample can be identified on the basis of the CLS results in different ways, and this provides different forms of linkages between the two surveys. The best solution depends on the particular situation and objectives, but primarily on two factors: (i) the time gap between the two surveys; (ii) the quality of screening provided by the CLS in identifying the presence of working children.

The diversity of options in CLS-LCS linkage include the following.

A. In relation to selection of the sample within the CLS sample areas, we may take all or select a sample of:

(1) previously identified working children individually;

(2) households identified previously to contain a working child;

(3) households identified previously to contain any child in the age range of interest;

(4) all households interviewed in the previous sample;

(5) or possibly, all households selected in the previous sample (including non-respondents);

(6) all households listed in the CLS sample areas (i.e., obtaining a new LCS sample from existing CLS household lists); or

(7) households from re-listing of the CLS sample areas (i.e., updating the area lists before selecting a new sample).

B. In relation to 'eligibility for inclusion' of the CLS areas, we may take as eligible:

(8) CLS sample areas containing at least $x \geq 1$ working child(ren), or containing at least $y \geq 1$ household(s) with a working child;

(9) CLS sample areas containing at least $x \geq 1$ child(ren) or at least $y \geq 1$ household(s) with a child in the age range of interest; or

(10) all CLS sample areas.

C. In relation to the selection of areas, we may:

(11) take a sub-sample of the CLS sample areas;

(12) take all CLS sample areas; or

(13) select additional areas for LCS, linked to the CLS sample areas.

Some of the options in the three sets A–C can be combined, for instance: households identified previously to contain any child in the age range of interest (option 3), but only from a sub-sample of areas (option 11), those areas selected from CLS sample areas which contain at least one working child (option 8).

Note that among (1)–(7), all options except (1) require fresh identification of labouring children in the households included in the LCS sample. Option (1) provides the tightest link between the two surveys: the LCS sample is confined to the particular children identified to be engaged in child labour during the CLS. This makes the option almost the same as an integrated CLS-LCS design, except for the possibility of sub-sampling between the two operations because of their operational separation in time. With this option, any working children unidentified during CLS remain unidentified during the subsequent LCS operation. The opposite error – non-working children identified as working – is less important and in any case can be identified during the subsequent operation. With options (2)–(7), the LCS has a greater potential to refine the CLS estimates of child labour.

The categories from (13) to (1) represent increasingly close linkage between the two surveys. Going upwards from (13) to (1), especially from (7) to (1), the options generally become: more restrictive (the LCS sample is increasingly restricted to units identified in the CLS sample); more focussed (on the particular population of interest, i.e. labouring children); and more efficient exploiter of information already collected or the design already implemented in the CLS (hence more cost-effective, and less burdensome for the respondent).

But the options increasingly require: more information to be collected during the CLS and then fed-forward to the LCS (hence the operation becoming more costly and time consuming); and also more sensitive to changes over time (hence less suitable in the presence of a long time gaps between the two surveys).

Exclusion of certain CLS areas

Information on the reported level of child labour in the CLS is often useful in determining the cut-off level below which the areas may be altogether excluded from selection into the LCS. This is often necessary for practical reasons – it may simply not be cost-effective to attempt the LCS in areas containing none or very few reported working children. The cost considerations have to be balanced against the bias which such exclusion introduces.

Sub-sampling of CLS areas

Sub-sampling of CLS sample areas may be introduced in order to reduce the LCS sample size, and specifically to make the LCS sample more concentrated, i.e. confined to fewer sample areas with higher levels of child labour.

Expanding the original CLS areas

A brief comment will be useful on option (13). This option may be used when the LCS sample needs to include additional areas, beyond the CLS areas. This may, for instance, be because the CLS sample areas do not yield a sufficient number of sample cases (labouring children) for the LCS. Another motivation can be to enhance the LCS sample by selecting more cases from and around CLS areas found to contain concentrations of labouring children. An obvious way to expand the LCS sample is to include in it additional areas in the neighbourhood of CLS sample areas. Various techniques (such as ‘adaptive cluster sampling’, see Tompson and Seber, 1996) can be used for expanding the sample in this way, while retaining its probability nature.

5. Examples of diverse structures of child labour surveys

Tables 1 and 2 provide some essential information on the samples and the structure of linkages between different components for around 30 surveys on child labour. The tables show the diversity of the survey structure encountered. The survey-structural concepts have been explained in the preceding sections.

5.1.Linkage of CLS to a base survey

Table 1 classifies the type of linkage between the CLS and its base, which may be a larger survey (such as FLS) or the household listing operation. It can be seen that half the surveys (15 of the surveys reviewed) are *stand-alone surveys*, meaning that the survey is exclusively or primarily concerned with child labour.

The other common arrangement (13 of the surveys reviewed) is an *integrated survey,* involving the collection of the base survey and CLS information during the same operation. Here we distinguish modular versus combined surveys in the sense described in Section 3. The LFS forms the base for most of the combined surveys. It is also the case for several modular CLS's, but a variety of other types of surveys have also served as the base as shown in Table 1.

Another possibility, but a rare one, is to have a *linked survey,* where the child labour survey interview is operationally separated from that of the base survey, but the sample and some substantive information is fed-forward from the base survey.

The table also shows that most commonly CLS is a single round, one-time survey, though in one of six cases it has involved multiple rounds, typically four rounds corresponding to quarters of the year. A multi-round CLS is not affordable in most circumstances. Other information shown concerns sample size (n households) and its division into number of clusters (a) and sample-take per cluster (b=n/a). Sample size varies greatly, from 6,000 to 48,000 in the cases reviewed. Sample-take per cluster is even more variable – mostly in the range 5–50.

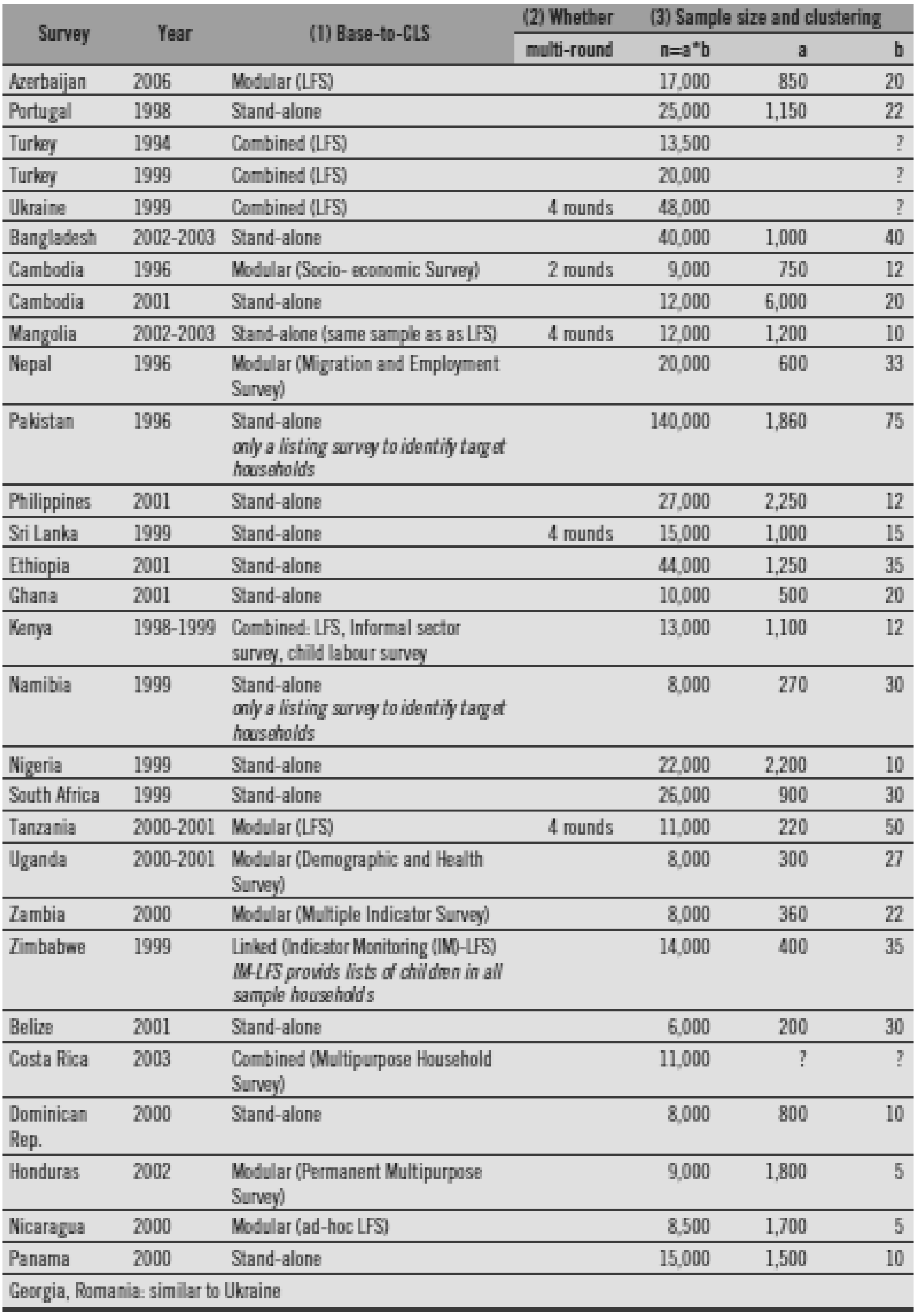
5.2.Linkage between CLS and LCS

Table 2 classifies the type of linkage between CLS and LCS components. It also indicates the substantive scope of the LCS – mainly whether it concerns primarily child labour (i.e. working children) or is a more inclusive survey of child activities or of children generally.

By far the predominant form has been an *integrated* CLS-LCS operation, in which information on prevalence of child labour and more detailed information on children who are found to be working are both collected during a single interview operation. A *linked* survey structure which permit a degree of operational separation between CLS and LCS has been used in only one-in-five of the surveys reviewed in the table.

Concerning the substantive scope of the LCS (irrespective of whether it is integrated or merely linked to the CLS), a majority are concerned primarily with economic activity of working children. However, almost one-half are broader in scope. A number are *child activity surveys* covering all types of activities of children including non-economic activities; some are even broader *children's surveys* covering in addition other areas such as children's health, housing and living conditions. This broader scope requires the coverage of a broader population – essentially of all children as in the case of the CLS.

Table 1. Child Labour Surveys: Base-to-CLS. Examples of diverse structures

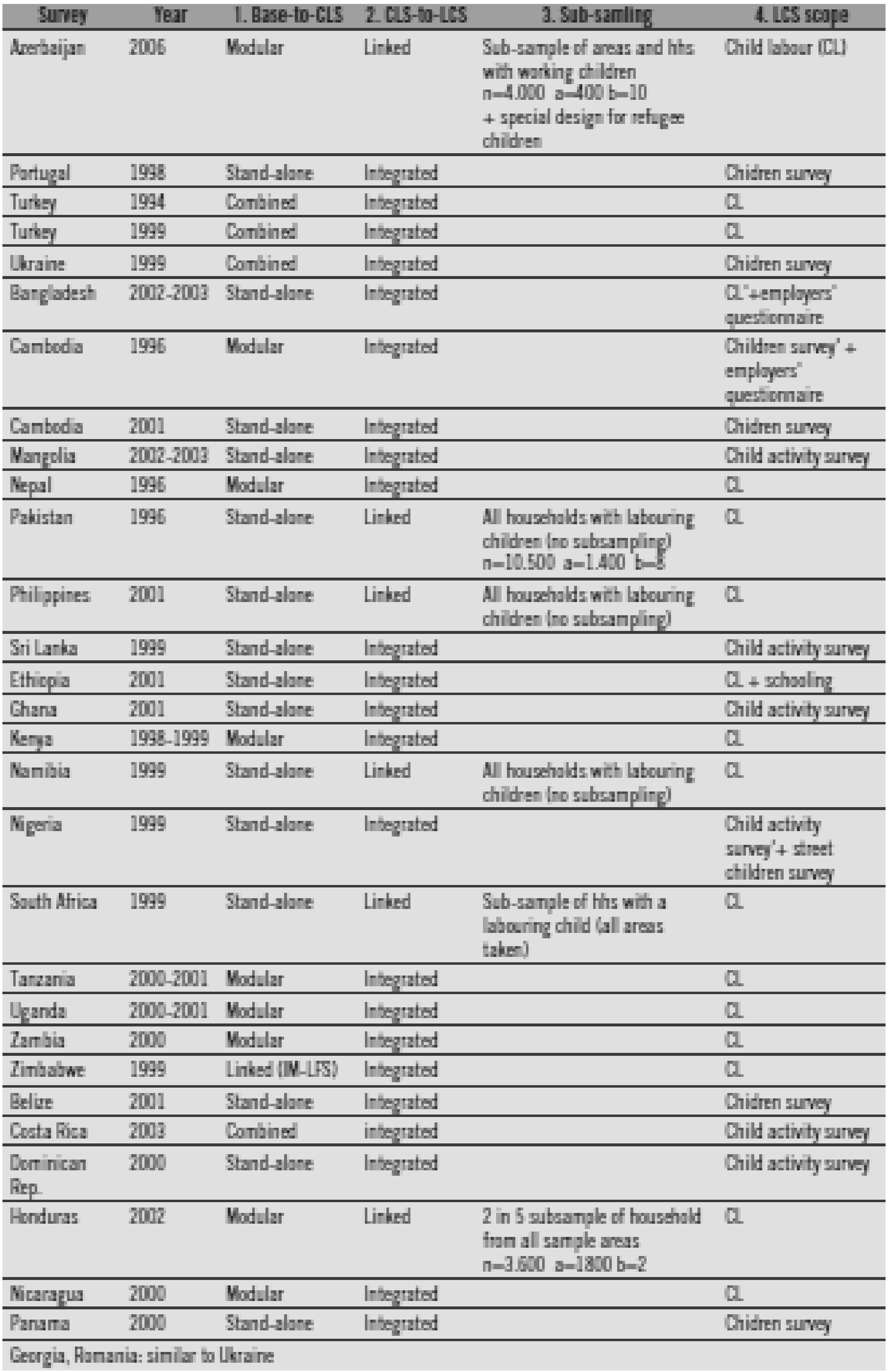


Survey	Year	(1) Base-to-CLS	(2) Whether	(3) Sample size and clustering		
			multi-round	n=a*b	a	b
Azerbaijan	2006	Modular (LFS)		17,000	850	20
Portugal	1998	Stand-alone		25,000	1,150	22
Turkey	1994	Combined (LFS)		13,500		?
Turkey	1999	Combined (LFS)		20,000		?
Ukraine	1999	Combined (LFS)	4 rounds	48,000		?
Bangladesh	2002-2003	Stand-alone		40,000	1,000	40
Cambodia	1996	Modular (Socio- economic Survey)	2 rounds	9,000	750	12
Cambodia	2001	Stand-alone		12,000	6,000	20
Mangolia	2002-2003	Stand-alone (same sample as as LFS)	4 rounds	12,000	1,200	10
Nepal	1996	Modular (Migration and Employment Survey)		20,000	600	33
Pakistan	1996	Stand-alone *only a listing survey to identify target households*		140,000	1,860	75
Philippines	2001	Stand-alone		27,000	2,250	12
Sri Lanka	1999	Stand-alone	4 rounds	15,000	1,000	15
Ethiopia	2001	Stand-alone		44,000	1,250	35
Ghana	2001	Stand-alone		10,000	500	20
Kenya	1998-1999	Combined: LFS, Informal sector survey, child labour survey		13,000	1,100	12
Namibia	1999	Stand-alone *only a listing survey to identify target households*		8,000	270	30
Nigeria	1999	Stand-alone		22,000	2,200	10
South Africa	1999	Stand-alone		26,000	900	30
Tanzania	2000-2001	Modular (LFS)	4 rounds	11,000	220	50
Uganda	2000-2001	Modular (Demographic and Health Survey)		8,000	300	27
Zambia	2000	Modular (Multiple Indicator Survey)		8,000	360	22
Zimbabwe	1999	Linked (Indicator Monitoring (IM)-LFS) *IM-LFS provids lists of children in all sample households*		14,000	400	35
Belize	2001	Stand-alone		6,000	200	30
Costa Rica	2003	Combined (Multipurpose Household Survey)		11,000	?	?
Dominican Rep.	2000	Stand-alone		8,000	800	10
Honduras	2002	Modular (Permanent Multipurpose Survey)		9,000	1,800	5
Nicaragua	2000	Modular (ad-hoc LFS)		8,500	1,700	5
Panama	2000	Stand-alone		15,000	1,500	10
Georgia, Romania: similar to Ukraine						

An '?' indicates information not available from published survey report or other documentation.

***Source**: Compiled from national reports on surveys of child labour.*

Table 2. Child Labour Surveys: CLS-to-LCS. Examples of diverse structures



Survey	Year	1. Base-to-CLS	2. CLS-to-LCS	3. Sub-samling	4. LCS scope
Azerbaijan	2006	Modular	Linked	Sub-sample of areas and hhs with working children n=4.000 a=400 b=10 + special design for refugee children	Child labour (CL)
Portugal	1998	Stand-alone	Integrated		Chidren survey
Turkey	1994	Combined	Integrated		CL
Turkey	1999	Combined	Integrated		CL
Ukraine	1999	Combined	Integrated		Chidren survey
Bangladesh	2002-2003	Stand-alone	Integrated		CL"+employers' questionnaire
Cambodia	1996	Modular	Integrated		Children survey' + employers' questionnaire
Cambodia	2001	Stand-alone	Integrated		Chidren survey
Mangolia	2002-2003	Stand-alone	Integrated		Child activity survey
Nepal	1996	Modular	Integrated		CL
Pakistan	1996	Stand-alone	Linked	All households with labouring children (no subsampling) n=10.500 a=1.400 b=8	CL
Philippines	2001	Stand-alone	Linked	All households with labouring children (no subsampling)	CL
Sri Lanka	1999	Stand-alone	Integrated		Child activity survey
Ethiopia	2001	Stand-alone	Integrated		CL + schooling
Ghana	2001	Stand-alone	Integrated		Child activity survey
Kenya	1998-1999	Modular	Integrated		CL
Namibia	1999	Stand-alone	Linked	All households with labouring children (no subsampling)	CL
Nigeria	1999	Stand-alone	Integrated		Child activity survey'+ street children survey
South Africa	1999	Stand-alone	Linked	Sub-sample of hhs with a labouring child (all areas taken)	CL
Tanzania	2000-2001	Modular	Integrated		CL
Uganda	2000-2001	Modular	Integrated		CL
Zambia	2000	Modular	Integrated		CL
Zimbabwe	1999	Linked (IM-LFS)	Integrated		CL
Belize	2001	Stand-alone	Integrated		Chidren survey
Costa Rica	2003	Combined	integrated		Child activity survey
Dominican Rep.	2000	Stand-alone	Integrated		Child activity survey
Honduras	2002	Modular	Linked	2 in 5 subsample of household from all sample areas n=3.600 a=1800 b=2	CL
Nicaragua	2000	Modular	Integrated		CL
Panama	2000	Stand-alone	Integrated		Chidren survey
Georgia, Romania: similar to Ukraine					

***Source**: Compiled from national reports on surveys of child labour.*

6. Sample selection for a child labour survey (CLS)

This section describes some technical procedures for drawing the sample for a child labour survey (CLS) on the basis of the sample used for a larger survey of the general population, in particular the LFS.

6.1.The base survey (LFS)

In order to facilitate concrete discussion, we will assume the following design for the base survey. This is by far the most commonly used procedure for selecting population-based samples, especially in developing countries. It involves the selection of area units in one or more stages (often in only one stage) with probability proportional to a measure of population size of the area (p_i), and within each selected area, the selection of ultimate units with probability inversely proportional to the size measure. Below, summation Σ is over all areas in the population. Parameter 'a' refers to the number of areas (strictly 'ultimate area units') selected. If the current size of the area exactly equals the size measure p_i used for its selection, then f is the constant selection probability for any ultimate unit, parameter b is the constant number of units selected from any sample area, and $n = a \cdot b$ is the resulting sample size.

Selection probability for an area unit

$$f_{1i} = \left(\frac{a}{\Sigma p_i} \right) p_i = \frac{p_i}{I}, \text{ say} \tag{1}$$

Selection of ultimate units within selected area unit

$$f_{2i} = \left(\frac{b}{p_i} \right) \tag{2}$$

Overall selection probability of an ultimate unit

$$f_i = f_{1i} \cdot f_{2i} = \left(\frac{b}{I} \right) = f \text{, a constant.} \tag{3}$$

We will also assume the commonly used procedure of selecting area units systematically with probability proportional to size measure p_i from a list ordered in some meaningful way (PPS sampling). The systematic selection interval is $I = \Sigma p_i / a$, as defined above.

Dealing with very large and very small units

Special treatment is required in the selection of units of extreme (very large or very small) size. ‘Very large’ in the context of PPS sampling means a unit whose size measure exceeds the sampling interval, i.e. $p_i > I$. Such units may be segmented (divided into smaller areas) such that no segment exceeds I in size. (The segmentation may be applied to all units in the frame prior to sample selection, or only to the selected units using some objective rule not dependent on which particular units happen to be selected.) An alternative (and generally recommended) procedure is to treat large units as automatically selected. In this case, the selection equations become $f_{1i} = 1; \quad f_{2i} = f$, the required overall constant rate. The sample size from the area is then proportional to its current size, $b_i = f \cdot p_i$.

‘Very small’ in the context of PPS sampling means a unit whose size measure is smaller than the required sample-take, i.e. $p_i < b$. Small units in the sampling frame may be grouped together (merged to form larger areas) such that no group is smaller than the required sample-take b. (As above, the grouping may be applied to all units in the frame prior to sample selection, or only to the selected units using some objective rule independent of which particular units happen to be selected.) Two commonly used alternative procedures are the following.

(1) Assigning small units a minimum size measure, $p_i = b$ for the area selection, and taking into the sample all final stage units in each selected area. The selection equations become $f_{1i} = (b/I) = f, f_{2i} = 1$. The original ultimate selection probabilities are retained unchanged, but, for a given sample size, the number of area units in the sample is increased. This can be inconvenient and expensive if there are too many small units in the frame.

(2) Excluding the smallest among small areas, with appropriate compensation. The following method has proved useful in a number of child labour surveys; it is particularly useful when the population includes many units, each of them with only a few (or even no) ultimate units of interest – as may well be the case in child labour surveys. In outline, the procedure is as follows. Let Σp_i be the total size measure of the set of ‘very small’ units to be sampled. The set of small units is ordered by unit size and divided into two parts. The first part is defined to consist of $A = \Sigma p_i / b$ *largest* units in this small-unit set. The size measure of each unit in this subset is increased to b, so that selection with interval I gives a sample of $a = A \cdot b/I = \Sigma p_i / I$ area units. All ultimate units in each selected area are retained in the sample. The second part consists of the remaining *smallest* of the small-unit set, with size measures $p_i < b_0$, say. For reasons of cost and practicality, these units are altogether excluded

from the sample, even though this is not properly a probability sampling procedure. It can be shown that a compensation for this exclusion is obtained by increasing the weight given to each selected unit in the first part by the factor (b/p_i), where p_i is the original size measure of the unit concerned.

6.2. Common structure but different design parameters between LFS and CLS

While the size measure in the LFS is the general population (or the population in the working ages), for a CLS it is an appropriately defined population of children exposed to the risk of child labour. It is generally the case that these two populations are closely related in size – the average difference between them being primarily a scaling factor. As noted, these similarities in the basic structure of the samples have important implications in the choice of appropriate survey structure for the CLS.

The procedure for selecting sample areas can also be the same in so far as the relevant base population sizes for the two surveys are nearly proportional to each other. In particular, if selection with probability proportional to size p_i is suitable for the LFS, it is reasonable to assume that, for practical purposes, the *same* size measures p_i as used in the LFS for PPS selection of areas are also appropriate for the same purpose in the CLS.

Despite similarity in the structure and distribution of the population to be sampled, the CLS and LFS will typically differ in a number of design requirements and parameters. In many countries, the LFS is a well established, regular or even a continuous survey. The CLS is more likely to be a new or recently instituted survey, conducted at best periodically and often only on an *ad hoc* basis. The resources available for the CLS are likely to be more limited, and often less certain. Increasingly, the LFS is required to produce separate estimates for different regions and sub-populations in the country, and produce these more frequently such as annually or even quarterly, while in most cases the primary objective of the CLS still has to be, at the first instance, the production of national-level estimates from time to time. In short, the LFS may be seen as a large, extensive and regular survey, and by comparison the CLS as a smaller, more intensive and less frequent survey. Consequently, the two types of surveys differ in relation to the choice of design parameters such as survey timing and frequency, sample size, the number of areas selected for the sample and the related sample size per area, allocation of the sample across different domains in view of different reporting requirements in the two surveys, the details of the stratification, and in the ultimate units for which data are collected.

6.3. Selection of CLS sample from a base survey such as the LFS

This subsection considers the procedure for selecting the sample for CLS from the sample for a base survey, typically the LFS.

One extremely important practical point should be noted at the outset when a sample is obtained by sub-sampling from an existing sample: full details must be recorded not only of the sub-sampling procedures, but must also be available for the existing sample used for sub-sampling. Unfortunately one finds examples of surveys in which details on the design of the existing sample were either not properly documented or had not been preserved. The most critical piece of information concerns probabilities of selection applied in the original sample. If such details are not available for an existing sample, it is desirable to look for alternative sources for sub-sampling.

Basic procedure

We begin with the simplest situations: for a given domain, a reduced number of areas are to be retained for the CLS from a given sample of LFS areas. Assume that the LFS is based on the commonly used PPS design described in Section 6.1. Let this sample contain 'a' areas selected with probability proportional to a measure of area population size p_i. The objective is to select for the CLS a reduced number of areas, say $a' = g\,a,\ g < 1$, also with probability proportional to the *same* measure of size p_i. Given the common size measures, the procedure for sampling of areas from LFS to CLS is straightforward: select a sub-sample of LFS areas with a *constant probability* $g = a'/a$. Selection with a constant probability g can be achieved simply by applying to the LFS sample areas the *equal probability* systematic sampling procedure with interval $k = 1/g$. The result for the CLS is a PPS sample of areas with probability proportional to the same population size measure p_i. The selection equations for the LFS and the CLS areas are:

LFS: $f_1 = \left(\frac{a}{\Sigma p_i}\right).p_i = \frac{p_i}{I}$, say, and CLS: $f_1' = \left(\frac{a'}{\Sigma p_i}\right).p_i = \left(\frac{a/k}{\Sigma p_i}\right).p_i = \left(\frac{p_i}{k\,I}\right)$,

where Σp_i is the sum of size measures for all areas in the population from which the LFS sample was selected.

This simple procedure involving subsampling of areas at a constant rate applies to any arbitrary choice of size measures p_i provided that they are the same in the two samples.

Dealing with 'very large' areas in sub-sampling

Additional steps are involved in the sub-sampling procedure in dealing with areas in the original sample which were selected using special procedures because they were considered to be too ‘small’ or too ‘large’ for the normal PPS procedure, in the sense described earlier.

We first consider very large units, i.e. units with $p_i > I$. Assume that in the LFS such areas have been selected with $f_{1i} = 1; \quad f_{2i} = f$ (see Section 6.1). For the purpose of selecting a $(1/k)$ sub-sample of these ‘self-representing’ LFS areas for the CLS, two groups among these need to be distinguished.

Group 1: $p_i > (k\, I)$. These are the largest units. All of these units must be retained in the CLS with probability =1, as in the LFS. At the final stage, ultimate units can be selected with the overall selection rate, say f', required for the CLS.

Group 2: $I \le p_i \le (k\, I)$. All these large units do not get selected into the CLS automatically, though that was the case in the LFS. For these units, a proper PPS sample of areas can be selected for the CLS with $f'_{1i} = (p_i / k\, I) \le 1$. For a self-weighting design with overall sampling rate f', the final stage selection probabilities would be $f'_{2i} = (f'/f'_{1i})$.

Dealing with ‘very small’ areas in sub-sampling

The sub-sampling procedure becomes more complex when dealing with very small units. In a two-stage design, the procedure for handling very small areas in LFS-to-CLS sub-sampling depends on the details of the LFS sampling at both stages. For describing the sub-sampling procedure, we assume the LFS sample selected according to equations (1)–(3).

For the CLS, sub-sampling from the LFS involves two steps: selection of 1 in k sample areas, $a' = (a/k)$, and then (an average or target) selection of an expected number b' of ultimate units per area included in the CLS, giving $n' = a'\, b'$ as the CLS target sample size. Note that the two (LFS and CLS) samples may not necessarily overlap as concerns the ultimate units, though often they do and $b' \le b$. Also, both for the LFS and the CLS, these parameters may differ from one sampling domain or stratum to another. It is sufficient here to describe the procedure for one such domain.

The full selection equations for self-weighting CLS design, corresponding to (1)–(3), are:

Selection probability for an area unit

$$f'_{1i} = \left(\frac{a'}{\Sigma p_i} \right) p_i = \frac{p_i}{k\, I} = \frac{p_i}{I'}, \text{ say} \tag{4}$$

Selection of ultimate units within selected area

$$f'_{2i} = \left(\frac{b'}{p_i} \right) \tag{5}$$

Overall selection probability of an ultimate unit

$$f_i' = f_{1i}' \, f_{2i}' = \left(\frac{b'}{I'}\right) = \left(\frac{1}{k}\,\frac{b'}{b}\right) f = f', \text{ a constant.} \tag{6}$$

The procedure for selecting for the CLS a sub-sample of very small LFS areas, i.e. areas with size measures $p_i < b$, depends on how these areas were selected into the LFS sample itself. In addition, we have also to consider the relationship of the size measure p_i to the required sample-take b' in the CLS. Table 3 shows the selection equations for very small areas in CLS, corresponding to the two procedures for their treatment in the LFS described in Section 6.1, namely:

Procedure (1). Assigning small units a minimum size measure, $p_i = b$ for the area selection, and taking into the sample all final units in each selected area.

Procedure (2). Excluding smallest of small areas, with appropriate compensation.

Table 3. Sub-sampling of LFS areas for the CLS

	Condition	LFS		CLS		$\frac{f_1'}{f_1}$
		f_1	f_2	f_1'	f_2'	
	Very large areas					
1	$p_i > (k\ I)$	1	b/I	1	$b'/k\ I$	1
2	$I \le p_i \le (k\ I)$	1	b/I	$p_i/k\ I$	b'/p_i	$p_i/k\ I$
	Normal areas (majority of the areas)					
3	$b, b' \le p_i < I$	p_i/I	b/p_i	$p_i/k\ I$	b'/p_i	$1/k$
	Very small areas – area selected in LFS using Procedure (1)					
4 *	$b \le p_i < b'$	p_i/I	b/p_i	$b'/k\ I$	1	$b'/k\ p_i$
5	$b' \le p_i < b$	b/I	1	$p_i/k\ I$	b'/p_i	$p_i/k\ b$
6	$p_i \le b, b'$	b/I	1	$b'/k\ I$	1	$b'/k\ b$
	Very small areas – difference from above if selected in LFS using Procedure (2)					
5	$b_0, b' \le p_i < b$	as in case 5 above				$1/k$
6	$b_0 \le p_i \le b, b'$	as in case 6 above				$b'/k\ p_i$
7	$p_i < b_0$	areas not included in LFS or CLS sample				--

** Note that only one of the two cases, 4 and 5, can apply in any particular situation. See Section 6.1 last paragraph, for definition of b_0: units with size smaller than this limit were dropped from the LFS.*

6.4. Sample allocation and reporting domains

Apart from differences in the required sample size and clustering, the CLS may differ from the 'parent' LFS also in the requirements concerning sample allocation and stratification.

For instance, the LFS may be allocated disproportionately for the purpose of producing sub-national estimates with over-sampling of small regions or other reporting domains, while this may not be required for a CLS when it is based on a smaller sample aimed primarily at producing national-level estimates, or producing breakdowns only for a few major domains.

In any case, we can generally expect the CLS to have a smaller number of reporting domains than a bigger survey like the LFS; furthermore, the sampling rates in the CLS sample are often more uniform. Typically, the reporting domains for the CLS would be groupings of the LFS reporting domains – e.g. major regions of the country rather than individual provinces or districts. It is unlikely to have the situation in which the more detailed LFS domains cut across boundaries of the more aggregated CLS domains.

Significant differences in the required stratification are unlikely. Certain common stratification criteria, such as geographic location and degree of urbanisation, are commonly used in all types of household surveys. Often these are practically the only stratification criteria available. Where available, additional useful criteria (ethnicity, predominating occupation, literacy rate, mean level of income, etc., for the sample area) also tend to be similar for different social surveys. The CLS and LFS are likely to be even more similar in terms of stratification because of their shared or similar subject matter. Generally, differences in stratification requirements are likely to arise only in relation to differing requirements between the two surveys in terms of sample allocation.

7. Sample selection for labouring children survey (LCS)

7.1. The required sample structure

As noted, a critical issue of practical importance is whether the distinct CLS and LCS objectives can be satisfactorily met through a single integrated survey, or it is better to organise them as two separate – but nevertheless linked – operations. In the latter case, a related question is whether the two components can be based on the same sample of units, or the LCS should be a subsample of the CLS, smaller in size and possibly also with a different structure.

The main difference between the CLS and LCS concerning the sample structure is that in the former the primary focus was on the measurement of prevalence of child labour among the population of all children, and consequently that population formed the base for the design and selection of the CLS sample. The base for the LCS is the population of working children. Consequently, the selection of LCS sample areas requires information on the number of labouring children in each area in the 'frame' from which the survey areas are to be selected. Such information is not normally available in general-purpose population-based frames. It is this reason that makes it necessary to select the LCS as a sub-sample of areas for which such information has been collected, such as sample areas from the larger CLS.

This section considers the procedure for selecting the sample for LCS from the sample for CLS as the base. For the development and exposition of the CLS-to-LCS sampling procedure, we will assume the basic sampling scheme for the CLS to be equations (4)–(6); actually for notational simplicity, we will use their equivalent, equations (1)–(3).

In a LCS, where the base population of interest is labouring children, an appropriate design will involve the selection of area units with *probability proportional to the number of labouring children* (c_i) in the area, and then the selection, within each selected area, of such children with probability inversely proportional to c_i:

Selection probability for an area unit

$$f'_{1i} = \left(\frac{a'}{\Sigma c_i}\right) c_i = \frac{c_i}{I'}, \text{ say} \tag{7}$$

Selection of ultimate units within selected area

$$f'_{2i} = \left(\frac{b'}{c_i}\right) \tag{8}$$

Overall selection probability of an ultimate unit

$$f'_i = f'_{1i}\ f'_{2i} = \left(\frac{b'}{I'}\right) = f', \text{ a constant.} \tag{9}$$

This design differs from that of the CLS discussed earlier in a number of respects.

(1) The LCS design depends on the categories of children included in the target population – for example the definition of child work or labour used, or the specific types of child activities included. The measure of size c_i refers to the numbers of children in the categories of interest. These values cannot be assumed known for all areas in the population. It is assumed that these are obtained or estimated from a base survey (LFS, or more appropriately a CLS), but only for the areas enumerated in that survey. Hence the LCS must be confined to a sub-sample of areas in the first survey. Within common sample areas, the samples of ultimate units may of course be different or overlapping.

(2) The base population of labouring children is likely to be much more unevenly distributed over sample areas compared to the general population of children. A few areas may contain high concentrations, and many areas only very low numbers of working children.
(3) In particular, there may be many ‘zeros’, i.e. areas containing no labouring children of interest in the LCS. Problems such as the presence of extreme (‘very large’ or ‘very small’) areas, defined here in terms of the number of working children the area contains, are likely to be much more widespread than those in the CLS.
(4) The sample size of the LCS is likely to be (or at least should be in a good quality survey) much smaller because of its intensive nature.

7.2.Selection of area units

Given a sample of areas selected according to equations (1)–(3) in the base survey, how to obtain a sample of areas of the type described by (7)–(9) for the LCS? This can be achieved by selecting a sub-sample of areas from the first sample with PPS, the area measures of size for this sub-sampling being the ratio (c_i/p_i). This can be expressed as:

$$g_i = a'\left(\frac{(c_i/p_i)}{\Sigma_s(c_i/p_i)}\right),\ f'_{1i} = g_i \cdot f_{1i} \tag{10}$$

where the sum is taken over all areas in the base (CLS or LFS) sample, as indicated by the subscript $'s'$; g_i is the probability of area i from the base sample being selected into the LCS; and a' is the number of areas to be selected for LCS. This with (1) gives:

$$f'_{1i} = g_i \cdot f_{1i} = a'\left(\frac{(c_i/p_i)}{\Sigma_s(c_i/p_i)}\right)\left(\frac{a}{\Sigma p_i}\right)p_i = \left(\frac{a'}{k_s}\right).\frac{c_i}{\Sigma c_i},$$

$$\text{with } k_s = \frac{\Sigma_s(c_i/p_i)/a}{\Sigma c_i/\Sigma p_i} \tag{11}$$

Hence sub-sampling procedure (10) results in a sample of areas selected with probabilities proportional to size measure c_i, as required. In the above equation, Σ is the sum over areas in the population, Σ_s is sum over areas in the base sample, and k_s is a constant determined by the population and base sample characteristics, independent of the LCS sample or particular area i.

Note that if the LCS sample of a' cluster were selected directly from the population, with area selection probability proportional to the size measure c_i (a function of the number of labouring children in the area), the selection equation would have been (7) instead of (11). Thus k_s is a factor surmising the effect of selecting this sample 'indirectly', via the base survey. If c_i is strictly proportionate to p_i in all areas, it can be seen that $k_s = 1$. In fact factor k_s is not known since c_i values are not known for *all* areas in the population. It also depends on the particular sample which happens to be selected in the first survey – hence (11) does not provide the true selection probabilities in the sense of expected values over all possible base samples. Nevertheless, the value of this factor is expected to be close to 1.0, since its numerator and denominator both estimate average of the ratio (c/p): the numerator is the average in the base sample of separate ratios (c_i/p_i) while the denominator is the combined ratio $\Sigma c_i/\Sigma p_i$ of the same quantities in the population. In any case, this factor does not affect the *relative probabilities* of the area units selected into the final sample, since the factor is the same for all these units. The units are therefore selected with relative probabilities proportional to their size measures c_i.

7.3.Dealing with 'very large' and 'very small' areas

Units with extreme characteristics are likely to occur in the LCS more often than in surveys like the LFS or CLS. Care is required to ensure that correct selection probabilities are achieved for such units. Such problems concerning unit size of course also occur in selecting the base sample. However, these aspects for the base sample generally do not make the treatment of such cases in the LCS more complicated. This is because different sets of areas are involved as 'extreme' cases in the base and the LCS samples since the two use different types of size measures. Hence, generally the two sets can be dealt with separately.

Very large areas

In the context of sub-sampling from the base sample to obtain a sample of areas for the LCS, 'very large' actually refers not to the area population size but to the *degree of concentration of child labour* in it, i.e. to very high values of the ratio (c_i/p_i). This is because this ratio is used as the size measure in (10). Very large are units for which this measure exceeds the sampling interval used in the selection of LCS areas from the base sample:

$$(c_i/p_i) \geq I_s \text{ where } I_s = \frac{1}{a'}\Sigma_s(c_i/p_i).$$

These areas can be treated in the same way as in other cases described in previous sections. For instance, one may redefine any measure of size $(c_i/p_i) \geq I_s$ as $= I_s$, so that any such area in the first sample is taken into the

LCS sample with certainty. These areas thus retain the original probabilities of selection into the base sample unchanged for the LCS. As to the ultimate stage of selecting households or persons in these sample areas, the ultimate stage sampling rate f'_{2i} in equation (8) may be correspondingly adjusted so as to keep the required overall selection probability f' unchanged.

Very small areas

For large areas the PPS sampling procedure needed adjustment only because the size measure (c_i/p_i) exceeded the sampling interval I_s. Therefore areas were identified as being large or not large on the bases of proportion (c_i/p_i) of working children among all children (or persons) in the area. By contrast, areas are defined as being 'very small' in terms of the number of ultimate units they possess or contribute to the sample, i.e. the expected absolute number of working children c_i in the area.

The presence of small c_i values has important practical consequences.

(1) First of all, it should be emphasised that in the design described above of selecting areas from the base survey with probability proportional to (c_i/p_i), areas for which no working children have been reported in the base survey, $(c_i = 0)$, are automatically excluded from the LCS sample. Formally, this of course is also true of areas with no population $(p_i = 0)$ for the selection of the sample areas in the base sample with probability proportional to p_i. But in practice the two situations are quite different. Areas with no population $(p_i = 0)$ tend to be rare and of no interest to the survey in any case, but areas with no working children $(c_i = 0)$ may be very common. Furthermore, the situation with respect to the later (working children) is likely to be much more changing compared to the situation with respect to the former (population). Hence the information on the presence or otherwise of working children in an area needs to be quite fresh.

(2) Even when not exactly zero, very small values of c_i are much more likely to occur than small p_i values. This is because of the uneven distribution of child labour across sample areas. The practical question arises as to whether a lower bound should be put for automatic exclusion from the sample of areas with c_i values below that limit.

With 'very small' areas defined in terms of the size measure c_i, small are the areas whose size measure is smaller than the required sample-take at the ultimate stage, i.e. $c_i < b'$ in equation (8). Procedures for dealing with very small areas are similar to those described earlier. However, different procedures may be used for the selection of households within sample areas; these procedures are probably more varied in labouring children surveys, compared to other, more general surveys of the population.

7.4. Obtaining sufficient sample size for the LCS

Take-all sampling

An obvious concern in LCS is to ensure that the required sample size of working children can be achieved in practice. This can be a problem if the prevalence of child labour is lower than what was assumed at the time of sample design, or if because of poor quality the previous survey missed a large proportion of units subject to child labour. When such problems exist, a 'compact cluster' or 'take-all' design can be an interesting option. In this design all relevant units (e.g. households with a working child) in a selected area are taken into the sample, possibly with an upper limit on the maximum number to be selected. Selection probabilities of ultimate units, as well as sample takes per area, will generally vary in such a design.

Expanding the size of first sample areas

It can happen that the type of areas originally selected in the base sample are generally too small to yield the required number of cases for the LCS. In such situations it may be necessary to consider whether some of the areas – perhaps those with high concentrations of labouring children, areas which are also likely to have such high concentrations in the neighbourhood – can be expanded in physical size to include additional neighbouring areas. One simple procedure is to group areas in the frame into exhaustive and non-overlapping larger areas. Each original smaller area selected brings into the sample the whole of the larger area it belongs to. In statistical terms, the resulting sample would be essentially equivalent to the selection of the larger units, with the probability of selection of such a unit equalling the sum of the selection probabilities of all the smaller units contained within it. Adaptive cluster sampling is another, more sophisticated approach which may be useful and feasible in certain circumstance.

REFERENCES

HUSSMANNS, R, MEHRAN, F. and VERMA, V., 1990. *Surveys of Economically Active Population, Employment, Unemployment and Underemployment: An ILO Manual on Concepts and Methods*. Geneva: International Labour Organisation.

INTERNATIONAL LABOUR ORGANISATION, 2004. *Child Labour Statistics: Manual on Methodologies for Data Collection through Surveys*. Geneva:

International Labour Organisation, International Programme for the Elimination of Child Labour (IPEC).

KORDOS, J., 2008. Bookreview: Sampling for Household-based Surveys of Child Labour, by Vijay Verma*, Statistics in Transition, 9(3), 587–590.*

PORTUGAL MINISTRY OF LABOUR AND SOLIDARITY, 1998. *Child labour in Portugal: Social Characterisation of School Age Children and Their Families*. Lisbon: MTS.

TOMPSON, S.K. and SEBER, G.A.F., 1996. *Adaptive Sampling*. John Wiley & Sons.

VERMA, V., 2008. *Sampling for Household-based Surveys of Child Labour*. Geneva: International Labour Organisation.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 465—502

ADULT EDUCATION AND TRAINING IN COMPARATIVE PERSPECTIVE – INDICATORS OF PARTICIPATION AND COUNTRY PROFILES

Bernhard von Rosenbladt[1]

ABSTRACT

Lack of data has been a serious deficiency for comparative research on adult learning until recently. The present paper explores new opportunities based on the European Adult Education Survey (AES). It argues that general indicators of participation in adult education and training must be complemented by more specific indicators revealing sectoral structures and behavioural patterns within the adult learning system. AES as a subject-specific survey can help to understand better "what is behind" a general level of participation in adult learning, as provided in other statistics. Country variations are explored across a set of 16 European countries.

Key words: Adult education; adult learning; continuing vocational training; international indicators; comparative research

Introduction

International comparability has gained growing importance in education statistics. This also holds for the field of adult education and training. However, lack of data has been a serious deficiency for comparative research until recently. The present paper explores new opportunities based on the Adult Education Survey (AES).

The AES is a new component of the European Statistical Programme. The objective is to generate comparable information on participation in adult education and training across European countries. Based on a 2008 Framework Regulation of the Council of the European Union, the first Europe-wide AES will be conducted in 2011–2012 and then repeated every five years. On a voluntary basis a pilot AES was conducted in 29 countries in the period of 2006–2009, for convenience in the following referred to as 'AES Pilot 2007'. An overview of

[1] TNS Infratest Sozialforschung, Munich. Email: bmrosenbladt@t-online.de.

results based on 18 countries was published by Eurostat in May 2009 (Boateng 2009), and results for more countries were released subsequently.

While in most countries the AES is carried out by the national statistical office there are other arrangements in some countries. In Germany the Federal Ministry of Education and Research (BMBF), following an established line of national surveys on the subject, commissions the AES by tender as a research project. The German AES 2007, after intensive reporting on concepts and results at national level (Rosenbladt and Bilger 2008; Gnahs, Kuwan and Seidel 2008) included a follow-up study to take a closer look at AES results in a comparative perspective. As a starting point, the German AES team organized an AES International Workshop, 9–10 November 2009 in Berlin. A first version of the present paper served as input to that workshop, and a revised version was issued after the workshop (Rosenbladt 2009). The author wishes to thank all participants for stimulating contributions. Special thanks are given to the AES team of Eurostat which supported the work in general and by special tabulations of AES data.[1] Responsibility for the analysis, of course, is fully the author's.

1. Adult education statistics and comparative analysis –new opportunities, new challenges

It is a relatively new development that adult learning, or adult education and training, is included as a strategic area for defining objectives and evaluating achievements in education (Sorvillo 2009). The objective is *"to enable all citizens to acquire, update and develop over a lifetime both job-specific skills and key competences…" (ET2020).* This objective is not easy to grasp in statistical terms. There are two conceptual lines. The first one, at international level followed by LFS (Labour Force Survey) as well as AES, is to measure the participation of adults in learning activities. The second one, at international level followed by IALS (International Adult Literacy Study) or PIAAC (Programme for the International Assessment of Adult Competencies) is output-oriented, measuring skills and competences of the adult population.

Political objectives have been defined in terms of participation levels in adult learning. Statistical indicators of participation provide the basis for monitoring progress towards defined target levels, as well as for ranking countries in terms of more or less developed systems of lifelong learning (Commission of the European Communities 2007 and 2009).

We must be aware of the fact, however, that adult learning is a heterogeneous field: learning activities may have different forms and objectives, may be of short or long duration, and may be supplied in different sub-sectors of the educational system, or even outside the educational system. Statisticians tend to cope with

[1] Eurostat published a set of standard tables for 24 countries on its online portal in June 2009. In addition to that, a scientific use file of the data was made available. The present paper, however, is based exclusively on Eurostat tabulations of AES data.

such problems by elaborating comprehensive definitions and classifications. For the AES the respective basis is the Classification of Learning Activities (CLA) (Eurostat 2006). Still the question remains as to what the statistical indicators really measure and how useful the information is.

This is particularly relevant for international comparisons. In a national framework we often have some common understanding of what we are talking about. Cross-country comparisons are more difficult due to country-specific institutional settings, cultural patterns of learning behaviour, and language-specific terminology.

In the European Statistical System, data on adult education and training in the last decade were based on the European Labour Force Survey (LFS, annually) and its 2003 module on lifelong learning. Indicators for international comparison were derived from this data source, in particular the EU benchmark indicator on participation in lifelong learning (Commission of the European Union 2009). The latest EU Council document on objectives and indicators ("ET2020") confirms that monitoring will continue to be based on the LFS-based indicator of participation in lifelong learning – but adding that "benefit can also be drawn from the information on adult participation in lifelong learning gathered by the Adult Education Survey" (Council of the European Union 2009, p.8). The OECD in its handbook of educational indicators used, for European countries, an LFS-based indicator on participation in job-related education and training (OECD 2005 and 2006) but switched to the new AES data base in its latest edition (OECD 2010, indicator A5).

The indicators used show large differences of the levels of participation in adult education and training across Europe. However, they provide little help in understanding and explaining these differences. Obviously then, it is not easy to draw policy conclusions.

The present paper argues that general indicators of participation in adult education and training must be complemented by more specific indicators revealing sectoral structures and behavioural or organizational patterns within the adult learning system. While the LFS as a general household survey cannot provide this, the AES as a subject-specific survey supplies more detailed information about the learning activities in which respondents participate. This can help to understand better "what is behind" a general level of participation.

One may use AES data in the same way as LFS data, providing an overall participation rate in adult education and training, and showing that countries as well as different groups of the population vary in regard to participation rates. In fact, this is what the latest OECD handbook of indicators does (OECD 2010). AES is thus introduced in the international statistical literature as a new data base for adult learning – though used in a traditional way.

Going beyond this, however, there is an innovative potential of the AES for comparative studies of adult learning. Its analytical advantage is that it provides more information about the dependent variable studied, that is, participation in adult learning. The present paper tries to set up a conceptual road map for using

AES data this way. The first part explores different kinds of indicators which can be used. The second part applies this to selected comparative issues in adult education and training.

We would like to add that statistical analysis as presented here can only be a first step, a basis for further debate. It must be supplemented by country-specific background knowledge about educational systems and traditions. National reporting and international comparative studies need linkages. Efforts to establish suitable frameworks or networks for this end should be supported.

In total, 29 countries have contributed data to the Pilot AES 2007. The analysis of the paper, however, will be confined to a short list of 16 countries (see table 1). This is adequate for the present objective, which is not to give a descriptive overview of European results, but to explore the analytical potential of the data for comparative studies. The list of selected countries is systematic insofar as it represents the full range of the main indicator, from the lowest to the highest level of participation in adult learning across Europe. From the Eastern European, formerly socialist countries, three are selected to represent the range of levels of participation: Hungary (lowest level), Slovakia (highest level) and Poland (medium level). Some countries are missing just because their data were not yet available at the time when the analysis was started. In the tables and figures, a European average (referred to as "EU") is given which normally is the "total" figure provided by Eurostat for the 24 countries in the database at that time. Country figures are weighted by population size in constructing the EU average.

No methodological criteria were applied for selecting countries. Though we feel that there are methodological issues of comparability in regard to the Pilot AES 2007, this is not subject of the present paper. We provide, therefore, only the required minimum of methodological information.

Table 1. Selected countries for analysis AES Pilot 2007, net sample size (persons aged 25–64)

Country		Net sample size
Austria	AT	4.675
Belgium	BE	4.850
Finland	FI	4.144
France	FR	15.350
Germany	DE	6.407
Greece	EL	6.510
Hungary	HU	7.494
Italy	IT	27.848
Netherlands	NL	3.326
Norway	NO	3.018
Poland	PL	24.817
Portugal	PT	9.854
Slovakia	SK	5.001
Spain	ES	16.968
Sweden	SE	3.632
United Kingdom	UK	3.528
EU 16 total (countries included)		151.009
EU 24 total (AES data available)		183.887

"EU" is used as abbreviation for "Europe", independent of membership in the European Union.

National AES surveys are population surveys using random sampling procedures according to country standards. The age range covered may have varied in original surveys but was standardized in the database to represent the population aged 25 to 64 years. National sample sizes in the AES Pilot 2007 differ a lot because some countries incorporated the AES programme in other surveys, such as the national LFS. In Italy, the AES was included in a larger survey of leisure and culture activities. In most countries, however, the AES was conducted as a stand-alone survey, interviewing one respondent per household. The interviewing method was face-to-face in most countries, but some used telephone or internet or a mix of methods.

There was a master questionnaire in English language which was to secure ex-ante harmonization of questionnaires, and a data checking program securing ex-post harmonization of variables. The level of achieved comparability may be lower than desirable, due to the pilot character of the survey, but it seems sufficient for substantial analysis being based on the data. For the first round of official AES surveys in 2011–2012 the degree of methodological standardization will be somewhat higher. Based on the lessons learned from the Pilot, the AES 2007 standard questionnaire was revised for the next round of surveys. The line of argument presented in this paper, however, will still be valid for the improved official AES.

The comparative methodology applied in the following analysis is most straightforward. There are figures or tables for all selected variables studied. Countries are always ordered from the lowest to the highest level of participation in adult learning so that a kind of "correlation" between the respective variable and the overall level of participation can be visually assessed.

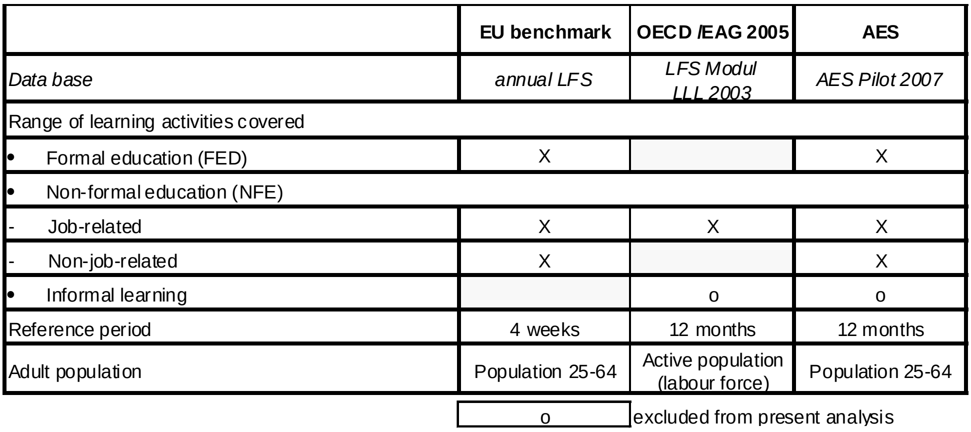
2. Participation in adult education and training – indicators for monitoring and analysis

2.1. Level of participation

The most common indicator of the level of participation in adult education and training is the "participation rate". It indicates the percentage of the adult population participating in specific learning activities in a defined period of time.

There are three parameters in this definition which may vary: the learning activities covered, the adult population covered, and the reference period. In fact, the major international indicators used in the field differ in their combination of these parameters:

Table 2. Scope of the information covered by different indicators of participation



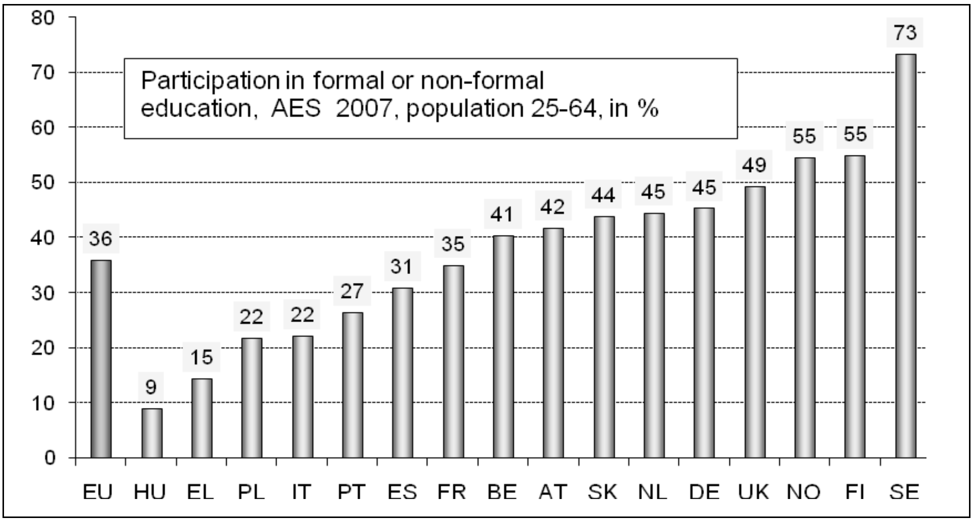
	EU benchmark	OECD /EAG 2005	AES
Data base	*annual LFS*	*LFS Modul LLL 2003*	*AES Pilot 2007*
Range of learning activities covered			
• Formal education (FED)	X		X
• Non-formal education (NFE)			
- Job-related	X	X	X
- Non-job-related	X		X
• Informal learning		o	o
Reference period	4 weeks	12 months	12 months
Adult population	Population 25-64	Active population (labour force)	Population 25-64

o excluded from present analysis

EU benchmark and AES cover formal and non-formal education, job-related as well as non-job-related. The OECD indicator focuses on a specific segment, that is, job-related non-formal education. Informal learning is covered by two of the surveys (LLL 2003 and AES). However, there is some debate as to whether, for conceptual as well as methodological reasons, it should be included in international indicators of adult learning. In this paper, therefore, we leave informal learning apart. The reference period of the indicators differs, which has substantial effects. Regarding the age range of the population, the choice of 25–64 is not theoretically grounded. In fact Eurostat now recommends the broader range of 18–64 (which is used in some countries, such as Germany), but this was not accepted as the general standard.

The core indicator for international comparison is "participation of the adult population (25–64) in formal or non-formal education". Both EU benchmark and AES use it, though with different reference periods. OECD EAG 2010, as well, uses it as its main indicator in the respective field. We may call this the "overall indicator" of participation in adult learning. Figure 1 shows this indicator, based on AES data, for the 16 countries.[1]

Figure 1. The overall participation rate in adult learning



On average across Europe, 36% of the adult population have participated in some kind of formal or non-formal education in the 12 months period before the survey. However, Europe has by no means achieved common standards in this area. There are huge differences across countries, ranging from 9% participation in Hungary to 73% in Sweden. The general pattern is that Scandinavian countries are at the top; the former socialist countries tend to rank low (HU, PL), but there are remarkable exceptions (SK); Mediterranean countries are all below the European average. The other Western European countries form the medium group, ranging between EU average (FR) and clearly above average (NL, DE, UK).

[1] Figures for all 24 countries are given in the Annex.

The overall indicator covers, as mentioned, both formal and non-formal education. These are normally quite distinct kinds of learning activity though the line between them may not be sharp in all cases, and national terminology may vary to some degree. This is an argument for using the overall indicator for international comparison. Yet one should be aware of the fact that different kinds of learning activities are put here in one basket. According to the Classification of Learning Activities (Eurostat 2006)

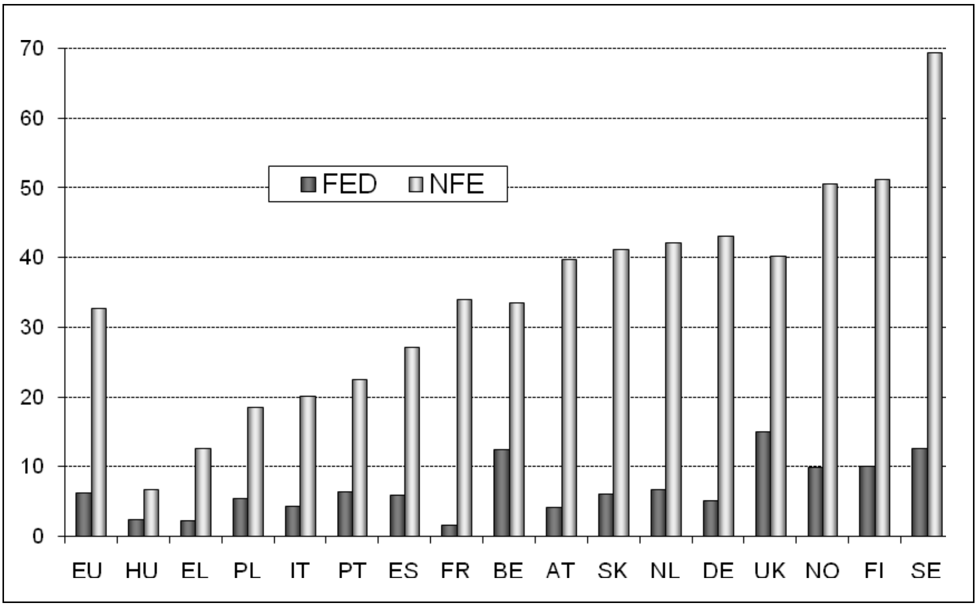
- formal education is defined as education provided in the system of schools, colleges, universities and other formal educational institutions and which normally constitutes a continuous "ladder" of full-time education for children and young people
- non-formal education is defined as an organised and sustained educational activity that may take place both within and outside educational institutions and cater to persons of all ages.

The AES questionnaire specifies the meaning of non-formal education by enumerating more concrete forms of learning activities such as courses, workshops and seminars, guided on-the-job training, private lessons and distance learning. Individual learning activities are identified according to these categories. In the course of the 12 months reference period, a person may have been involved in more than one learning activity. Those engaged in formal education may also report non-formal learning activities. Therefore, if participation rates for more specific categories of adult learning are summed up, the sum is higher than the overall participation rate.

Figure 2 shows specific, or disaggregated participation rates for formal and non-formal education. Considerably more adults participate in non-formal learning activities (EU average 33%) than in formal education (6%). This holds for all countries, but the relative weight of the two forms of learning varies. In some countries, enrolment in formal education is very low among adults (France 2%, Germany 5%). This means that the overall participation rate mainly reflects non-formal learning activities. By contrast, participation in formal education amounts to 15% in UK and 13% in Belgium. One would like to know what is behind these figures. Do they reflect real differences of the educational systems, or just different ways of classifying certain learning activities? Section 3.1 of this paper will explore this issue further.

Saying that one would like to know what is behind the figures applies even more to the non-formal learning activities. We know by definition that they cover courses, seminars, workshops, guided on-the-job training and private lessons. But, more exactly, what do people do – or do not do – under non-formal learning activities in different countries with different levels of participation – such as in Sweden (69%) or Finland (51%) or less in Germany (43%) or France (34%) or Poland (22%), and hardly any in Hungary (7%). Section 2.3 of this paper will explore this issue further.

Figure 2. Participation in formal education (FED) and in non-formal learning activities (NFE) AES 2007, population 25–64, in %

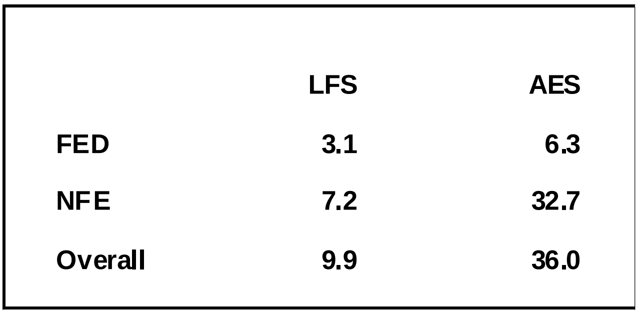


Going back to the overall indicator of participation in adult learning, we can compare the two versions of it that co-exist now in the European Statistical System. One is the AES-based indicator used here as well as in the OECD handbook of educational indicators (OECD 2010), the other is the LFS-based indicator of "lifelong learning" in the EU benchmark system of educational objectives (Council of the European Union 2009; Commission of the European Communities 2009). The range of adult learning activities covered is roughly the same in the two indicators. They differ in (a) the reference period for reported learning activities, which is 12 months in the case of AES and 4 weeks for the EU benchmark, and (b) the data source, which is an issue-specific population survey in one case and a general household survey in the other.

The shorter reference period of the LFS indicator (EU benchmark) has two effects. Nearly by definition, one effect is a lower level of reported participation in adult learning: 9.5% (LFS) versus 36% (AES) of the adult population. Of course, statistical analysis may use one or the other reference period. As a measure of individual learning behaviour, however, the 36% rate of the AES seems to provide a more meaningful picture. It shows that substantial numbers of adults do engage in some organized learning activity at some time during a year. Non-learners in the short term perspective of the LFS may well be learners in a more extended time perspective.

Besides the level effect there is a structural effect. Formal education (FED) is normally a long-term activity, while non-formal learning activities (NFE) are often short-term and distributed over time (see section 2.2 for empirical evidence). The longer the duration of the activity is, the longer the probability of an activity to come into sight in a short observation period rises. By consequence, FED activities have a higher probability of being represented in the LFS indicator than NFE activities. Prolonging the reference period to 12 months, as AES does, brings up the reported participation only moderately in FED, but much more so in NFE (see table 3). Partly this may also be caused by a more detailed specification of NFE activities in the AES. Consequently, the LFS indicator reflects relatively more the level of participation in FED, while the AES indicator reflects relatively more the level of participation in NFE.

Table 3. Components of the overall indicator: formal (FED) and non-formal education (NFE). Participation rates in %, EU average



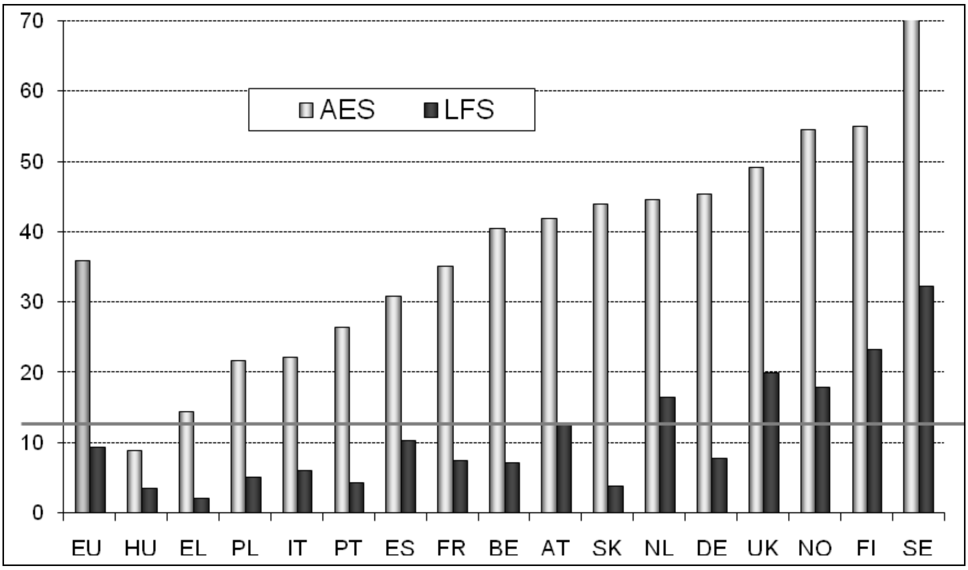
	LFS	AES
FED	3.1	6.3
NFE	7.2	32.7
Overall	9.9	36.0

Source for LFS: Commission of the European Communities (2007b), p. 81

Figure 3 compares the two indicators across the 16 countries. The general pattern of the EU benchmark is similar to the AES indicator, showing large differences within Europe. Participation rates range from 2% in Hungary to 32% in Sweden. The grey line in the chart marks the 12.5% line which was the EU target for 2010 (now substituted by a 15% target for 2020) (Council 2009). Most of the countries are far below the line.[1]

[1] Exact figures are given in the Annex, showing the full list of countries with the latest available results.

Figure 3. Overall participation according to AES and LFS (EU benchmark) 2007, population 25–64, in %
AES: past 12 months, LFS: past 4 weeks



Some countries score surprisingly low in the EU benchmark, compared to their relative position according to AES. This mainly holds for Germany, Slovakia, Belgium, Portugal and Greece. There must be measurement problems behind this (either in the AES or the LFS, or in both), which cannot be resolved here. It is clear, however, that diverging results such as these cause irritation, as they suggest different conclusions for policy-makers.

To illustrate the meaning of divergent results one may spell out the message of the two statistics for Germany. The EU benchmark would read, and in fact is read, as follows: “Lifelong learning is a key issue for a competitive, knowledge based economy and an inclusive society. Germany is a poorly developed country in this respect, and far from reaching the European targets. There is an urgent need of catching up with international standards in the field.” By contrast, AES results would read as follows: “In Germany, compared to other European countries, participation in adult education and training is quite well developed. Some countries, in particular the Scandinavian countries and UK, show higher levels of participation. This may encourage all actors involved to follow the government’s policy line of further increasing the participation in adult learning.”

The best way of dealing with divergent results of alternative indicators would be to disaggregate the overall participation rate in order to find out which segments of adult education and training are covered differently in the two sources. The problem is that LFS data do not include sufficient detail to show what is behind the overall indicator.

Finally we just mention the third international indicator of participation in adult learning which has gained importance, i.e. the LLL 2003 indicator as reported in the OECD handbook of educational indicators (OECD 2005). The Lifelong Learning module (LLL) of LFS 2003 is a kind of forerunner of AES, using the same classification of learning activities (formal, non-formal, informal) and a reference period of 12 months. The most prominent indicator derived from the data is the participation rate of adults (labour force) in job-related non-formal education. (See overview of indicators in the Appendix.) Results can be compared to those of an AES-based indicator defined correspondingly (OECD 2010, table A5.5). Levels of participation according to both indicators are nearly the same in some countries (e.g. Finland, UK) but diverge heavily in others. In particular Germany, Spain, Portugal and Italy score much lower using the LLL 2003 data compared to AES 2007 data. It seems that the LFS-based LLL module captured respondents' learning activities of the past 12 months more successfully in some countries than in others. This may be regarded as statistical history. However, it is worth mentioning because recent comparative research literature on adult learning still refers to LLL 2003 as data source (e.g. Müller/ Kogan 2010, p.264; O'Conell/ Jungblut 2008, p. 112).

2.2.Volume of participation

The participation rate does not distinguish between longer or shorter learning activities. A student in a full-time educational programme of three years, covering all 12 months of the reference period, is counted in the same way as a person attending a seminar of three hours. This is an apparent weakness of the "participation rate" indicator.

Information on the duration or volume of learning activities may be used for two objectives. First, it is a fundamental aspect in descriptive terms. We need volume information in order to know what we are talking about. Second, individual cases of participation may be "weighted" by the time factor. The "volume of participation" may be regarded as a more comprehensive indicator of participation than the participation rate.

As a measure of volume, the AES uses "instruction hours", that is, not the total time devoted to learning activities[1]. Instruction hours are the basic information, representing also a relevant organizational aspect of the supply side of adult education.

The number of instruction hours asked in the survey refer to single learning activities as described by the respondent. Starting from this, other indicators can be generated via aggregation. Relevant dimensions of volume are:

1. the duration of single learning activities (instruction hours)

[1] The AES Pilot included, as a separate variable, information on preparatory time as well, but this was skipped for the AES 2011/12 and will not be considered here.

2. the number of learning activities an individual participates in within the reference period of 12 months
3. the cumulated annual instruction hours of individual respondents
4. the percentage of the population participating in adult learning within the reference period of 12 months (participation rate)
5. the aggregated volume of participation, that is, the number of instruction hours per head of the population (per capita).

Starting at the first level, table 4 shows the duration of single learning activities, with classes defined identically for formal and non-formal learning activities.

Table 4. Duration of learning activities (instruction hours) AES 2007, population 25–64, all-European sample, in %

number of hours	NFE	FED
1-10	39	6
11-40	39	16
41-100	14	16
101-200	5	15
201-500	2	20
500-800	1	10
801+	0	15
Total (valid cases)	100	100
Item non-response rate (%)	5	21

Non-formal education and training (NFE), covering courses, seminars, workshops, on-the-job training and private lessons, is mostly of short duration. Nearly 40% of the reported learning activities are short activities of up to 10 hours, which will normally be a one day or two day activity at most. Another nearly 40% has a duration of 11 to 40 hours (the upper value meaning a week if organized full-time). But there are longer courses also: 22% above 40 instruction hours, 3% above 200 instructions hours. So there is a large range, but mostly we are talking here of short units of instruction.

By contrast, formal education programmes are more long-term. Certainly there a measurement problems here when the duration of participation is to be given in hours. In the questionnaire this may be broken up into weeks or months plus average hours per week. Yet there are a lot of missing values in the data

(21%)[1]. One must also be aware of the fact that for any programme lying partly outside of the reference year, only the duration falling within that year is recorded. The actual length of the programmes may be longer than reported here.

Most FED activities are of medium or long duration. Nearly 80% are above the line of 40 instruction hours, and 25% are above 500 hours. More surprising are the short durations: 22% of the reported FED activities are in the range of up to 40 instruction hours, and 6% even in the range of up to 10 hours. This is to a large extent country-specific, indicating a non-uniform understanding of what kind of learning activities are "formal" (which will be explored further in section 3.1). Not surprisingly, the borderline between formal and non-formal education is empirically not as sharp as in the abstract definition. This may be an argument for using an overall indicator, including both forms of adult learning, in comparative studies across countries.

Returning to non-formal education, table 5 provides the detail of information required for analysing the volume dimension. For the European average, figures read as follows: One third of the population has participated in the past 12 months in some kind of non-formal learning activity. Those who did, report two different learning activities on average. The average duration of individual NFE activities is 35 instruction hours. For two activities in the 12 month period, this amounts to an annual average of 70 instruction hours per participant. Relating this to the total adult population, including non-participants in such learning activities, the aggregate volume of participation in non-formal education and training per capita is 23 hours.

[1] The level of missing values varies considerably across countries. For NFE the range is between zero (DE, HU) and 23% (BE). For FED it is between 1% (PL) and about 30% (EL, BE, DE) or even 88% in France. The Netherlands did not provide any data for this variable. For the next AES, better data quality at this point should be strived for. This includes methods of imputation which can be used here. Imputation may have been used for AES 2007 in individual countries, but there is no detailed information on national procedures.

Table 5. Components of the volume dimension in non-formal education and training AES 2007, population 25–64

	(1)	(2)	(3)	(4)=(2)*(3)	(5)=(4)*(1)
	Participation rate NFE %	Average number of activities	Average hours per activity	annual hours per participant	annual volume (hours) per capita
EU	33	2,0	35	70	23
HU	7	1,2	95	111	8
EL	13	1,3	65	86	11
PL	19	1,5	54	82	15
IT	20	3,4	14	48	10
PT	23	1,5	61	93	21
ES	27	1,5	74	112	30
BE	34	2,2	52	114	38
FR	34	2,3	25	57	19
AT	40	1,9	49	92	37
UK	40	1,6	30	48	19
SK	41	1,8	33	58	24
NL	42	2,1	28	59	25
DE	43	2,0	39	76	33
NO	51	2,7	29	78	39
FI	51	2,2	43	95	49
SE	69	3,1	24	73	51

Countries are ranked here according to their participation rate in NFE (column 1).

Country variations show the following patterns:

- The higher the country's participation rate in non-formal adult learning (column 1), the higher tends to be the reported average number of learning activities per participant within the past year (column 2). Hungary has the lowest participation rate (7%) and the lowest number of reported activities per participant (1.2). At the other end, Sweden has the highest participation rate (69%) and at the same time a very high average number of learning activities (3.1 per participant). One exception is Italy which has a relatively low participation rate but the highest number of reported learning activities (3.4 per participant).
- The more learning activities are reported (column 2), the shorter tends to be the duration of the single learning activity (column 3). Hungary has the longest average duration (95 hours), whereas Sweden (24 hours) and Italy (14 hours) have the shortest.

- This means, by and large, that an increasing level of participation in non-formal education is not just “more of the same”. It goes along with a tendency to have more but shorter learning activities.
- Because of the inverse relationship between frequency and duration of learning activities, the cumulated annual instruction hours of participants (column 4) show no clear correlation with the country’s level of participation.
- The aggregate volume of participation in terms of instruction hours per head of the population (column 5) is determined by the amount of time devoted to learning activities as well as the participation rate among the population. The latter tends to be the strongest determinant. Therefore, the ranking of countries based on the volume indicator is similar to that based on the participation rate (column 1). At the lower end there is Hungary with 8 hours of non-formal learning annually per head of the population. At the upper end there is Sweden with 51 hours. However, for some countries the time dimension changes their comparative position.
- UK is an interesting case in this respect. UK seems to spread the overall volume of non-formal education and training to a high number of people, but with a relatively low intensity as measured in terms of instruction time. The participation rate is high, but the aggregate volume of participation is below the European average.
- On the other hand, there are countries like Belgium, Austria and Spain. Based on the participation rate they rank at a medium level of participation in non-formal education and training. However, due to relatively long instruction time per participant, they score clearly above the European average in regard to the aggregate volume of non-formal adult learning.

Similarly, one can analyse the volume dimension of formal education (FED). The European average figures read as follows (table 6):

- Among adults of 25–64 years 6% have participated in some programme of formal education during the past 12 months (column 1). The average number of instruction hours falling in this period is 382 hours (column 2). Assuming a weekly duration of 30 hours this would amount to three months.
- Calculated per head of the population the aggregate volume of participation in formal education is 24 hours (column 3). This is about the same volume as for non-formal education where it is 23 hours. The composition of the overall volume, however, is different in the two sectors: In FED the volume is made up of few participants with a long duration of learning activities, in NFE it is made up of many participants with a short duration of learning activities.

Table 6. Components of the volume dimension in formal education (FED) AES 2007, population 25–64

	participation rate %	annual instruction: hours per participant	annual volume: hours per capita
	(1)	(2)	(3)=(1)*(2)
EU	**6**	**382**	**24**
FR	2	339	6
EL	2	405	9
HU	3	487	12
AT	4	532	22
IT	4	367	16
DE	5	905	47
PL	6	420	23
ES	6	413	25
SK	6	326	20
PT	7	543	35
NL	7	:	:
NO	10	379	38
FI	10	399	41
BE	13	230	29
SE	13	515	65
UK	15	121	18

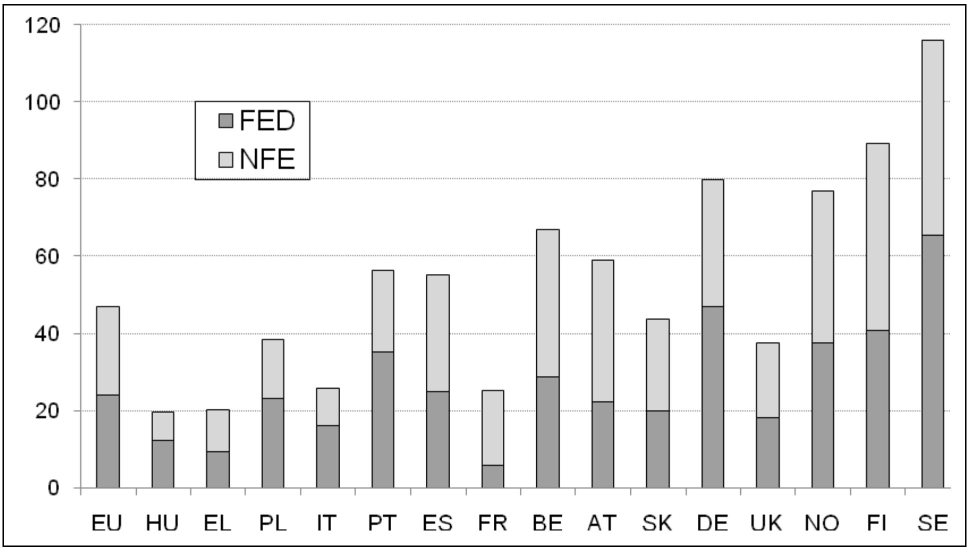
Countries are ranked here according to their participation rate in FED (column 1).

Country variations show the following patterns:

- There is no clear correlation between the participation rate in formal education (column 1) and duration in terms of instruction hours (column 2).
- At the lower end, with short durations, there is the UK again with annual instruction hours of 121 hours per participant. This is a surprisingly short duration which some other countries (ES, BE) show for non-formal learning activities.
- At the upper end there is Germany with 905 instruction hours on average. Assuming a weekly duration of 30 hours, this amounts to about seven months, which is not unrealistic if most participants are full-time students in long-term educational programmes.
- The aggregate annual volume of formal education (column 3) varies from 6 hours per head of the population in France up to 65 hours in Sweden. The contrast between Sweden and UK is interesting: the two represent equally high levels of participation, but the average duration of programme participation is much longer in Sweden than in the UK.

The aggregate volumes of formal and of non-formal education may be summed up to form one overall indicator of the volume of participation in adult learning. It is true that, when putting formal and non-formal education in one basket, we mix things which mostly are quite different. Yet the overall volume indicator seems meaningful, representing the total amount of instruction time for adult learning activities, formal and non-formal. Figure 4 shows the aggregate indicator with its two components (FED and NFE) across countries. Countries are ranked according to the overall participation rate.

Figure 4. Aggregate volume of formal and non-formal education (FED/NFE): annual instruction hours per head of the population (25–64)



The European average of the overall volume of participation in adult learning is 47 hours per head annually. Country variations range from 20 hours (HU, EL) up to 116 hours (SE). The position of countries in the comparative ranking is by and large similar to the picture we know from the overall participation rate (figure 1). However, some countries rank much lower if the time dimension is included (IT, FR, SK, UK).

For international comparative studies, volume based indicators of participation in adult learning may be attractive because they incorporate more information than just the number of people involved.

In policy terms, one may ask whether 47 hours of adult learning is a satisfying volume. Related to the volume of annual working hours, for instance, the amount of time invested in learning activities is still marginal. It is true that simple quantitative measures such as these are not sufficient for evaluating lifelong learning and skill formation in modern societies (for a broader discussion see Müller and Jakob 2008). However, comparative research needs some simple measures. Adding the time dimension to statistical indicators of participation in adult learning would at least bring in the aspect of intensity of learning – one step towards including quality issues in the debate.

2.3.Profiles of participation

Overall indicators of participation in adult learning – referring to participation rates or volume – are quantitative measures in the sense of higher or lower, more or less. Countries are located at a specific point on a scale which is interpreted as a continuum from poor to excellent.

This is an important item of information. When trying to understand the countries' adult education system, however, we must look for more qualitative information. This is often derived from other types of policy studies. The present paper argues that a database such as the AES has a potential for more qualitative information in itself. The comparative analysis can reveal "profiles of participation" in individual countries. The more we know about such profiles, the more we can evaluate the significance of the different levels of participation observed.

Profiles can be explored regarding different aspects or dimensions of adult learning. To illustrate the concept, we are selecting three aspects which are relevant in terms of policy options:

1. Time-related profiles (patterns of intensity)
2. Sector-related profiles (patterns of institutional structure)
3. Group-related profiles (patterns of social inequality)

The first of these profiles builds on the analysis of the preceding section regarding the volume of participation. We will only briefly take up some thoughts here.

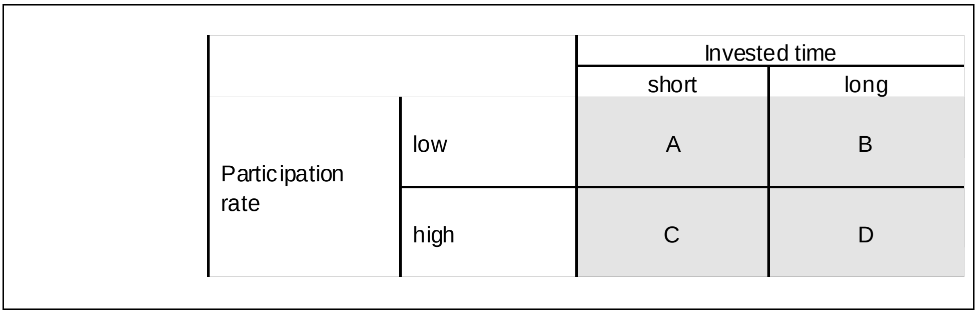
The second profile builds on a typology of adult education and training specifying the main sub-sectors. We will briefly present the typology here. Sections 3.1 (role of formal education) and 3.2 (role of enterprises) will look more closely into selected types or sectors.

The third profile refers to the fact that we observe very different levels of participation in adult learning not only across countries, but also across groups of the population within countries. Age groups will be examined in the present section, while section 3.3 will look at the importance of educational background.

We are discussing the three aspects separately. In fact they may be interrelated. This would have to be explored in more elaborated studies.

Time-related profiles (patterns of intensity)

A certain volume of adult education and training can be distributed over the population in different ways: either relatively few persons are involved in time-intensive studies, or many persons take part in shorter learning activities. This is a two-dimensional perspective, considering the number of persons involved (participation rate) and the amount of time devoted to learning activities by participants. Constructing a simple typology leads to four possibilities:



		Invested time	
		short	long
Participation rate	low	A	B
	high	C	D

Tables 5 and 6 above have shown examples of all four combinations. In countries with low levels of participation there is a tendency towards type B. Distinct cases are Hungary in regard to non-formal education (NFE) and Germany in regard to formal education (FED). However, there are also cases of type A: France in regard to FED, and Italy in regard to NFE. The UK was already identified as a distinct case of type C (in FED as well as NFE). The Scandinavian countries tend towards type D, with most pronounced values in Sweden. It should not be precluded, however, that some of the observed differences across countries may have been affected by non-uniform measurement and classification.

It is evident that adult education policies in a country cannot just switch from one type to the other. In principle, however, the time-related patterns can be viewed as options for policy directions.[1]

The most advanced country in terms of participation rates in adult learning is Sweden. Figures tell that it is quite normal for an adult to engage in some course or seminar. The individual learning activities are fairly brief units of instruction, but participants attend several of them in the course of a year. This is a pattern one may call “the normality of lifelong learning”.

Sector-related profiles (patterns of institutional structure)

The adult education system is more heterogeneous than other fields of the educational system, say schools or higher education. Often it is not even perceived as one policy field, and political competences may be split for different sub-sectors such as vocational training and general interest education for adults. There is a multitude of actors and providers in the field. Transparency is often low, even at national level. Across countries, the picture will be even more heterogeneous.

[1] For example, as noted by Steve Leman, “Turning to the policy aspect: is UK practice an optimal situation – spreading the resources widely so that more people benefit? Or is it a poor use of resources – not enabling individuals to make significant advances in their knowledge and skills? What can we learn from other countries’ experience?“ (Leman 2009, p.2).

The conclusion drawn from this could be to refrain from any attempt to structure the field, apart from distinguishing between formal and non-formal education. On the other hand, this would mean we do not know a lot when saying that country X has a certain participation rate. Participation in what? And is the spectrum of learning activities behind the participation rate the same in country X and country Y? Or are we comparing apples and oranges? This is a question in all comparative studies, but certainly it seriously applies to adult education.

We therefore need some empirically feasible structuring of the field. Criteria for defining sub-sectors should reflect institutional structures, as far as possible. Our proposal for a sector typology is shown in table 7.

Table 7. Sector typology of adult learning

Level 1	Level 2	Level 3
All learning activities (FED and NFE)	Formal education (FED)	Sector 1: FED below tertiary level
		Sector 2: FED at tertiary level
	Non-formal education (NFE)	Sector 3: Employer-sponsored training
		Sector 4: Other job-related training (off-the-job)
		Sector 5: Other adult education (general interest)

The typology maintains the basic distinction between formal and non-formal education (level 2). It is true that the two forms of learning as introduced in the international Classification of Learning Activities (CLA) are originally not meant to define "sectors" of the adult education system. Yet the distinction between the two forms is mainly based on institutional criteria, and one may well regard formal and non-formal education as two sectors within the adult education system (keeping in mind that formal education is mainly targeted to younger age groups). For the next level of disaggregation (level 3) criteria must be sought in the data set so that operational definitions can be made.

In formal education, sub-sectors can be defined according to the level of educational programmes. For the age group studied here it is sufficient to distinguish two levels: below tertiary ("sector 1") and tertiary or higher education ("sector 2").[1]

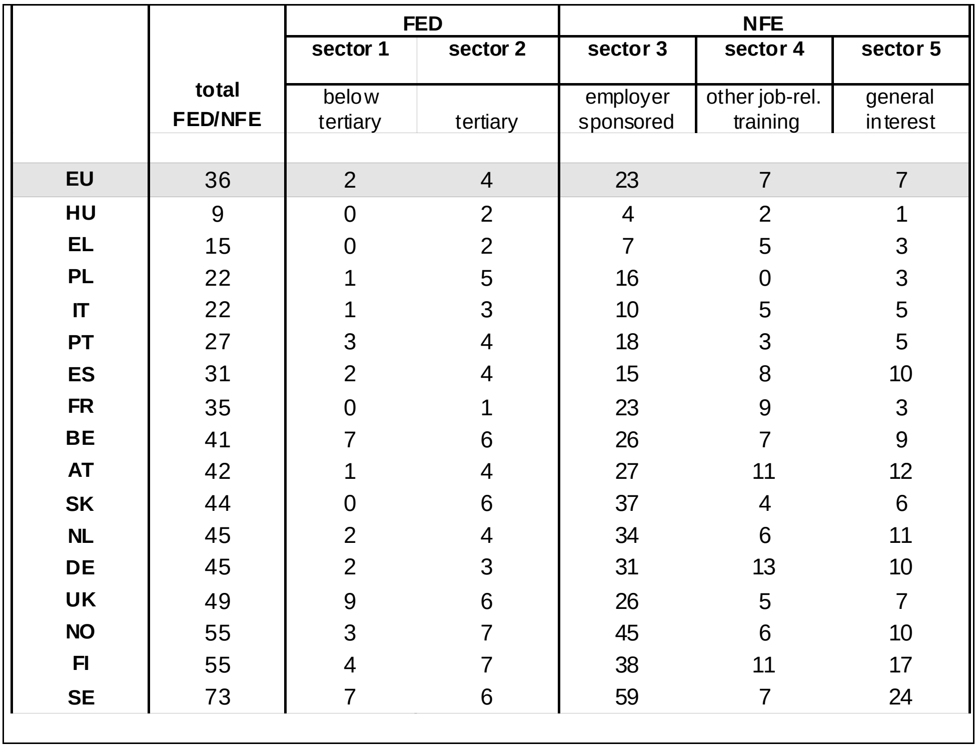
[1] Alternatively, one could distinguish between vocational and non-vocational programmes. This variable, however, is not available in the Pilot AES.

In non-formal education, the main requirement is to identify company-based or employer-sponsored training (“sector 3”). Criteria are whether the learning activity takes place during paid working time or the employer pays for it in another way (discussed more closely in section 3.2). The remaining non-formal activities are sub-divided according to the criterion of being job-related or not. The distinction is based on the subjective view of the respondent – whether the reasons for participating are “mainly job-related” or “mainly personal, non-job-related”. It must be underlined that this does not necessarily correspond to the distinction of “vocational training” versus “general adult education”. Off-the-job but job-related training is “sector 4”, all other adult education (general interest) is “sector 5”.

The overall participation rate in adult learning can be disaggregated to show participation rates in individual sectors. Table 8 provides the numbers. The sum of the sector-related participation rates at EU level is 43%, that is, it is higher than the overall participation rate of 36%. The reason is that individual respondents, in the course of 12 months, may participate in more than one learning activity, and different activities may fall in different sectors.[1]

[1] This is an aspect of the methodological design which has important implications for data analysis. In the Pilot AES 2007, the general rule was that respondents describe up to three non-formal learning activities in detail. If they reported more than three activities, the selection of three activities was made at random. This affects the results concerning a disaggregated participation rate. This measure can be calculated only for the set of activities which are described in detail. If a respondent has reported more activities than described, one cannot say reliably whether he or she participated in a certain sub-sector. Therefore, the calculated participation rates for sub-sectors represent minimum values. This problem gets more serious if less than three activities are described in detail. Unfortunately, in the AES Pilot, there were national surveys describing only one activity in detail (FR, IT, UK). For the next round of AES surveys in 2011–2012 the rule is that a minimum of two activities (if reported) must be described.

Table 8. Participation rate in adult learning, disaggregated according to sectors AES 2007, population 25–64, in %



	total FED/NFE	FED		NFE		
		sector 1	sector 2	sector 3	sector 4	sector 5
		below tertiary	tertiary	employer sponsored	other job-rel. training	general interest
EU	36	2	4	23	7	7
HU	9	0	2	4	2	1
EL	15	0	2	7	5	3
PL	22	1	5	16	0	3
IT	22	1	3	10	5	5
PT	27	3	4	18	3	5
ES	31	2	4	15	8	10
FR	35	0	1	23	9	3
BE	41	7	6	26	7	9
AT	42	1	4	27	11	12
SK	44	0	6	37	4	6
NL	45	2	4	34	6	11
DE	45	2	3	31	13	10
UK	49	9	6	26	5	7
NO	55	3	7	45	6	10
FI	55	4	7	38	11	17
SE	73	7	6	59	7	24

Results of the sector-related disaggregation may be summarized as follows:

1. Employer-sponsored or company-based training (sector 3) is by far the largest sector. At EU level, 23% of the population take part in such training within the 12 month reference period. Participation rates in all other sectors are much lower. As regards country variations, the sector of company-based or employer-sponsored training shows the largest range across countries, from 4% participation in Hungary up to 59% in Sweden. It is this sector which determines most strongly the overall participation rate in adult education. The conclusion is that, when we are looking for core issues of adult education in a comparative perspective, we have to explore the role of the employer- or company-based sector. Section 3.2 of this paper will take up this subject.
2. Within this overall pattern, countries have their individual profiles. For instance:
   - Regarding formal education, in most countries adult education is mainly located at tertiary level. In some countries, however, the focus is more on the below-tertiary level (UK, BE, SE).

- Regarding non-formal education, participation in off-the-job training (sector 4) is below 10% in most countries. In some countries, however, this sector plays a larger role (DE, AT, FI).
- The sector of non-job-related, general interest education (sector 5) has a very different size across countries, ranging from a participation level of 1% in Hungary up to 17% in Finland and 24% in Sweden.

The sector typology provides relevant information on what is behind the overall participation rate in adult education. Knowing a country's sector profile contributes to a more profound understanding of its comparative position.

Group-related profiles (patterns of social inequality)

Disaggregating the participation rate in adult education by groups of the population is the most common line of analysis. A lot of studies show that there are differential participation rates in adult learning according to social characteristics of the respondents. For instance, Eurostat standard tables on AES results, as published on the Eurostat portal, comprise breakdowns of participation rates (for FED, for NFE and for FED/NFE) by sex, by age groups, by education attained, by labour status, by occupation and by degree of urbanisation. Other breakdowns can be found in the OECD handbook of educational indicators (OECD 2010).

When studying the comparative tables, the first impression is that general patterns are fairly similar across countries. However, for two reasons this needs a second look:

- Descriptive figures at group level may be misleading because of underlying factors. A common example is the participation of men and woman, or of different age groups, in adult learning. The "real" explaining factors of group differences must be explored, e.g. by further disaggregating the figures or by multivariate analysis.
- The pattern of group-related differences may be similar across countries, but the magnitude of difference may vary. If we regard group differences of access to adult education as a form of social inequality, the magnitude of differences is a measure of the degree of inequality.

We focus on the second aspect. The relevance of educational will be explored later in section 3.3. Here we take a look at age groups. Age is an important dimension because it directly relates to the "lifelong" in lifelong learning. To what extent can we identify patterns of participation across countries in this aspect?

Table 9 provides the figures. Countries are ranked, again, according to their overall participation rate in adult education (FED and/or NFE). Participation rates of age groups are shown in their relative position in regard to the country's total. At EU level, the younger age group (25–34 years) has a score of 1.26, that is, the overall participation rate of the group is 26% higher than that of the total population. The middle age group is near the average at large. The elder group

(55–64) has a score of 0.60, meaning that its participation level is 60% of the total population average.

In all countries the participation level of the elderly is lower than that of the total population. However, their relative position varies considerably across countries, in a range of 0.28 in Hungary to 0.83 in Sweden. The relative position of the youngest age groups varies inversely. Adult learning in Hungary is strongly concentrated on early stages of adult biographies, while in Sweden it is maintained at a relative stable level over the (working) life course.

It would need more in-depth studies to analyse background factors, such as the employment rate of the elderly in the respective countries. For the present paper we just note that there is a clear relationship of the age-group related pattern with the overall level of participation in adult learning. The higher the overall participation rate, the more equal is the distribution of participation across age groups. One may take this as another aspect of what we have called "the normality of lifelong learning".

Table 9. Relative participation rate of age groups in adult learning (FED/NFE) AES 2007, population 25–64

Country	Participation rate (%)	Relative participation rate (index, total = 100)		
	Total	Age 25–34	Age 35–54	Age 55–64
EU	33	1,26	1,04	0,60
HU	7	1,76	1,00	0,28
EL	13	1,57	0.97	0,35
PL	19	1,56	0.95	0,31
IT	20	1,37	1,04	0,53
PT	23	1,52	0,96	0,41
ES	27	1,28	1,00	0,55
BE	34	1,37	1,02	0,46
FR	34	1,39	1,04	0,58
AT	40	1,12	1,09	0,61
UK	40	1,16	1,10	0,54
SK	41	1,34	1,01	0,65
NL	42	1,17	1,07	0,62
DE	43	1,19	1,02	0,75
NO	51	1,19	1,02	0,75
FI	51	1,20	1,07	0,69
SE	69	1,10	1,04	0,83

3. Comparative issues in adult education and training

The set of indicators described in previous sections is a kind of tool kit which can be applied to the comparative analysis of more specific issues in adult education and training. “Comparative issues” as discussed here relate to aspects or sub-sectors of the overall field of adult learning which appear to be crucial for identifying what exactly is different in particular countries. This is the prerequisite for evaluating different levels of participation, and to ask what kind of policy conclusions may be drawn from the figures.

Naturally a multitude of comparative issues can be defined. We are selecting three for more in-depth discussion here, building on starting points provided above: (1) the role of formal education in adult learning, (2) the role of enterprises in adult education and training, and (3) social inclusion – participation of the lower educated in adult education and training.

3.1. Comparative issue (1): the role of formal education in adult learning

According to the AES Manual, following the UNESCO Manual for statistics on non-formal education, a learning activity is considered to be formal when upon its completion it leads to a learning achievement (qualification or award) that can be positioned to the National Framework of Qualification (NFQ). By and large, this means enrolment in regular institutions of the educational system, such as schools, colleges, universities or (in some countries) apprenticeships.

The overall participation rate in formal and non-formal adult education mainly reflects the non-formal part of the spectrum. Enrolment in formal education is relatively rare in the age group of 25–64. The interesting point is, however, that there is considerable variation across countries in this respect (see figure 2 above and table 10 below).

The average participation rate in formal education (FED), using a reference period of 12 months, is 6%. Participation rates in individual countries range from below 2% in France or Greece up to 15% in the UK. Other countries with participation rates above average are Belgium and Sweden (12–13%) as well as Norway and Finland (10%). Such figures are surprising, looked at from the perspective of countries at the lower end, including Austria or Germany, where the level of participation in formal education among adults is reported to be much lower (about 4–5%).

AES data provide additional information, or FED profiles, put together in table 10. The first aspect is the age distribution of participation in FED programmes. Table 10 presents this in the form of relative participation rates (like in table 9).

Not surprising, participation in formal education is mainly found in the lower age group of the 25–34 years old respondents, but it also occurs in the middle and even in the higher age group. In the lower age group it may partly be the final stage of “initial education”, e.g. for university students. Higher participation rates in this age group might indicate ineffective organisation of tertiary education, instead of continued learning in adulthood. Unfortunately, the comparative AES data do not include information on this crucial distinction. The German AES 2007 showed that about one third of formal education in the age group above 18 (!) years is not “initial” but “further” education, that is, part of a second phase of education after some time of employment (Rosenbladt and Bilger 2008, pp. 48–52).

Countries with a higher overall participation rate in FED tend to show higher participation rates in all three age groups. Findings clearly indicate different patterns of the role of formal education in peoples’ educational biographies. While in some countries formal education is largely restricted to the early years of adulthood, in other countries it is part of the repertoire of options for lifelong learning.

One would expect that this goes along with changing forms and content of the respective educational programmes. We examine this by looking at the volume of annual instruction hours and the ISCED level of the respective educational programme (right-hand columns in table 10).

Table 10. Profiles of participation in formal education (FED) AES 2007, population 25–64

	Relative Participation rate for age group (index, total=1.0)			Total participation rate (%)	Average annual instruction hours	% distribution of all FED activities by level of programme	
	25–34	35–54	55–64			below tertiary	tertiary level
EU	2.3	0.8	0.3	6	382	38	62
FR	2.5	0.5	0.0	2	339	19	81
EL	3.0	0.5	0.0	2	405	12	88
HU	2.3	0,7	0.0	3	487	16	84
AT	3.0	0,5	0.0	4	532	22	78
IT	3.3	0.8	0.3	4	367	44	56
DE	3.0	0,6	0.4	5	905	35	65
PL	2.2	0.5	0.0	6	420	55	45
ES	2.0	0.7	0.3	6	413	15	85
SK	2.2	0.8	0.0	6	326	13	87
PT	2.0	0.7	0.1	7	543	58	42
NL	2.1	0.7	0.3	7	:	8	92
NO	2.1	0.8	0.2	10	379	32	68
FI	2.4	0.9	0.1	10	399	35	65
BE	1.7	0.8	0.5	13	230	27	73
SE	2.1	0.8	0.2	13	515	36	64
UK	1.5	1.0	0.5	15	121	59	41

The volume aspect was already discussed above in section 2.2. Regarding the level of educational programmes, the majority of FED activities take place at tertiary level (62%). In a number of countries, all of them with low or medium participation rates, the share of tertiary level programmes is above 80% (up to 92% in NL). By contrast, there are three countries where the majority of FED activities in adult age (25+) take place at below tertiary levels, namely Poland, Portugal and the UK.

There are no clear correlations between participations rates and volume or level of the FED activities, but there are clusters or types of countries. Low or medium participation rates tend to go along with higher instruction hours and restriction to tertiary level programmes (seen most clearly in EL, HU, AT). At the upper end of the comparative list, the group of countries with high participation rates in formal education does not show a consistent pattern at all. There is the distinct case of the UK where a large number of adults are involved in formal education, with low volumes of instruction and the majority of programmes below tertiary level. There is Belgium with relatively low instructions hours, though focus on tertiary level programmes. There are the Scandinavian countries where about one third of the programmes are below tertiary level and instruction hours at a medium position, except Sweden with a higher volume of instruction hours.

At this point of the analysis it is difficult to draw conclusions in terms of evaluating the different systems. Should countries with lower participation rates in formal education among adults move in the direction of countries with higher participation rates? The Scandinavian countries seem to provide the most inspiring model. Their high participation rates in formal adult education are clearly supported by policy measures, such as obligatory supply of adult learning centres at communal level, financial support of participants, and guaranteed return to the job after an educational leave (for Sweden see, for instance, Stenberg 2009). It is a “culture of second chance” which may well serve as a model for other countries.

For further analysis of the subject one would also have to look at problems of comparability. Countries may differ to some degree in classifying educational programmes as “formal” or “non-formal”. There is evidence that AES data in Sweden, for instance, classify some learning activities as “formal” which in other countries probably would be reported as non-formal education (e.g. Swedish for foreigners, labour market training via the employment office) (Löfgren and Svenning 2009). The same may hold for the UK. One may regard this as a methodological source of non-perfect comparability. On the other hand, it may reflect differences in the educational system which are real, incorporating a broader set of learning activities in the National Framework of Qualifications and thus giving a higher level of recognition to them.

Finally, formal education institutions may also play a role in non-formal education and training (NFE). Schools, colleges, universities etc. provide “regular” educational programmes as their main business, but they may in addition supply specific courses for other target groups. AES data include information about the provider of reported NFE activities. For 10% of all non-formal education and training activities the provider is a "formal education institution". The relative role of formal education institutions in NFE is the smallest in Sweden (4%) and Germany (5%). It is clearly above average in Greece, Belgium and Slovakia (15–17%), and in Belgium it arrives at a remarkable percentage of 38%.

One would like to have more information at this point, for instance, about the role of universities in the market of non-formal education and training. Unfortunately, the comparative data do not show the required detail (though national data sets may offer additional information). More specific studies are needed here.

3.2. Comparative issue (2): the role of enterprises in adult education and training

Enterprises play a prominent role in providing continuing training for adults. The European system of educational statistics comprises a separate survey on this issue, the "Continuing Vocational Training Survey" (CVTS) which will be conducted parallel to the AES every 5 years. Yet the AES has to play a role for this topic as well. Whereas the CVTS is based on a survey of enterprises (of at least 10 employees) and thus limited to training activities of enterprises, the AES has a comprehensive approach. It can show the contribution of enterprise-based training within the overall field of adult education. Moreover, AES data view this segment from the individual employee's perspective, in contrast to the employer's perspective reflected in the CVTS. The two surveys can complement each other in a fruitful way, provided they are sufficiently harmonized.[1]

The first task is to identify, in the AES data, which learning activities shall be classified as "company-based" or "employer-sponsored". Results were provided already in section 2.3 where employer-sponsored training is presented as type 3 within the sector typology of adult learning (tables 7 and 8). Here we discuss more closely matters of definition and content.

A basic point is that employer-sponsored or company-based training is conceptualized as a sub-category of non-formal education (NFE). This is not self-evident, and it is a restriction. The workplace is a major environment for informal learning as well. This may be learning-by-doing or it may be semi-organized in work teams, and providing opportunities for informal learning may be part of a company's training policy. Employees tend to report this as their most important source of learning while more formal learning arrangements (here covered as "non-formal learning") rank far below (Müller and Jakob, 2008, p.139). It is difficult, however, to include these forms of learning in a statistical concept which

[1] An improved concept of harmonizing the two surveys was discussed in the two parallel Eurostat Taskforces for AES and CVTS and will be reflected in Commission Regulations for the next round of surveys in 2011–2012.

tries to identify individual learning activities, as AES does.[1] Company based learning analysed here is restricted to learning activities that can be classified as "non-formal". Similarly, CVTS focuses on non-formal training in "courses" though some information on learning activities other than courses (as far as they are organized by the employer) is included.

Non-formal education, according to the AES Manual, requires "an organization providing structured arrangements (which must include a student-teacher-relationship), especially designed for education and learning". More specifically, the AES questionnaire addresses three kinds of organized learning which are relevant here: (1) courses, (2) workshops and seminars, (3) guided on-the-job training.

While it is evident that guided on-the-job training is a company-based learning activity, this does not apply to courses, seminars and workshops. Additional criteria must be sought, therefore, to identify the segment of company-based learning. AES data include three potential criteria:

(A) the provider of the learning activity is the employer
(B) the learning activity takes place during working time
(C) costs of the learning activity are paid for by the employer.

Table 11 shows for what percentages of all non-formal learning activities these characteristics apply. Figures do not represent participation rates, in this case, but refer to the structure of the total of NFE activities.

Employers are the leading providers of non-formal education and training activities.[2] 38% of the activities are directly provided by the employer. One can add to this another 9% for suppliers of the employer (which we assume covers mainly training for new equipment or software). Together they account for nearly one half of all NFE activities (column A). Beyond this, the employer may have contracted out training activities to professional providers. More comprehensive, therefore, is criterion (B): About 60% of all non-formal education and training activities take place fully or mostly during paid working hours. In about as many cases costs of the training activity ("for tuition, registration, exam fees") are fully or partly paid for by the employer (criterion C).

The conclusion is that we define a non-formal learning activity as being employer-based training ("sector 3") if either criterion B or criterion C applies.[3]

[1] AES questions on informal learning do cover learning at the workplace, but this is explicitly restricted to intentional learning ("to teach yourself anything at work or during your free time"), and informal learning at the workplace cannot be identified as such.

[2] See Boateng (2008) for more information on providers.

[3] Note that for the purpose of definition we do not explicitly include criterion (A) because country-specific institutional arrangements would need more specific investigation. Yet this is an important aspect. In a number of countries the employer's role as a direct provider of training is small even within the sector of employer-sponsored training. According to the data, this applies in particular to Hungary and Poland. In Hungary, the main providers are employers' organisations and (less so) trade unions. In Poland the main providers are (formal and non-formal) education and training institutions. More country-specific analysis is required to explore the institutional aspects.

Applying this definition, 70% of all non-formal learning activities are assigned to this sector (column D in table 11). We call it "employer-sponsored training" because this term indicates best the full range of what is covered. Taking place during paid working hours is the main criterion (contributing 59%), another 11% are contributed by including the criterion of financing. An advantage of this broad definition is that it comes close to the range of training activities covered by CVTS.

The essential finding is that non-formal education is largely work-related and employer-based. This is true in all countries. The minimum share of employer-sponsored training in non-formal education is found in Italy, Greece and Spain, where it amounts to about 50%. The percentage rises up to more than 80% in Norway and Sweden on the one hand, and Slovakia and Poland on the other hand.

Table 11. Defining "employer-sponsored training" – Characteristics of all reported NFE activities, in % AES 2007, population 25–64

	A	B	C	D
	Provider: Employer or supplier	Fully or mainly during working hours	Costs: Paid for by employer (fees etc.)	Sector 3: „Employer-sponsored training"
EU	47	59	61	70
IT	36	46	29	48
EL	50	48	41	55
ES	25	42	50	55
HU	4	58	42	62
AT	40	53	57	64
UK	50	63	62	66
FR	:	:	66	66
DE	56	62	54	69
FI	37	66	59	70
BE	51	71	65	75
NL	39	65	69	76
PT	49	74	73	78
SE	63	77	63	81
NO	:	67	80	84
SK	48	85	85	86
PL	27	86	87	87

Opportunities for adults to take part in organized learning processes are mainly bound to the work environment. As already shown in section 2.3 (table 8), employer-sponsored training is by far the largest sector of adult learning. On average in Europe, 23% of the adult population have participated in employer-sponsored training during the past year. However, country variations are enormous, ranging from 4% in Hungary to 59% in Sweden.

Findings may also be summarized by saying that the leading position of the Scandinavian countries in regard to the overall participation in adult education and training is not only, but mainly due to their well developed supply of work-related, employer-provided training opportunities. By contrast, the lack of employer-provided training in countries like Hungary and Greece is a severe deficiency in their development of opportunities for lifelong learning.

A next step of analysis could be to investigate which groups of the workforce are involved to a higher or lower degree in employer-provided training (for Germany see Rosenbladt and Bilger, 2008, pp. 82–85). Here we instead follow the guideline of "looking behind the figures". Which kind of learning activities are supplied in employer-sponsored training?

Learning activities are not necessarily organized in the form of "courses" or "seminars". Enterprise-based training activities can be planned and conducted "near to the workplace". This tends to facilitate access and to support motivation as well as transfer of what participants have learnt. A common term for such forms is "guided on-the-job training" (GOJT). This type of training is characterised "by planned periods of training, instruction or practical experience, using normal tools of work, either in the immediate place of work or in the work-situation or with the presence of a tutor. It is usually organised by the employer to facilitate adaptation of (new) staff, including transferred, re-hired and seasonal/temporary staff in their new or current jobs. It may include general training about the company (organisation, operating procedures, etc.) as well as specific job related instructions (safety and health hazards, working practices)." (AES Manual 2009, p.9). One may note that the boundary line between "non-formal education" and "informal learning" is certainly not clear-cut at this point. But it may be exactly this feature which makes the training effective. Experts regard this form of learning often superior to traditional classroom instructions.

AES data permit to define three sub-categories of employer-sponsored training:

Type 3a: Guided on-the-job training (GOJT)

Type 3b: Courses/seminars conducted during working hours

Type 3c: Courses/seminars outside working hours, but sponsored by the employer.

Findings are provided in table 12. Column 1 shows, as a reminder, the very different levels of participation in employer-sponsored training. Columns 2 to 5 change the perspective to show the structure within the sector, taking all reported activities of employer-sponsored training as the basis for percentages in the three sub-categories.

Table 12. Three main forms of employer-sponsored training ("sector 3") AES 2007, population 25–64

	participation rate in sector 3	all learning activities in sector 3	of which		
			Type 3a	Type 3b	Type 3c
	%	%	%	%	%
EU	23	100	39	47	15
HU	4	100	61	33	6
EL	7	100	55	33	12
IT	10	100	40	54	6
ES	15	100	34	42	24
PL	16	100	98	0	1
PT	18	100	80	15	5
FR	23	100	26	74*	:
BE	26	100	47	47	6
UK	26	100	68	28	5
AT	27	100	22	62	16
DE	31	100	21	69	10
NL	34	100	32	53	15
SK	37	100	92	6	2
FI	38	100	15	80	6
NO	45	100	24	55	21
SE	59	100	18	77	4

* France: Sum of type 3b and 3c (variable "during working time" is lacking).

The largest part of employer-sponsored training activities consists of courses, seminars or workshops conducted during working hours (47%, type 3b). Nearly as many are reported as "guided on-the-job training" (39%, type 3a). A smaller part consists of courses or seminars conducted outside working hours but fully or partly sponsored by the employer (15%, type 3c).

Regarding country variations one must be aware of the fact that the category "guided on-the-job training" is not self-explanatory as a response category in a survey, so that results will be affected by the wording in national questionnaires. Figures in table 12 provide empirical evidence of the problem. In some countries, nearly all employer-based training is reported as "guided on-the-job training": 98% in Poland, 92% in Slovakia, 61% in Hungary. This may be due to terminological traditions in ex-socialist countries. But there are also other countries where the GOJT category makes up for the majority of employer-sponsored training activities, such as Portugal (80%), UK (68%) and Greece (55%). In the majority of countries, by contrast, the most common form of training covers courses, seminars or workshops, while guided-training-on-the job is regarded as a specific form of instruction. This is clearly the case in Scandinavian countries as well as in Germany, Austria and France where the share of GOJT varies in a range of 15% to 26%.

Large country variations also apply to the third sub-category (type 3c), that is, training activities outside working hours but sponsored by the employer (type 3c). In a number of countries this form hardly plays a role, while in Norway (21%) and in Spain (24%) it seems to be a common way of organizing training activities. Data do not show exactly what kind and amount of financial contribution the employer makes. Nor do they show whether the training activity is initiated by the employer (but conducted outside working hours) or initiated by the employee (but sponsored by the employer). However, findings draw our attention to a point of interest, calling for additional information to explore the matter.

3.3. Comparative issue (3): social inclusion – participation of the lower educated in adult education and training

All studies on the subject show that participation in adult education and training is not equally distributed across social groups of the population, and that educational background and qualification are strong determinants to explain the inequality. The general pattern is that less educated persons participate less in adult learning. While this may be true in all countries, it does not apply to the same degree in all countries. Some countries are more successful than others in including the lower educated groups of the population in adult education and training.

AES data strongly confirm these findings (table 13). As before in table 10, figures for sub-groups are shown as relative participation rates, i.e. relative to the total. Three levels of educational attainment are distinguished, defined by ISCED levels 1–2, 3–4 and 5–6. This may not capture the stratification of educational levels very well in all countries but is sufficient for the present analysis. Figures refer to non-formal education and training; including formal education would even accentuate the pattern. The index may be regarded as a measure of inequality within countries regarding access to adult learning.

Table 13. Participation rate in non-formal education and training by highest level of education attained (ISCED levels) AES 2007, population 25–64

Country	Participation rate (%)	Relative participation rate (index, total = 1,0)		
	Total	Education levels 0–2	Education levels 3–4	Education levels 5–6
EU	33	0,5	1,0	1,6
HU	7	0,3	0,9	2,1
EL	13	0,3	1,0	2,2
PL	19	0,2	0,7	2,5
IT	20	0,4	1,3	2,3
PT	23	0,6	1,6	2,6
ES	27	0,6	1,1	1,6
BE	34	0,5	0,9	1,6
FR	34	0,6	1,0	1,6
AT	40	0,5	1,0	1,6
UK	40	0,7	1,1	1,3
SK	41	0,3	0,9	1,4
NL	42	0,6	1,0	1,4
DE	43	0,4	1,0	1,4
NO	51	0,7	1,0	1,3
FI	51	0,7	0,9	1,4
SE	69	0,8	1,0	1,2

In all countries the well-educated participate considerably more in adult learning than the population at large. Their relative index at EU level is 1.6, which means that the participation rate is about 60% above average. Correspondingly, in all countries the lower-educated participate much less in adult learning than the population at large. Their relative index at EU level is 0.5, which means that the participation rate is about 50% of the average.

Absolute values (not shown in the table) indicate an interesting pattern of inequality across European countries. Differences across countries are less pronounced among the higher educated than for the population at large (not regarding the extreme lower end, i.e. HU and EL). This indicates a “convergence of the elites”, while societies at large differ more strongly. Differences across countries are even more pronounced if you look at the lower-educated population. Participation rates here range from 2% in Hungary to 52% in Sweden. Inequality across Europe is stronger within the low-educated segment of populations than for the population at large.

The most marked inequality in access to adult education and training is found in Poland. The relative participation index ranges from 0.2 for the lower-educated to 2.5 for the well-educated. At the upper end, again, there is Sweden. Here, the relative participation index only ranges from 0.8 for the lower-educated to 1.2 for the well-educated. Looking at the whole list of countries there is a clear pattern, showing an inverse relationship between inequality and the overall level of participation in adult education and training. The higher the overall participation rate, the lower the inequality.

In policy terms one may turn it the other way round: social inclusion of the lower-educated is a prerequisite for reaching higher overall participation rates in adult education and training. Policy makers have recognized this and devoted special attention in their programmes to fostering participation of the lower-educated in adult education and training. This is true at EU level ("Action Plan on Adult Learning", 2007) as well as for policies at national level in individual countries, e.g. in Germany.

As mentioned, the relationship between educational background and participation in adult learning over the life-course has been widely noted and discussed in the research literature on adult education and training. A pronounced summary of findings reads like this: "Continuous Vocational Training thus usually does not compensate for a lack of initial training. During the life course there is no equalization in educational and training differences between individuals; differences rather increase, and this is true for all countries" (Müller and Kogan 2010, p. 268). One may not disagree with this. However, one may wish to arrive at a more specific insight. Country variations show that political, economic and cultural conditions can modify the general pattern to a great extent.

4. Concluding remarks

Adult education and training is a key element in lifelong learning. Increasing levels of participation are an objective of European educational policies. Statistical indicators have been set up to monitor the situation in individual countries as well as the development over time. They show that Europe is far from realising common achievements in this field. There are enormous differences in the level of participation in adult learning across European countries. Moreover, indicators show that progress in this field is not easily accomplished. Europe-wide, the level of participation in adult education and training has hardly increased in recent years.

The present paper argues that statistical indicators should not just monitor the development. They should contribute to understanding what is behind overall participation rates in adult learning, and what are the meanings and effects of different levels of participation as measured by overall indicators. To this end, data must be sufficiently detailed to enable more specific indicators and analysis. The newly established Adult Education Survey provides a data base that opens new prospects for comparative research along these lines.

The AES database as set up by Eurostat is accessible for research purposes (notwithstanding any technical problems). This is highly welcome. Recent research literature articulates needs that call for a comparative database: "Although research exploring the huge differences between countries in the extent of further training participation and lifelong learning is growing, more work is needed on the systematic explanation of these differences. Research needs not only long-term longitudinal observations, but also more specific measures (e.g., the general or vocational character or the curricula, field of study, or attributes of the educational institution)" (Müller and Jakob 2008, p.144 and p.162). The present paper outlined a number of specific measures and ways to analyse country profiles, trying to demonstrate the potential of the new statistical database for a better understanding of adult education and training across Europe.

REFERENCES

AES International Workshop (2009), 9–10 December 2009 in Berlin. Documentation, Munich: TNS Infratest.

BOATENG, SADIQ KWESI (2009): Significant country differences in adult learning. Eurostat, Statistics in Focus 44/2009.

COMMISSION OF THE EUROPEAN COMMUNITIES (2009): Progress towards the Lisbon objectives in education and training. Indicators and benchmarks 2009. Commission Staff Working Document, SEC(2009) 1616 final.

COMMISSION OF THE EUROPEAN COMMUNITIES (2007a): Delivering lifelong learning for knowledge, creativity and innovation. Draft 2008 joint progress report of the Council and the Commission on the implementation of the "Education & Training 2010 work program", {SEC(2007) 1484}.

COMMISSION OF THE EUROPEAN COMMUNITIES (2007b): Progress towards the Lisbon objectives in education and training. Indicators and benchmarks 2007. Commission Staff Working Document, SEC(2007) 1284.

COUNCIL OF THE EUROPEAN UNION (2009): Council conclusions on a strategic framework for European cooperation in education and training ("ET 2020"). Brussels, May 2009.

EUROSTAT (2006): Classification of learning activities – Manual. Luxembourg.

GNAHS, DIETER/ KUWAN, HELMUT/ SEIDEL, SABINE (ed.)(2008): Weiterbildungsverhalten in Deutschland. Band 2: Berichtskonzepte auf dem Prüfstand. Bielefeld: wbv.

LEMAN, STEVE (2009): Role of the AES in the United Kingdom. In: AES International Workshop (2009).

LÖFGREN, JOHAN/ SVENNING, ANDREAS (2009): Paper on the AES pilot. In: AES International Workshop (2009).

MÜLLER, WALTER/ KOGAN, IRENA (2010): Education. In: Immerfall, Stefan/ Therborn, Göran (ed.): Handbook of European Societies: Social Transformations in the 21st Century. New York: Springer, pp. 217–289.

MÜLLER, WALTER/ JAKOB, MARITA (2008): Qualifications and the Returns to Training Across the Life Course. In: Mayer, Karl-Ulrich/ Solga, Heike (ed.): Skill Formation. New York: Cambridge, pp. 126–172.

O'CONNELL, PHILIP/ JUNGBLUT, JEAN-MARIE (2008): What Do We Know about Training at Work? In: Mayer, Karl-Ulrich/ Solga, Heike (ed.): Skill Formation. New York: Cambridge, pp. 109–125.

OECD (2005): Education at a Glance. OECD Indicators 2005. Paris.

OECD (2006): Education at a Glance. OECD Indicators 2006. Paris.

OECD (2010): Education at a Glance. OECD Indicators 2010. Paris.

ROSENBLADT, BERNHARD VON/ BILGER, FRAUKE (2008): Weiterbildungsverhalten in Deutschland. Band 1. Berichtssystem Weiterbildung und Adult Education Survey 2007. Bielefeld: wbv.

ROSENBLADT, BERNHARD VON (2009): Adult education in comparative perspective – understanding differences across countries. Research paper, Munich: TNS Infratest.

SORVILLO, MARIA-PIA (2009): Adult Learning in the 2009 Progress Report. In: AES International Workshop (2009).

STENBERG, ANDERS (2009): Upgrading the Low Skilled: Is Public Provision of Formal Education a Sensible Policy? Research paper, Stockholm University.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 503—516

SMALL AREA ESTIMATION UNDER A MIXTURE MODEL

Hukum Chandra, HVL Bathla and U C Sud[1]

ABSTRACT

Small area estimation (SAE) under a linear mixed model may not be efficient if data contain substantial proportion of zeros than would be expected under standard model assumptions (hereafter zero-inflated data). We discuss the SAE for zero-inflated data under a mixture model (Fletcher *et al*., 2005 and Karlberg, 2000) that account for excess zeros in the data. Our results from simulation studies show that mixture model based approach for SAE works well and produces an efficient set of small area estimates. An application to real survey data from the National Sample Survey Organisation of India demonstrates the satisfactory performance of the approach.

Key Words*:* Linear mixed model, Small area estimation, EBLUP, Zero-inflated data, mixture model.

1. Introduction

In recent years, demand for reliable small areas statistics has greatly increased worldwide due to their growing use in formulating policies and programs, allocation of government funds, regional planning, and marketing decisions at local level. Sample surveys are usually planned to produce estimates for larger domains or areas and are therefore not appropriate to produce small area statistics due to small sample sizes. Due to cost and operational considerations, it is seldom possible to procure a large enough overall sample size to support direct estimates of adequate precision for all areas of interest. It is often necessary to employ indirect estimates for small areas that can increase the effective area sample size by borrowing strength from related areas through linking models, using census and administrative data and other auxiliary data associated with the small areas. The linear mixed models have been widely used in SAE. The empirical best linear unbiased predictor (EBLUP) is the most popular approach for estimation under these models, see Rao (2003). However, the EBLUP is model dependent and

[1] Division of Sample Survey, Indian Agricultural Statistics Research Institute, PUSA Campus, New Delhi-110012, India. Email: hchandra@iasri.res.in.

sensitive to model failure. To guard against model failure, the model-assisted approaches are proposed in the literature. For example, the Pseudo-EBLUP described in Prasad and Rao (1999) and the model-assisted empirical best predictor of Jiang and Lahiri (2006), hereafter JL-EBP. If linear mixed model is true, neither will be as efficient as the EBLUP. For the robustness, both estimators rely on the design consistency. Relying on a large sample property of a small sample statistic seems rather optimistic.

In practice, survey data often contain large proportion of zero values (for example, agricultural and environmental surveys etc.) than would be expected under standard model assumptions. Presence of excess zeros in the data makes the model assumptions invalid, see McCullagh and Nelder (1989). Consequently, problems with inference are liable to occur by ignoring this feature of the data. In classical regression literatures, mixture models which separately model the non-zero values and the occurrence of zero values are widely used to account for excess zeros in data, see, for example, Lambert (1992), Welsh *et al.* (1996) and Fletcher *et al.* (2005). In survey estimation Karlberg (2000) applied mixture model to estimate the population total for highly skewed data with many zeros. In the context of SAE, Chandra and Chambers (2006) and Chandra *et al.* (2007) observed that the EBLUP is ill-suited for small areas with large proportions of zeros. This indicates that standard methods of SAE under a linear mixed model may not be efficient for such data. In this article we explore the small area estimation under the mixture model for zero-inflated data. Following Fletcher *et al.* (2005) and Karlberg (2000), our approach for SAE works in three steps. First, a linear mixed model is fitted for positive values and then, in the second stage, a generalized linear mixed model (GLMM) is fitted for probability of positive values. Finally, two models are combined in estimation.

The structure of the paper is as follows. In the next section we first illustrate a linear mixed model and related estimators for small areas and then we introduce the mixture model and small area estimator under this model. In section 3 we present results from model-based as well as design-based simulation to evaluate the proposed approach of SAE, with the latter based on real survey data from Debt-Investment Survey 2002–03 of the National Sample Survey Organisation (NSSO) for the rural areas of the state of Uttar Pradesh in India. Finally, section 4 is devoted to concluding remarks and further research topics.

2. Estimation of Small Area Means

2.1. Small area estimation under a linear mixed model

Let U denote a population of size N and assume that population is partitioned into D small areas (or areas) $U_i (i = 1, ..., D)$. Let N_i and n_i is population and sample size respectively for area *i*. The total number of units in the population is

$N=\sum_{i=1}^{D}N_i$, with corresponding total sample size $n=\sum_{i=1}^{D}n_i$. Let $\mathbf{y}_i$ denote the N_i-vector of population values of a characteristic *Y* of interest and $\mathbf{x}_i$ denote the corresponding $N_i \times p$ matrix of population values in area *i*. Throughout, we use *i* to index the *D* small areas of interest, and *j* to index the distinct population units in these areas. We use *s* to denote the collection of units in a sample, with s_i the subset drawn from area *i*. Our aim is estimation of population mean of *Y* in small area *i*, i.e. $m_i = N_i^{-1}\sum_{j\in U_i} y_j$.

A commonly used class of models in small area inference is the class of linear mixed models. We consider the following linear mixed model for the distribution of $\mathbf{y}_i$ given $\mathbf{x}_i$ in area *i*:

$$\mathbf{y}_i = \mathbf{x}_i\boldsymbol{\beta} + \mathbf{g}_i\mathbf{u}_i + \mathbf{e}_i \tag{1}$$

where $\boldsymbol{\beta}$ is a $p\times 1$ vector of fixed effects, $\mathbf{g}_i$ is a $N_i \times q$ matrix of known covariates characterising differences between small areas, $\mathbf{u}_i$ is a q-vector of random area effects associated with area *i* and $\mathbf{e}_i$ is a $N_i \times 1$ vector of individual level random errors. The area specific effects $\{\mathbf{u}_i; i=1,\ldots,D\}$ are assumed to be independent and identically distributed realisations of a random vector of dimension *q* with zero mean and covariance matrix Σ_u. Similarly, the scalar individual effects making up $\mathbf{e}_i$ are assumed to be independent and identically distributed realisations of a random variable with zero mean and variance σ_e^2, with area and individual effects mutually independent. The covariance matrix of $\mathbf{y}_i$ is $\mathbf{v}_i = \sigma_e^2\mathbf{I}_{N_i} + \mathbf{g}_i\Sigma_u\mathbf{g}_i'$, depends on a vector of parameters $\theta = \left(\Sigma_u, \sigma_e^2\right)$ typically referred to as the variance components of (1). We assume that the sampling method used is uninformative given the values of the auxiliary variables, so the sample data also follow the population model (1).

By aggregating the area-specific models (1) over the *D*-small area, we are led to the population level linear mixed model

$$\mathbf{y}_U = \mathbf{x}_U\beta + \mathbf{g}_U\mathbf{u} + \mathbf{e}_U \tag{2}$$

where $\mathbf{y}_U = (\mathbf{y}_1', \ldots, \mathbf{y}_D')'$, $\mathbf{x}_U = (\mathbf{x}_1', \ldots, \mathbf{x}_D')'$, $\mathbf{g}_U = diag\{\mathbf{g}_i; 1\le i\le D\}$, $\mathbf{u}' = \left(\mathbf{u}_1', \ldots, \mathbf{u}_D'\right)$ and $\mathbf{e}_U = (\mathbf{e}_1', \ldots, \mathbf{e}_D')'$. Under (2), the covariance matrix of $\mathbf{y}_U$ is $\mathbf{v}_U = diag(\mathbf{v}_i; 1\le i\le D)$. Given a sample *s* of size *n* from this population, we can partition $\mathbf{v}_U = \begin{bmatrix} \mathbf{v}_{ss} & \mathbf{v}_{sr} \\ \mathbf{v}_{rs} & \mathbf{v}_{rr} \end{bmatrix}$ into their sample and non-sample components. Here, $r = U - s$ denotes the population units that are not in the sample. In particular, under (2) we have $\mathbf{v}_{ss} = diag\left\{\mathbf{v}_{iss}; i = 1, \ldots, D\right\} = diag\left\{\mathbf{g}_{is}\Sigma_u\, \mathbf{g}'_{is} + \sigma_e^2 \mathbf{I}_{is}; i = 1, \ldots, D\right\}$ and $\mathbf{v}_{sr} = diag\left\{\mathbf{v}_{isr}; i = 1, \ldots, D\right\} = diag\left\{\mathbf{g}_{is}\Sigma_u\, \mathbf{g}'_{ir}; i = 1, \ldots, D\right\}$. Here $\mathbf{g}_{is}$ and $\mathbf{g}_{ir}$ denote the restriction of $\mathbf{g}_i$ to sampled and non-sampled units in area *i* respectively. Given estimated values $\hat{\theta} = \left(\hat{\Sigma}_u, \hat{\sigma}_e^2\right)$ of the variance components we can substitute these to obtain estimates $\hat{\mathbf{v}}_{ss}$ and $\hat{\mathbf{v}}_{sr}$ of $\mathbf{v}_{ss}$ and $\mathbf{v}_{sr}$ respectively.

Under (2), the EBLUP for small area *i* mean of *Y* is (Rao, 2003, section 6.2.3)

$$\begin{aligned} \hat{m}_i^{EBLUP} &= \hat{E}\left\{m_i \left| \mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir} \right.\right\} \\ &= N_i^{-1}\left[\sum\nolimits_{j \in s_i} y_j + \mathbf{1}'_{ir}\left\{\mathbf{x}_{ir}\hat{\beta} + \hat{\mathbf{v}}_{irs}\hat{\mathbf{v}}_{iss}^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\right] \\ &= N_i^{-1}\left[n_i \bar{y}_{is} + (N_i - n_i)\left\{\bar{\mathbf{x}}'_{ir}\hat{\beta} + \bar{\mathbf{g}}'_{ir}\hat{\Sigma}_u \mathbf{g}'_{is}\left(\mathbf{g}_{is}\hat{\Sigma}_u \mathbf{g}'_{is} + \hat{\sigma}_e^2 \mathbf{I}_{is}\right)^{-1}(\mathbf{y}_{is} - \mathbf{x}_{is}\hat{\beta})\right\}\right]. \end{aligned} \tag{3}$$

Here $\hat{E}$ denotes the expectation operator under (2) with unknown parameters replaced by estimates, $\mathbf{x}_{is}$ and $\mathbf{x}_{ir}$ are the matrices of sample and non-sample values of **X** in area *i*, $\mathbf{y}_{is}$ is the vector of sample values of *Y* in the same area, $\hat{\beta}$ is the 'empirical' BLUE of β, $\hat{\mathbf{v}}_{irs}$ is the transpose of the estimated value of $\mathbf{v}_{isr}$ with $\hat{\mathbf{v}}_{iss}$ the corresponding estimate of $\mathbf{v}_{iss}$, and $\mathbf{1}_{ir}$ is a vector of ones of length $N_i - n_i$. Note that the estimator (3) is model dependent and works well under (1).

Two alternative approaches in the literature under linear mixed model (1) are the pseudo-EBLUP (Rao, 2003, section 7.2.7) and the estimator of Jiang and Lahiri (2006). Recollect from (3) that the EBLUP is defined by replacing the unknown area *i* mean m_i by an estimate of its expected value given the observed sample values of *Y* in area *i* and the area *i* values of **X**. Let π_{ij} denote the sample inclusion probability of population unit *j* in small area *i*. The pseudo-EBLUP is then defined by replacing m_i by an estimate of its expected value given the value of its design-consistent estimate

$$\hat{m}_i^{\pi} = \left(\sum\nolimits_{j \in s_i} \pi_{ij}^{-1}\right)^{-1} \sum\nolimits_{j \in s_i} \pi_{ij}^{-1} y_{ij} = \sum\nolimits_{j \in s_i} \tilde{w}_{ij} y_{ij} \tag{4}$$

and the area *i* values of **X**. That is, under (1) the pseudo-EBLUP of m_i is (Rao, 2003, section 7.2.7)

$$\begin{aligned}\hat{m}_i^{psuedoEBLUP} &= \hat{E}\left\{m_i \left|\hat{m}_i^{\pi}, \mathbf{x}_{is}, \mathbf{x}_{ir}\right.\right\} \\ &= \bar{\mathbf{x}}_i'\hat{\beta}_{\tilde{w}} + \left(\bar{\mathbf{g}}_i'\hat{\Sigma}_{u\tilde{w}}\bar{\mathbf{g}}_{i\tilde{w}}\right)\left(\bar{\mathbf{g}}_{i\tilde{w}}'\hat{\Sigma}_{u\tilde{w}}\bar{\mathbf{g}}_{i\tilde{w}} + \hat{\sigma}_{e\tilde{w}}^2\sum\nolimits_{j\in s_i}\tilde{w}_{ij}^2\right)^{-1}(\hat{m}_i^{\pi} - \bar{\mathbf{x}}_{i\tilde{w}}'\hat{\beta}_{\tilde{w}})\end{aligned} \tag{5}$$

where $\hat{\beta}_{\tilde{w}}$, $\hat{\Sigma}_{u\tilde{w}}$ and $\hat{\sigma}_{e\tilde{w}}^2$ are pseudo-maximum likelihood estimates based on the weights $\tilde{w}_{ij}$ and $\bar{\mathbf{g}}_{i\tilde{w}}$ and $\bar{\mathbf{x}}_{i\tilde{w}}'$ are design-consistent estimates of $\bar{\mathbf{g}}_i$ and $\bar{\mathbf{x}}_i$ that are defined in exactly the same way as $\hat{m}_i^{\pi}$ above. Under the same model the Jiang and Lahiri (2006) approach leads to an estimator that is also defined by conditioning on the value of $\hat{m}_i^{\pi}$,

$$\begin{aligned}\hat{m}_i^{JL} &= \sum\nolimits_{j\in s_i}\tilde{w}_{ij}\hat{E}\left\{\hat{E}\left(y_{ij}\left|\mathbf{x}_{ij}, \mathbf{u}_i\right.\right)\left|\hat{m}_i^{\pi}, \mathbf{x}_i\right.\right\} \\ &= \bar{\mathbf{x}}_{i\tilde{w}}'\hat{\beta} + \left\{\tilde{\mathbf{w}}_{is}'\left(\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}_{is}' + \hat{\sigma}_e^2\mathbf{I}_{is}\right)\tilde{\mathbf{w}}_{is}\right\}^{-1}\left\{\tilde{\mathbf{w}}_{is}'\mathbf{g}_{is}\hat{\Sigma}_u\mathbf{g}'\tilde{\mathbf{w}}_{is}\right\}\left(\hat{m}_i^{\pi} - \bar{\mathbf{x}}_{i\tilde{w}}'\hat{\beta}\right)\end{aligned} \tag{6}$$

where $\tilde{\mathbf{w}}_{is}$ is the vector of sample weights $\tilde{w}_{ij}$ in area *i*. Note that in (6) we use optimal (i.e. maximum likelihood (ML) or restricted maximum likelihood (REML)) estimates for model parameters. Both (5) and (6) are essentially motivated by the idea of estimating the area *i* mean by its conditional expectation under (1) given the value of the usual design-consistent estimator (4) for this quantity. Under (1), neither will be as efficient as the EBLUP.

2.2.Small area estimation under a mixture model

Let us consider that survey variable *Y* is zero-inflated. We then introduce a mixture model to accommodate excess zeros in the data. In the dual regime model described in section 1, following the ideas of Fletcher *et al.* (2005) and Karlberg (2000), the variable of interest *Y* is expressed as product:

$$\mathbf{y}_i = \mathbf{y}_i^*\boldsymbol{\delta}_i \tag{7}$$

where $\mathbf{y}_i^*$ is the linear component and assume to follow a linear mixed model, like (1) and $\boldsymbol{\delta}_i = I\left(\mathbf{y}_i > 0\right)$, is a binary (0/1) variable, assume to follow a generalized linear mixed model (GLMM) with logit link function (Breslow and Clayton, 1993), i.e. logistic linear mixed model, referred as the logistic component of (7). We assume that δ_j given $\mathbf{x}_j$ are independent Bernoulli random variables with probability $p_j = \mathrm{P}(y_j > 0) = \mathrm{P}(\delta_j = 1)$. With these notations, the model linking the probability of positive values with the covariates is the logistic linear mixed model in small area *i* of the form (Manteiga *et al.*, 2007):

$$\log it(\mathbf{p}_i) = \ln\{\mathbf{p}_i/(1-\mathbf{p}_i)\} = \boldsymbol{\eta}_i = \mathbf{x}_i\boldsymbol{\alpha} + \mathbf{g}_i\mathbf{b}_i \ (i = 1,...,D) \qquad (8)$$

with $\mathbf{p}_i = \exp(\boldsymbol{\eta}_i)\{1+\exp(\boldsymbol{\eta}_i)\}^{-1} = \exp(\mathbf{x}_i\boldsymbol{\alpha}+\mathbf{g}_i\mathbf{b}_i)\{1+\exp(\mathbf{x}_i\boldsymbol{\alpha}+\mathbf{g}_i\mathbf{b}_i)\}^{-1}$. Here $\boldsymbol{\alpha}$ is a vector of unknown fixed effects parameters and $\mathbf{b}_i$ $(i = 1,...,D)$ is the random area effect associated with area *i* which is assumed to be normal with zero mean and constant variance.

For estimation of parameters for linear-component, we denote by $s_+ = \{j \in s, y_j > 0\}$ the subset of the sample for which the survey variable is non-zeros, and $n_+ = \sum_{j\in s}\delta_j$ denotes the number of non-zeros sample units. As below (2), we denote by $\mathbf{y}^*_{s_+}, \mathbf{x}_{s_+}$, $\mathbf{g}_{s_+}$ and $\mathbf{V}_{s_+s_+}$ the corresponding vector and matrices related to non-zeros survey variable values of the sample. Accordingly, at area level we use similar notation by introducing an extra subscript *i*. Assuming model (1), the empirical-BLUE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{D}\mathbf{x}'_{is_+}\hat{\mathbf{V}}^{-1}_{iss_+}\mathbf{x}_{is_+}\right)^{-1}\left(\sum_{i=1}^{D}\mathbf{x}'_{is_+}\hat{\mathbf{V}}^{-1}_{iss_+}\mathbf{Y}^*_{is_+}\right)$ with $E(\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{\delta}_j) = \boldsymbol{\beta}$ and $V(\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{\delta}_j) = \left(\sum_{i=1}^{D}\mathbf{x}'_{is_+}\hat{\mathbf{V}}^{-1}_{iss_+}\mathbf{x}_{is_+}\right)^{-1}$. The expectation and variance are under the model. Under (1), the predicted values for the linear component of (7) are (Henderson, 1953):

$$\hat{E}(y^*_j) = \hat{\mu}_j = \mathbf{x}_j\hat{\boldsymbol{\beta}} + \mathbf{g}_j\hat{u}_i;\ j \in i\text{, with } \hat{\mathbf{u}}_i = \hat{\Sigma}_u\boldsymbol{\beta}\mathbf{g}'_{is_+}\hat{\mathbf{v}}^{-1}_{is_+s_+}(\mathbf{y}_{is_+} - \mathbf{x}'_{is_+}\hat{\ }). \qquad (9)$$

For the estimation of unknown parameters of logistic components in (8), we used an iterative procedure that combines the Penalized Quasi-Likelihood (PQL) estimation of $\boldsymbol{\alpha}$ and $\mathbf{b}_i$ with REML estimation of variance component parameters as described in Saei and Chambers (2003) and Manteiga *et al.* (2007). The predicted probabilities of the logistic component of (7) are:

$$\hat{p}_j = \exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}} + \mathbf{g}_j\hat{\mathbf{b}}_i)\left\{1+\exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}} + \mathbf{g}_j\hat{\mathbf{b}}_i)\right\}^{-1};\ j \in i. \qquad (10)$$

Using results from appendix and collecting (10) and (11), an approximately model-unbiased estimate of $E(y_j)$ is

$$\hat{E}(y_j\,|\,\mathbf{x}_j, \mathbf{g}_j, i) = \boldsymbol{\beta}\hat{\vartheta}_j = \left\{\mathbf{x}_j\hat{\boldsymbol{\beta}} + \mathbf{g}_j\hat{\Sigma}_u\mathbf{g}'_{is_+}\hat{\mathbf{v}}^{-1}_{is_+s_+}(\mathbf{y}_{is_+} - \mathbf{x}'_{is_+}\hat{\ })\right\}\cdot$$

$$\left\{ \mathrm{e}^{\mathbf{x}_j\hat{\boldsymbol{\alpha}}+\mathbf{g}_j\hat{\mathbf{b}}_i}\left(1+\mathrm{e}^{\mathbf{x}_j\hat{\boldsymbol{\alpha}}+\mathbf{g}_j\hat{\mathbf{b}}_i}\right)^{-1}\right\}. \tag{11}$$

Consequently using (11) the estimator for population mean of *Y* in area *i* is

$$\hat{m}_i^{mix} = \hat{E}\left\{m_i \left| \mathbf{y}_{is}, \mathbf{x}_{is}, \mathbf{x}_{ir}\right.\right\} = N_i^{-1}\left[\sum\nolimits_{j\in s_i} y_j + \sum\nolimits_{j\in r_i} \hat{y}_j\right] = N_i^{-1}\left[\mathbf{1}'_{is}\mathbf{y}_{is} + \mathbf{1}'_{ir}\hat{\boldsymbol{\theta}}_{ir}\right] \tag{12}$$

Besides (10), we also consider the estimated probabilities under a logistic linear model, with no area effect in (8):

$$\hat{p}_j = \exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}})\left\{1+\exp(\mathbf{x}_j\hat{\boldsymbol{\alpha}})\right\}^{-1};\, j\in i \tag{13}$$

3. Empirical Evaluations

In this section we present results from two simulation studies that were used to contrast the performance of different small area estimators set out in Table 1 and described in section 2. The first is a model based simulation in which small area population and sample data were simulation under the model. The second is a design based simulation in which a fixed population containing number of small areas was repeatedly sampled, holding the sample size in each small area fixed.

The performance of different small area estimators were evaluated with respect to two basic criteria–the relative bias and the relative root mean squared error, both expressed as percentages, of area mean estimates. The bias was measured as

$$\%AvRB = \underset{i}{mean}\left|\left\{M_i^{-1}\left(K^{-1}\sum\nolimits_{k=1}^{K}\hat{m}_{ik}\right)-1\right\}\right|\times 100$$

Note that the subscript of *k* here indexes the *K* simulations, with m_{ik} denoting the value of the small area *i* mean in simulation *k* and $\hat{m}_{ik}$ denoting the area *i* estimated value in simulation *k*. The actual area *i* mean value (the average over the simulations) is denoted by $M_i = K^{-1}\sum\nolimits_{k=1}^{K} m_{ik}$. The root mean squared error was measured as

$$\%AvR\mathrm{R}MSE = \underset{i}{mean}\left[M_i^{-1}\left\{\sqrt{K^{-1}\sum\nolimits_{k=1}^{K}\left(\hat{m}_{ik}-m_{ik}\right)^2}\right\}\right]\times 100\,.$$

3.1.Model-Based Simulations

In these simulations we fixed population size $N = 15{,}000$ and number of small areas *D* = 30. Population sizes in the small areas were uniformly distributed over the interval [443, 542] and were kept fixed over simulations. In each simulation we generated *N* population values of auxiliary variable x from the chi-square distribution with 10 degrees of freedom. For each area, population values for y were generated under the two-level model $y_{ij} = 10 + 2x_{ij} + u_i + e_{ij}$ $\left(j = 1,...,N_i; i = 1,..,30 \right)$. The area-specific random effects u_i and individual random effects e_{ij} were independently drawn from $N(0, \sigma_u^2 = 4)$ and $N(0, \sigma_e^2 = 16)$ distributions respectively. Zero-inflated population values for *y* were generated by first generating Bernoulli (0/1) random variable with fixed probability *p* (i.e. proportion of non-zero values in the data) and multiplying them to y for each area. A sample of size $n = 600$ was then selected from the simulated population, with area sample sizes proportional to the fixed area populations. Sampling was via stratified random sampling, with the strata defined by the small areas with average sample size of 25. All population values independently regenerated at each simulation and an independent sample drawn from the population each time. A total of *K* = 1000 simulations were carried out. We used simulations scenario with different proportion of non-zero values *p* = 0.90, 0.75, 0.65 and 0.50 for all small areas.

The average relative bias and the average relative root mean squared error over these simulations for different estimators are presented in Table 2. These results show an increase in biases and RMSEs for all estimators with increase in proportion of zeros in the data. We noted relatively unstable performance of the direct estimator (DIR), so it is not an option for SAE. Hereafter we do not discuss DIR. Among the linear mixed model based estimators, in terms of biases and RMSEs, the EBLUP is better overall. Both the EBLUP and Pseudo have smaller biases and RMSE compared to JL. However, EBLUP is either dominating marginally or at par with Pseudo. In contrast, among mixture model based estimators, MIX1 and MIX2 are almost identical. The area-specific relative bias and relative RMSE for p=0.65 are shown in Figure 1. The results in Figure 1 reflect the consistently better performance of the proposed approach. Similar conclusions are also true for other values of *p*. Overall, mixture model based approach is superior than the standard linear mixed model based estimators for zero-inflated data (see Table 1 and Figure 1).

3.2.Design-Based Simulations

In these simulations we used the data of Debt-Investment Survey 2002–03 for rural areas of the state of Uttar Pradesh in India conducted by National Sample Survey Organisation in the year 2002–03. In the original survey there were 11,814 sample households (with population of 22,145,951) spread across 69 districts of Uttar Pradesh that participated in the Survey. However, in our simulation studies (to avoid computation difficulties arising out of computer memory space) we used the sample of 1693 households from $D = 10$ districts only and further divided the survey weights by 10 to reduce the overall population size. This sample of 1693 households was bootstrapped to create a realistic population of $N = 327,481$ households by re-sampling with replacement with probability proportional to a household's sample weight. A total of $K = 1000$ independent stratified random samples were then drawn from this bootstrap population, with total sample size equal to that of the original sample and with districts defining the strata. Sample sizes within districts were the same as in the original sample. Districts were the small area of interest. The Y variable of interest was amount of loan outstanding per household (with 43 percent zero values in the original sample) and the auxiliary variable X was land owned by household. The aim was to predict the district level average value of amount of loan outstanding per household.

Table 3 set out the districts-wide as well as average relative biases and relative RMSEs of different estimators based on the 1000 repeated independent stratified samples. We note that the districts have different proportions of non-zeros in the data which gives a realistic scenario to apply our approach. These results show that the mixture model based MIX2 estimator for small area mean is severely biased and relatively unstable. The GLM (13) used to estimate probability for occurrence of zero in MIX2 is not fitting well since it does not capture the area effects. In contrast, average bias and RMSE of the MIX1 is consistently smaller than linear mixed model based EBLUP, Pseudo and JL estimators for small mean. Further, among linear mixed model based estimator, in terms of average bias and RMSE, the EBLUP performs better than JL which dominates the Pseudo. Furthermore, district-wide results reveal that in most of the districts MIX1 have smaller biases and RMSEs than alternative estimators. Overall the results in Table 3 clearly indicate that the mixture model based estimation is working well and better than the linear mixed model estimation if data contain substantial proportion of zeros.

4. Conclusions

The results set out in section 3 conclude that commonly used methods of SAE under a linear mixed model lead to biased and unstable estimates for small area with many zeros. In this case, proposed mixture model based SAE takes care of excess zeros and provides more efficient sets of small area estimates. An application to real survey data from the NSSO too shows satisfactory performance of the proposed approach. In this article we have not addressed the mean squared error (MSE) estimation for different estimator for small area means described in sections 2. The MSE of various estimators under the linear mixed model (EBLUP, Pseudo-EBLUP and JL) can be estimated via their analytical MSE expression already available in the literatures, see, for example, Rao (2003) and Jiang and Lahiri (2006). However, the MSE of the proposed estimators under mixture model still need to be developed. Alternatively, re-sampling based methods like Jacknife or bootstrap methods can be explored. Authors are currently working on the MSE estimation.

Appendix

From (7) we see that

$$\begin{aligned} E(y_j) &= \Pr(\delta_j = 1)E(y_j \mid \delta_j = 1) + \Pr(\delta_j = 0)E(y_j \mid \delta_j = 0) \\ &= \Pr(\delta_j = 1)E(y_j \mid \delta_j = 1) = p_j\mu_j \end{aligned}$$

where $\mu_j = E(y_j \mid \delta_j = 1)$. This leads to $\hat{E}(y_j) = \hat{p}_j\hat{\mu}_j = \hat{\theta}_j$. Assuming that $\hat{\mu}_j$ and $\hat{p}_j$ are uncorrelated (see Karlberg [7]) leads to

$$\begin{aligned} E(\hat{\theta}_j) &= E(\hat{p}_j\hat{\mu}_j) = E\left\{E(\hat{p}_j\hat{\mu}_j \mid \delta_j)\right\} \\ &= E\left\{\hat{p}_j E(\hat{\mu}_j \mid \delta_j)\right\} = E(\hat{p}_j)E\left\{E(\hat{\mu}_j \mid \delta_j)\right\} \\ &= p_j\mu_j = E(y_j). \end{aligned}$$

Table 1. Estimators evaluated in simulation studies

Estimators	Description
DIR	Direct estimation
Under linear mixed model (1)	
EBLUP	EBLUP (3)
Pseudo	Pseudo-EBLUP (5)
JL	JL-EBP (6)
Under mixture model (7)	
MIX1	Estimator (12) with estimated probabilities (10)
MIX2	Estimator (12) with estimated probabilities (13)

Table 2. Percentage average relative bias (*%AvRB*) and percentage average relative RMSE (*%AvRRMSE*) of different estimators under model based simulations

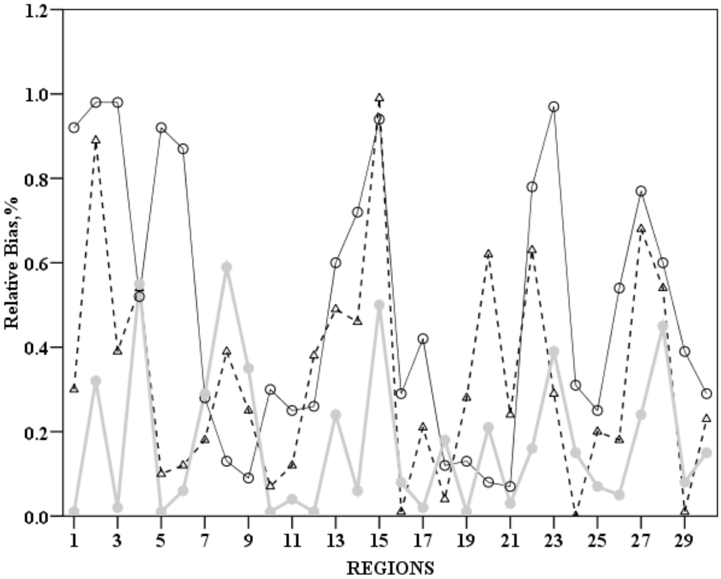
	p			
Estimators	0.90	0.75	0.65	0.50
	%AvRB			
DIR	0.48	0.54	0.61	0.90
EBLUP	0.24	0.28	0.33	0.37
Pseudo	0.25	0.30	0.33	0.39
JL	0.41	0.44	0.49	0.51
MIX1	0.18	0.16	0.21	0.30
MIX2	0.17	0.17	0.18	0.29
	%AvRRMSE			
DIR	10.74	15.14	18.00	24.40
EBLUP	5.78	7.26	8.02	9.72
Pseudo	5.87	7.32	8.06	9.68
JL	8.67	9.77	10.39	11.77
MIX1	3.91	5.19	6.03	8.04
MIX2	3.62	4.73	5.50	7.25

Table 3. Districts-wide (and average) percentage relative bias and percentage relative RMSE of different estimators under design based simulations

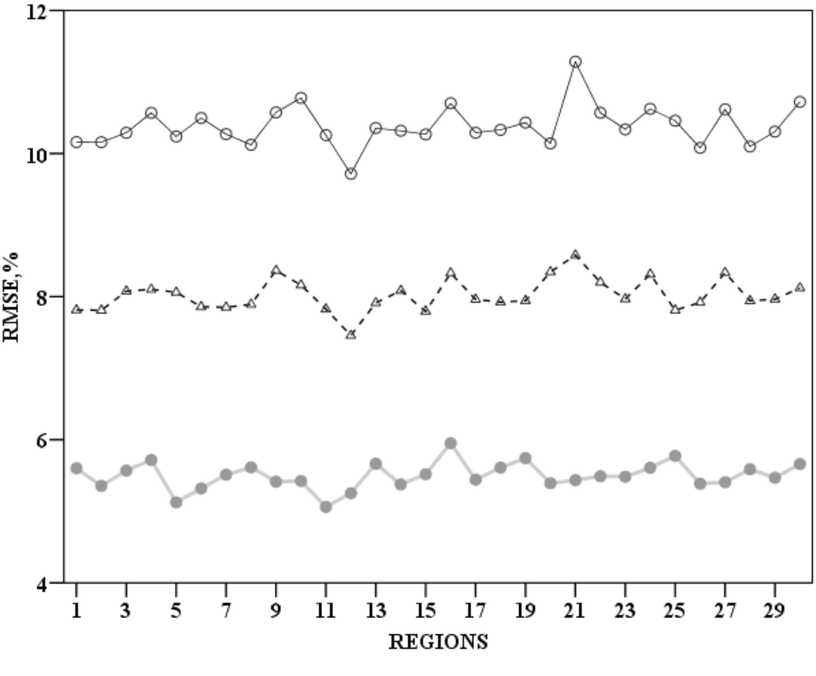
Districts	*p*	EBLUP	Pseudo	JL	MIX1	MIX2
% Relative bias						
1	0.59	4.7	6.6	4.9	4.7	21.6
2	0.62	4.0	5.0	4.0	4.0	24.3
3	0.42	5.8	6.4	5.5	0.3	10.9
4	0.46	9.1	7.4	9.0	2.7	5.6
5	0.50	3.3	3.2	3.2	5.0	10.5
6	0.31	30.0	33.4	30.0	19.4	67.1
7	0.44	5.7	8.0	6.0	15.1	21.9
8	0.52	19.1	17.0	19.2	18.0	24.5
9	0.27	2.4	5.8	2.5	7.0	47.8
10	0.49	0.7	1.4	0.4	5.4	2.7
Average	0.46	8.5	9.4	8.5	8.2	23.7
% Relative RMSE						
1	0.59	16.5	17.9	17.2	16.0	24.5
2	0.62	16.6	17.0	17.3	15.6	26.4
3	0.42	15.6	17.3	15.9	14.5	18.7
4	0.46	15.7	17.1	15.8	14.3	14.6
5	0.50	21.0	22.2	22.4	19.8	20.1
6	0.31	42.1	46.2	43.2	31.8	75.6
7	0.44	15.0	17.4	15.1	20.8	25.4
8	0.52	29.3	28.9	29.1	28.7	30.6
9	0.27	24.3	26.9	24.3	26.9	60.8
10	0.49	12.0	13.7	11.7	12.0	11.2
Average	0.46	20.8	22.4	21.2	20.0	30.8

Figure 1. Region-specific performance measures of the EBLUP (dashed line,Δ), JL (thin line, O) and MIX2 (solid line,●) for p=0.65 under model based simulations.

Relative Bias



Relative RMSE



REFERENCES

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistics Association,* **88**, 9–25.

CHANDRA, H. and CHAMBERS, R. (2006). Multipurpose Weighting for Small Area Estimation. *Journal of Official Statistics*, accepted for publication.

CHANDRA, H., SALVATI, N. and CHAMBERS, R. (2007) Small Area Estimation for Spatially Correlated Populations. A Comparison of Direct and Indirect Model-Based Methods. *Statistics in Transition,* **8**, pp. 887–906.

FLETCHER, D., MACKENZIE, D. and VILLOUTA, E. (2005). Modelling Skewed Data With Many Zeros: A Simple Approach Combining Ordinary and Logistic Regression. *Journal of Environmental and Ecological Statistics,* **12** (1), 45–54.

HENDERSON, C.R (1953). Estimation of Variance and Covariance Components. *Biometrics,* **9**, 226–252.

JIANG, J. and LAHIRI, P. (2006). Estimation of Finite Population Domain Means: A Model-Assisted Empirical Best Prediction Approach. *Journal of the American Statistical Association*, **101**, 301–311.

KARLBERG, F. (2000). Survey Estimation for Highly Skewed Populations in the Presence of Zeroes. *Journal of Official Statistics*, **16**, 229–241.

LAMBERT, D. (1992). Zero-Inflated Poisson Regression, With An Application To Defects in Manufacturing. *Technometrics*, 34, 1–14.

MCCULLAGH, P. and NELDER, J.A. (1989) *Generalized Linear Models* (New York: Chapman and Hall)

MANTEIGA, G.W., LOMBARDÌA, M.J., MOLINA, I., MORALES, D., and SANTAMARÌA, L. (2007). Estimation of the Mean Squared Error of Predictors of Small Area Linear Parameters under a Logistic Mixed Model. *Computational Statistics & Data Analysis*, **51**(5): 2720–2733.

PRASAD, N.G.N. and RAO, J.N.K. (1999). On Robust Small Area Estimation Using a Simple Random Effects Model. *Survey Methodology*, **25**, 67–72.

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

SAEI, A. and CHAMBERS, R. (2003). Small Area Estimation under Linear and Generalized Linear Mixed Models with Time and Area Effects. Methodology Working Paper No. M03/15. University of Southampton, UK.

WELSH, A.H., CUNNINGHAM, R.B., DONNELLY, C.F., and LINDENMAYER, D.B. (1996). Modelling the Abundance of Rare Species: Statistical Models for Counts with Extra Zeros. *Ecological Modelling*, 88, 297–308.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 517—538

FITTING GENERAL LINEAR MODEL FOR LONGITUDINAL SURVEY DATA UNDER INFORMATIVE SAMPLING

Abdulhakeem A. H. Eideh[1]

ABSTRACT

The purpose of this article is to account for informative sampling in fitting superpopulation model for multivariate observations, and in particular multivariate normal distribution, for longitudinal survey data. The idea behind the proposed approach is to extract the model holding for the sample data as a function of the model in the population and the first order inclusion probabilities, and then fit the sample model using maximum likelihood, pseudo maximum likelihood and estimating equations methods. As an application of the results, we fit the general linear model for longitudinal survey data under informative sampling using different covariance structures: the exponential correlation model, the uniform correlation model, and the random effect model, and using different conditional expectations of first order inclusion probabilities given the study variable. The main feature of the present estimators is their behaviours in terms of the informativeness parameters.

Key words: General Linear Model, Informative sampling, Longitudinal Survey Data, Maximum Likelihood , and Sample distribution.

1. Introduction

Sampling designs for surveys are often complex and informative, in the sense that the selection probabilities are correlated with the variables of interest, even when conditioned on explanatory variables. In this case conventional analysis that disregards the informativeness can be seriously biased, since the sample distribution differs from that of the population. Most of the studies in social surveys are based on data collected from complex sampling designs. Standard analysis of survey data often fails to account for the complex nature of the

[1] On Sabbatical Leave at Palestine Polytechnic University, Hebron, Palestine. *Address for correspondence*: Dr. ABDULHAKEEM A.H. EIDEH, Department of Mathematics, College of Science and Technology, Al-Quds University, Abu-Dies campus, Palestine, P.O. Box 20002, Jerusalem. E-mail: msabdul@science.alquds.edu.

sampling design such as the use of unequal selection probabilities, clustering, and post-stratification. The effect of the sample design on the analysis is due to the fact that the models in use typically do not incorporate all the design variables determining the sample selection, either because there may be too many of them or because they are not of substantive interest. However, if the sampling design is informative in the sense that the outcome variable (variable of interest) is correlated with the design variables not included in the model, even after conditioning on the model covariates, standard estimates of the model parameters can be severely biased, leading possibly to false inference. Pfeffermann (1993, 1996) reviews many examples reported in the literature that illustrate the effects of ignoring the sampling process when fitting models to survey data and discusses methods that have been proposed to deal with this problem, see also Skinner, Holt, and Smith (1989), Kasprzyk, Duncan, Kalton and Singh (1989), Hoem, (1989), and Chambers and Skinner (2003). It should be emphasized that standard inference may be biased even when the original sample is a simple random sample, due to non-response, attrition and imperfect frames that results in de facto a posterior differential inclusion probabilities.

To overcome the difficulties associated with the use of classical inference procedures for cross sectional survey data, Pfeffermann, Krieger and Rinott (1998) proposed the use of the sample distribution induced by the assumed population models, under informative sampling, and developed expressions for its calculation. Similarly, Eideh and Nathan (2006) fitted time series models for longitudinal survey data under informative sampling. Furthermore, Eideh (2008) fitted random effects or subject-specific effects models for analyzing normal data, which are assumed to be correlated, under the concept of informative sampling.

The plan of this paper is as follows. In Section 2 we define sample distribution and sample likelihood. In Section 3 we extract the sampled distribution of the multivariate normal distribution under informative sampling. In Section 4 we fit the general linear model for longitudinal survey data. Section 5 provides a discussion of the results.

2. Sample distribution and sample likelihood

Let $U = \{1,...,N\}$ denote a finite population consisting of N units. Let y be the target or study variable of interest and let y_i be the value of y for the ith population unit. Let x_i, $i \in U$ be the value of an auxiliary variable(s), x, and $\mathbf{z} = \{z_1,...,z_N\}$ be the values of a known design variable, used for the sample selection process but not included in the working model under consideration. In what follows we consider a sampling design with selection probabilities $\pi_i = \Pr(i \in s) > 0$, and sampling weight $w_i = 1/\pi_i$; $i = 1,...,N$. In practice, the π_i's may depend on the population values (x,y,z). We express this by writing: $\pi_i = \Pr(i \in s \mid x,y,z)$. The sample s consists of the subset of U selected at random by the sampling scheme with inclusion probabilities $\pi_1,...,\pi_N$. Denote by $\mathbf{I} = (I_1,...,I_N)'$ the N by 1 sample indicator (vector) variable such that $I_i = 1$ if unit $i \in U$ is selected to the sample and $I_i = 0$ if otherwise. The sample s is defined accordingly as $s = \{i \mid i \in U, I_i = 1\}$ and its complement by $c = \bar{s} = \{i \mid i \in U, I_i = 0\}$. We assume probability sampling, so that $\pi_i = \Pr(i \in s) > 0$ for all units $i \in U$.

We now consider the population values $y_1,...,y_N$ as random variables, which are independent realizations from a distribution with probability density function $f_p(y_i \mid x_i, \theta)$, indexed by a vector parameter θ.

According to Krieger and Pfeffermann (1997), the (marginal) sample probability density function of y_i is defined as:

$$\begin{aligned} f_s(y_i \mid x_i,\theta,\gamma) &= f_p(y_i \mid x_i,\theta,\gamma,i \in \mathrm{s}) \\ &= \frac{\Pr(i \in \mathrm{s} \mid x_i, y_i, \gamma) f_p(y_i \mid \mathbf{x}_i, \theta)}{\Pr(i \in \mathrm{s} \mid x_i, \theta, \gamma)} \\ &= \frac{E_p(\pi_i \mid x_i, y_i, \gamma) f_p(y_i \mid x_i, \theta)}{E_p(\pi_i \mid x_i, \theta, \gamma)} \end{aligned} \tag{1}$$

where θ is the parameter of the population distribution, γ is the parameter indexing $\Pr(i \in s \mid x_i, y_i, \gamma)$ and

$$E_p(\pi_i \mid x_i, \theta, \gamma) = \int E_p(\pi_i \mid x_i, y_i, \gamma) f_p(y_i \mid x_i, \theta) dy_i$$

Having derived the sample distribution, Pfeffermann, Krieger and Rinott (1998) proved that if the population measurements y_i are independent, then as $N \to \infty$ (with n fixed, where n is the sample size), the sample measurements are asymptotically independent, so we can apply standard inference procedures to complex survey data by using the marginal sample distribution for each unit. Based on the sample data $\{y_i, x_i, w_i;\ i \in s\}$, we can estimate the parameters of the population model in two steps:

Step-one: According to Pfeffermann and Sverchkov (1999), estimate the informativeness parameters γ using the following relationship:

$$E_s(w_i \mid x_i, y_i, \gamma) = 1 / E_p(\pi_i \mid x_i, y_i, \gamma) \tag{2}$$

Thus the informativeness parameters can be estimated using regression analysis. Denoting the resulting estimate of γ by $\tilde{\gamma}$.

Step-two: Substitute $\tilde{\gamma}$ in the sample log-likelihood function, and then maximize the resulting sample log-likelihood function with respect to the population parameters, θ:

$$\begin{aligned} l_{rs}(\theta,\tilde{\gamma}) &= l_{srs}(\theta) - \sum_{i=1}^{n} \log E_p(\pi_i \mid \mathbf{x}_i, \theta, \tilde{\gamma}) \\ &= l_{srs}(\theta) + \sum_{i=1}^{n} \log E_s(w_i \mid \mathbf{x}_i, \theta, \tilde{\gamma}) \end{aligned} \tag{3}$$

where $l_{rs}(\theta,\tilde{\gamma})$ is the sample log-likelihood after substituting $\tilde{\gamma}$ in the sample log-likelihood function and where

$$l_{srs}(\theta) = \sum\nolimits_{i \in s} \log\{f_p(y_i \mid x_i, \theta)\}$$

is the classical log-likelihood obtained by ignoring the sample design.

3. Multivariate normal distribution under informative sampling

Most classical methods of multivariate analysis of continuous data are based on an examination of the structure of population mean vectors and covariance matrices. We consider the problem of estimating the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$ of a vector of study variables $\mathbf{y}$, from survey data obtained under informative sampling. The original theoretical basis for this is the multivariate normal distribution. The existing work in this area deals with the estimation problem when the sampling scheme is noninformative; see for example Smith and Holmes (1989).

The following theorem focuses on multivariate normal distribution under different modelling of the population conditional expectation of first order inclusion probabilities.

Theorem 1. Assume that the population distribution is q-dimensional multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$, that is:

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})' \underset{p}{\sim} N_q(\boldsymbol{\mu}, \mathbf{V}), i = 1, \ldots, N$$

Let $E_p(y_{ij}) = \mu_j$ and $Cov_p(y_{ij}, y_{ik}) = v_{jk}$, $i = 1, \ldots, N;\ j, k = 1, 2, \ldots, q.$

1. Under the exponential inclusion probability model – exponential sampling:

$$\begin{aligned} E_p(\pi_i \mid \mathbf{y}_i) &= \exp(a_0 + \mathbf{a}'\mathbf{y}_i) \\ &= \exp(a_0)\prod_{j=1}^{q}\exp(a_j y_{ij}) \end{aligned} \tag{4}$$

The sample probability density function of $\mathbf{y}_i$ is:

$$f_s(\mathbf{y}_i) = (2\pi)^{-0.5q}|\mathbf{V}|^{-0.5}\exp\left[-0.5\{\mathbf{y}_i - (\boldsymbol{\mu} + \mathbf{Va})\}'\mathbf{V}^{-1}\{\mathbf{y}_i - (\boldsymbol{\mu} + \mathbf{Va})\}\right] \tag{5}$$

That is, $\mathbf{y}_i = (y_{i1},\ldots,y_{iq})' \underset{s}{\sim} N_q(\boldsymbol{\mu} + \mathbf{Va}, \mathbf{V}), i = 1,\ldots,n$. So that,

$$\begin{aligned} E_s(\mathbf{y}_i) &= E_p(\mathbf{y}_i \mid i \in s) \\ &= \boldsymbol{\mu} + \mathbf{Va} = E_p(\mathbf{y}_i) + \mathbf{Va} \end{aligned}$$

and

$$\begin{aligned} Var_s(\mathbf{y}_i) &= Var_p(\mathbf{y}_i \mid i \in s) \\ &= \mathbf{V} = Var_p(\mathbf{y}_i) \end{aligned}$$

2. Under the linear inclusion probability model – linear sampling:

$$\begin{aligned} E_p(\pi_i \mid \mathbf{y}_i) &= b_0 + \mathbf{b}'\mathbf{y}_i \\ &= b_0 + \sum_{t=1}^{q} b_j y_{ij} \end{aligned} \tag{6}$$

The sample probability density function of $\mathbf{y}_i$ is:

$$f_s(\mathbf{y}_i) = \frac{(b_0 + \mathbf{b}'\mathbf{y}_i)(2\pi)^{-0.5q}|\mathbf{V}|^{-0.5}\exp\{-0.5(\mathbf{y}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\}}{b_0 + \mathbf{b}'(\boldsymbol{\mu})} \tag{7}$$

Furthermore,

$$\begin{aligned} E_s(\mathbf{y}_i) &= \boldsymbol{\mu} + \frac{\mathbf{Vb}}{b_0 + \mathbf{b}'\boldsymbol{\mu}} \\ &= E_p(\mathbf{y}_i) + \frac{\mathbf{Vb}}{b_0 + \mathbf{b}'\boldsymbol{\mu}} \end{aligned} \tag{8a}$$

and

$$\begin{aligned} Var_s(\mathbf{y}_i) = Cov_s(\mathbf{y}_i) &= \mathbf{V} - \frac{(\mathbf{Vb})(\mathbf{Vb})'}{(b_0 + \mathbf{b}'\boldsymbol{\mu})^2} \\ &= Var_p(\mathbf{y}_i) - \frac{(\mathbf{Vb})(\mathbf{Vb})'}{(b_0 + \mathbf{b}'\boldsymbol{\mu})^2} \end{aligned} \tag{8b}$$

Proofs:

1. As an extension of equation (1), the sample probability density function of $\mathbf{y}_i$ is given by:

$$f_s(\mathbf{y}_i) = f_p(\mathbf{y}_i \mid i \in \mathrm{s}) = \frac{E_p(\pi_i \mid \mathbf{y}_i) f_p(\mathbf{y}_i)}{E_p(\pi_i)} \tag{9}$$

So that,

$$\begin{aligned} f_s(\mathbf{y}_i) &= \frac{\exp(\mathbf{a}'\mathbf{y}_i)(2\pi)^{-0.5q}|\mathbf{V}|^{-0.5}\exp\left\{-0.5(\mathbf{y}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}_i - \boldsymbol{\mu})\right\}}{\exp(\mathbf{a}'\boldsymbol{\mu} + 0.5\mathbf{a}'\mathbf{V}\mathbf{a})} \\ &= (2\pi)^{-\frac{q}{2}}|\mathbf{V}|^{-0.5}\exp(-0.5\mathbf{a}'\mathbf{V}\mathbf{a})\exp(-0.5Q) \end{aligned} \tag{10}$$

where $Q = (\mathbf{y}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) - 2\mathbf{a}'(\mathbf{y}_i - \boldsymbol{\mu})$.

Setting $\mathbf{C}_i = \mathbf{y}_i - \boldsymbol{\mu}$, then Q can be written as: $\mathrm{Q} = \mathbf{C}_i'\mathbf{V}^{-1}\mathbf{C}_i - 2\mathbf{a}'\mathbf{C}_i$ .Using theorems in multivariate statistical analysis, see Johnson and Wichern (1998), page 68, we have:

$$\mathrm{Q} = \mathbf{C}_i'\mathbf{V}^{-0.5}\mathbf{V}^{-0.5}\mathbf{C}_i - 2\mathbf{a}'\mathbf{V}^{0.5}\mathbf{V}^{-0.5}\mathbf{C}_i$$

Now let $\mathbf{D}_i = \mathbf{V}^{-0.5}\mathbf{C}_i$, then:

$$\begin{aligned} \mathrm{Q} &= \mathbf{D}_i'\mathbf{D}_i - 2\mathbf{a}'\mathbf{V}^{0..5}\mathbf{D}_i \\ &= \mathbf{D}_i'\mathbf{D}_i - 2\mathbf{a}'\mathbf{V}^{0.5}\mathbf{D}_i + \mathbf{a}'\mathbf{V}^{0.5}\mathbf{V}^{0.5}\mathbf{a} - \mathbf{a}'\mathbf{V}^{0.5}\mathbf{V}^{0.5}\mathbf{a} \\ &= (\mathbf{D}_i - \mathbf{V}^{0.5}\mathbf{a})'(\mathbf{D}_i - \mathbf{V}^{0.5}\mathbf{a}_1) - \mathbf{a}'\mathbf{V}\mathbf{a} \end{aligned}$$

But $\mathbf{C}_i = \mathbf{y}_i - \boldsymbol{\mu}$ and $\mathbf{D}_i = \mathbf{V}^{-0.5}\mathbf{C}_i$, so that Q can be expressed equivalently as:

$$\mathrm{Q} = \left(\mathbf{V}^{-0.5}\left((\mathbf{y}_i - \boldsymbol{\mu}) - \mathbf{V}^{0.5}\mathbf{V}^{0.5}\mathbf{a}\right)\right)'\left(\mathbf{V}^{-0.5}\left((\mathbf{y}_i - \boldsymbol{\mu}) - \mathbf{V}^{0.5}\mathbf{V}^{0.5}\mathbf{a}\right)\right) - \mathbf{a}'\mathbf{V}\mathbf{a}$$
$$= \left(\mathbf{y}_i - (\boldsymbol{\mu} + \mathbf{V}\mathbf{a})\right)'\mathbf{V}^{-1}\left(\mathbf{y}_i - (\boldsymbol{\mu} + \mathbf{V}\mathbf{a})\right) - \mathbf{a}'\mathbf{V}\mathbf{a}$$

Thus, after substituting this expression of Q in (10), we get (5).

Hence the multivariate normal distribution in the sample is the same as in the population, except that the mean is shifted by the constant $\mathbf{Va}$. Notice that the sample probability density function does not depend on a_0. Note also:

$$E_s(y_{ij}) = \mu_j + a_1 v_{j1} + \ldots + a_j v_{jj} + \ldots + a_q v_{jq} \qquad (11a)$$

and

$$Var_s(y_{ij}) = v_{jj}, Cov_p(y_{ij}, y_{ik}) = v_{jk}, \, j \neq k = 1,\ldots,q \qquad (11b)$$

2. Substituting (6) in (9) we get (7).

Let us now compute the first two moments of this sample pdf. In order to do this we will use the moment generating function technique. The moment generating function of the sample probability density function is given by:

$$M_s(\mathbf{u}_i) = E(\exp(\mathbf{u}_i'\mathbf{y}_i)) = \int (\exp(\mathbf{u}_i'\mathbf{y}_i))\frac{b_0 + \mathbf{b}'\mathbf{y}_i}{b_0 + \mathbf{b}'\boldsymbol{\mu}} f_p(\mathbf{y}_i)d\mathbf{y}_i$$
$$= \frac{b_0}{b_0 + \mathbf{b}'\boldsymbol{\mu}} M_p(\mathbf{u}_i) + \frac{\mathbf{b}'\dfrac{dM_p(\mathbf{u}_i)}{d\mathbf{u}_i}}{b_0 + \mathbf{b}'\boldsymbol{\mu}}$$

where

$$M_p(\mathbf{u}_i) = \exp(\mathbf{u}_i'\boldsymbol{\mu} + .5\mathbf{u}_i'\mathbf{V}\mathbf{u}_i)$$

and

$$\frac{dM_p(\mathbf{u}_i)}{d\mathbf{u}_i} = M_p(\mathbf{u}_i)(\boldsymbol{\mu} + \mathbf{V}\mathbf{u}_i)$$

Thus, we have:

$$M_s(\mathbf{u}_i) = \frac{M_p(\mathbf{u}_i)}{b_0 + \mathbf{b}'\boldsymbol{\mu}}\left(b_0 + \mathbf{b}'(\boldsymbol{\mu} + \mathbf{V}_0\mathbf{u}_i)\right) \qquad (12)$$

This equation gives explicitly the relationship between the population and the sample moment generating functions. Notice that the sample and population moment generating functions are different unless $\mathbf{b} = \mathbf{0}$, in which case the sampling mechanism is noninformative.

Let $R_s(\mathbf{u}_i) = \log M_s(\mathbf{u}_i)$. Differentiating this expression twice and setting $\mathbf{u}_i = \mathbf{0}$, we get (8a) and (8b).

From (8a) and (8b), we can see that:

$$E_s(y_{ij}) = \mu_j + \frac{b_1 v_{j1} + \ldots + b_j v_{jj} + \ldots + b_q v_{jq}}{b_0 + b_1\mu_1 + \ldots + b_j\mu_j + \ldots + b_q\mu_{iq}}, \tag{13a}$$

$$Var_s(y_{ij}) = v_{jj} - \frac{(b_1 v_{j1} + \ldots + b_j v_{jj} + \ldots + b_q v_{jq})^2}{(b_0 + b_1\mu_1 + \ldots b_j\mu_j + \ldots + b_q\mu_q)^2} \tag{13b}$$

$$Cov_p(y_{ij}, y_{ik}) = v_{jk}$$

$$-\frac{(b_1 v_{j1} + \ldots + b_j v_{jj} + \ldots + b_q v_{jq})(b_1 v_{k1} + \ldots + b_j v_{kj} + \ldots + b_q v_{kq})}{(b_0 + b_1\mu_1 + .. b_j\mu_j + \ldots + b_q\mu_q)^2} \tag{13c}$$

for $i = 1,\ldots,n,\ j \neq k = 1,2,\ldots,q.$

Also $Var_s(y_{ij}) \leq Var_p(y_{ij})$ and the equality holds if and only if $\mathbf{b} = \mathbf{0}$, that is when the sampling mechanism is noninformative. On the other hand, if **b≠0**, then the means the variances and the covariances change, in contradiction to what happens for the sample probability density function (5), where only the means change.

To illustrate the results, from now on, we only consider the particular cases of the exponential inclusion probability model – exponential sampling, see equation (4). The following are special cases of Theorem 1.

Corollary 1. (Univariate normal distribution, $q = 1$)

Let $E_p(y_{i1}) = \mu_1$ and $Var_p(y_{i1}) = v_{11}$. Under the exponential sampling:

$$E_p(\pi_i \mid y_{i1}) = \exp(a_0 + a_1 y_{i1}) \tag{14}$$

We get the following result:

$$y_{i1} \underset{s}{\sim} N(\mu_1 + a_1 v_{11}, v_{11}) \tag{15}$$

Corollary 2. (Bivariate normal distribution, $q = 2$).

Let $E_p(y_{i1}) = \mu_1, E_p(y_{i2}) = \mu_2$, $Var_p(y_{i1}) = v_{11}$, $Var_p(y_{i2}) = v_{22}$, $Cov_p(y_{i1}, y_{i2}) = v_{12}$ and $Cor_p(y_{i1}, y_{i2}) = \rho_{12}$. Under the exponential sampling:

$$E_p(\pi_i \mid y_{i1}, y_{i2}) = \exp(a_0 + a_1 y_{i1} + a_2 y_{i2}) \tag{16}$$

and using the properties of multivariate normal distribution, see Johnson and Wichern (1998), page 171, we obtain the following results:

1. The joint sample probability density function of (y_{i1}, y_{i2}) is:

$$(y_{i1}, y_{i2})' \underset{s}{\sim} N_2 \left\{ \begin{pmatrix} \mu_1 + a_1 v_{11} + a_2 v_{12} \\ \mu_2 + a_1 v_{21} + a_2 v_{22} \end{pmatrix}, \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix} \right\} \tag{17a}$$

2. The sample marginal probability density functions of y_{i1} and y_{i2} are respectively given by:

$$y_{i1} \underset{s}{\sim} N(\mu_1 + a_1 v_{11} + a_2 v_{12}, v_{11}) \tag{17b}$$

and

$$y_{i2} \underset{s}{\sim} N(\mu_2 + a_1 v_{21} + a_2 v_{22}, v_{22}) \tag{17c}$$

3. The conditional sample probability density function of y_{i1} given y_{i2} is univariate normal distribution with conditional mean:

$$\begin{aligned} E_s(y_{i1} \mid y_{i2}) &= E_s(y_{i1}) + \rho_{12}\sqrt{\frac{v_{22}}{v_{11}}}(y_{i2} - E_s(y_{i2})) \\ &= \mu_1 + a_1 v_{11} + a_2 v_{12} + \rho_{12}\sqrt{\frac{v_{11}}{v_{22}}}(y_{i2} - (\mu_2 + a_1 v_{21} + a_2 v_{22})) \\ &= \mu_1 + \rho_{12}\sqrt{\frac{v_{11}}{v_{22}}}(y_{i2} - \mu_2) + a_1\left(v_{11} - \frac{v_{12}^2}{v_{22}}\right) \\ &= \mu_1 + \rho_{12}\sqrt{\frac{v_{11}}{v_{22}}}(y_{i2} - \mu_2) + a_1 v_{11}(1 - \rho_{12}^2) \end{aligned} \tag{17d}$$

and conditional variance:

$$V_s(y_{i1} \mid y_{i2}) = v_{11}(1 - \rho_{12}^2) = V_p(y_{i1} \mid y_{i2}) \tag{17e}$$

That is,

$$y_{i1} \mid y_{i2} \underset{s}{\sim} N\{\mu_1 + \rho_{12}(v_{11}/v_{22})^{0.5}(y_{i2} - \mu_2) + a_1 v_{11}(1 - \rho_{12}^2), v_{11}(1 - \rho_{12}^2)\} \tag{17f}$$

Notice that:

$$E_s(y_{i1} \mid y_{i2}) = E_p(y_{i1} \mid y_{i2}) + a_1 V_p(y_{i1} \mid y_{i2}) \tag{17g}$$

4. Similarly, the conditional sample probability density function of y_{i2} given y_{i1} is:

$$y_{i2} \mid y_{i1} \underset{s}{\sim} N\left\{\mu_2 + \rho_{12}\left(v_{22}/v_{11}\right)^{0.5}\left(y_{i1} - \mu_1\right) + a_2 v_{22}\left(1-\rho_{12}^2\right), v_{22}\left(1-\rho_{12}^2\right)\right\} \quad (17h)$$

Thus, the sample and population probability density functions are different, but belong to the same family of distribution, which is normal. Also the change occurs only for the means and conditional means, whereas the variances, conditional variances, and covariance do not change.

In particular if $a_2 = 0$, that is the inclusion probabilities depend only on y_{i1} (which is the case in panel surveys), then we have:

1. The joint sample probability density function of $\left(y_{i1}, y_{i2}\right)$ is:

$$\left(y_{i1}, y_{i2}\right)' \underset{s}{\sim} N_2\left\{\begin{pmatrix} \mu_1 + a_1 v_{11} \\ \mu_2 + a_1 v_{12} \end{pmatrix}, \begin{bmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{bmatrix}\right\} \quad (18a)$$

2. The marginal sample probability density functions of y_{i1} and y_{i2} are respectively given by:

$$y_{i1} \underset{s}{\sim} N\left(\mu_1 + a_1 v_{11}, v_{11}\right) \quad (18b)$$

and

$$y_{i2} \underset{s}{\sim} N\left(\mu_2 + a_1 v_{12}, v_{22}\right) \quad (18c)$$

3. The conditional sample probability density function of y_{i1} given y_{i2} is:

$$y_{i1} \mid y_{i2} \underset{s}{\sim} N\left\{\mu_1 + a_1 v_{11}\left(1-\rho_{12}^2\right) + \rho_{12}\left(v_{11}/v_{22}\right)^{0.5}\left(y_{i2} - \mu_2\right), v_{11}\left(1-\rho_{12}^2\right)\right\} \quad (18d)$$

4. The conditional sample probability density function of y_{i2} given y_{i1} is:

$$y_{i2} \mid y_{i1} \underset{s}{\sim} N\left\{\mu_2 + \rho_{12}\left(v_{22}/v_{11}\right)^{0.5}\left(y_{i1} - \mu_1\right), v_{22}\left(1-\rho_{12}^2\right)\right\} \quad (18e)$$

Notice that the sample and population probability density functions of $y_{i2} \mid y_{i1}$ are the same, while the other sample and population distributions are different.

Birnbaum et al. (1950) studied the effect of selection performed on some coordinates of a multi-dimensional population, but from different point of view.

Notice that the sample and population pdf's of $y_{i2} \mid y_{i1}$ are the same, while the other sample and population distributions are different.

Corollary 3. (Bivariate normal distribution, $q = 2$).

Let $E_p(y_{i1}) = \mu_1, E_p(y_{i2}) = \mu_2$, $Var_p(y_{i1}) = v_{11}$, $Var_p(y_{i2}) = v_{22}$, $Cov_p(y_{i1}, y_{i2}) = v_{12}$ and $Cor_p(y_{i1}, y_{i2}) = \rho_{12}$. Under the linear model:

$$E_p(\pi_i \mid y_{i1}, y_{i2}) = b_0 + b_1 y_{i1} + b_2 y_{i2}$$

and using the properties of multivariate normal distribution, we have the following results:

1. The joint sample probability density function of (y_{i1}, y_{i2}) is:

$$f_s(y_{i1}, y_{i2}) = \frac{b_0 + b_1 y_{i1} + b_2 y_{i2}}{b_0 + b_1 \mu_1 + b_2 \mu_2} f_p(y_{i1}, y_{i2}) \tag{19a}$$

2. Integrating (19a), with respect to y_{i2}, we have:

$$\begin{aligned}
f_s(y_{i1}) &= \int f_s(y_{i1}, y_{i2}) dy_{i2} = \int \frac{b_0 + b_1 y_{i1} + b_2 y_{i2}}{b_0 + b_1 \mu_1 + b_2 \mu_2} f_p(y_{i1}, y_{i2}) dy_{i2} \\
&= \frac{1}{b_0 + b_1 \mu_1 + b_2 \mu_2} \left(\int b_0 f_p(y_{i1}, y_{i2}) dy_{i2} + \int b_1 y_{i1} f_p(y_{i1}, y_{i2}) dy_{i2} \right) + \\
&\quad \frac{1}{b_0 + b_1 \mu_1 + b_2 \mu_2} \int b_2 y_{i2} f_p(y_{i1}, y_{i2}) dy_{i2} \\
&= \frac{1}{b_0 + b_1 \mu_1 + b_2 \mu_2} \left(\int b_0 f_p(y_{i1}, y_{i2}) dy_{i2} + \int b_1 y_{i1} f_p(y_{i1}, y_{i2}) dy_{i2} \right) + \\
&\quad \frac{1}{b_0 + b_1 \mu_1 + b_2 \mu_2} \int b_2 y_{i2} f_p(y_{i1}) f_p(y_{i2} | y_{i1}) dy_{i2} \\
&= \frac{1}{b_0 + b_1 \mu_1 + b_2 \mu_2} \left(b_0 f_p(y_{i1}) + b_1 y_{i1} f_p(y_{i1}) + b_2 f_p(y_{i1}) E_p(y_{i2} | y_{i1}) \right) + \\
&\quad \frac{1}{b_0 + b_1 \mu_1 + b_2 \mu_2} \int b_2 y_{i2} f_p(y_{i1}) f_p(y_{i2} | y_{i1}) dy_{i2} \\
&= \left\{ \frac{b_0 + b_1 y_{i1} + \left(\mu_2 + \rho_{12} \left(\frac{v_{22}}{v_{11}} \right)^{0.5} (y_{i1} - \mu_1) \right)}{b_0 + b_1 \mu_1 + b_2 \mu_2} \right\} f_p(y_{i1})
\end{aligned} \tag{19b}$$

3. Similarly, integrating (19a), with respect to y_{i1}, we obtain the following marginal sample probability density function of y_{i2} :

$$f_s\left(y_{i2}\right)=\left\{\frac{b_0+b_2y_{i2}+b_1\left(\mu_1+\rho_{12}\left(\frac{v_{11}}{v_{22}}\right)^{0.5}\left(y_{i2}-\mu_2\right)\right)}{b_0+b_1\mu_1+b_2\mu_2}\right\}f_p\left(y_{i2}\right) \qquad \text{(19c)}$$

4. Using (19a), (19b), and (19c), we get the following conditional sample probability density function of y_{i1} given y_{i2}:

$$\begin{aligned}
f_s\left(y_{i1}\mid y_{i2}\right)&=\frac{f_s\left(y_{i1},y_{i2}\right)}{f_s\left(y_{i2}\right)}\\
&=\frac{b_0+b_1y_{i1}+b_2y_{i2}}{b_0+b_1\mu_1+b_2\mu_2}f_p\left(y_{i1},y_{i2}\right)\div\frac{b_0+b_2y_{i2}+b_1E_p\left(y_{i2}\middle|y_{i1}\right)}{b_0+b_1\mu_1+b_2\mu_2}f_p\left(y_{i2}\right) \qquad \text{(19d)}\\
&=\frac{b_0+b_1y_{i1}+b_2y_{i2}}{b_0+b_1E_p\left(y_{i1}\mid y_{i2}\right)+b_2y_{i2}}\frac{f_p\left(y_{i1},y_{i2}\right)}{f_p\left(y_{i2}\right)}\\
&=\left\{\frac{b_0+b_1y_{i1}+b_2y_{i2}}{b_0+b_1E_p\left(y_{i1}\mid y_{i2}\right)+b_2y_{i2}}\right\}f_p\left(y_{i1}\mid y_{i2}\right)
\end{aligned}$$

where

$$E_p\left(y_{i1}\mid y_{i2}\right)=\mu_1+\rho_{12}\left(\frac{v_{11}}{v_{22}}\right)^{0.5}\left(y_{i2}-\mu_2\right)$$

5. Similarly, the conditional sample probability density function of y_{i2} given y_{i1} is:

$$f_s\left(y_{i2}\mid y_{i1}\right)=\left\{\frac{b_1y_{i1}+b_2y_{i2}+b_0}{b_2E_p\left(y_{i2}\mid y_{i1}\right)+b_1y_{i1}+b_0}\right\}f_p\left(y_{i2}\mid y_{i1}\right) \qquad \text{(19e)}$$

where

$$E_p\left(y_{i2}\mid y_{i1}\right)=\mu_2+\rho_{12}\left(\frac{v_{22}}{v_{11}}\right)^{0.5}\left(y_{i1}-\mu_1\right)$$

Thus, in this case we see that the sample probability density functions and population probability density functions are very different. Also notice that formulas (19b, c) are true, not only for the bivariate normal distribution, but for any joint probability density function of $\left(y_{i1},y_{i2}\right)$, provided that the marginal and conditional distributions and the corresponding moments exist.

The following theorem provides the maximum likelihood estimators of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$.

Theorem 2. Assume $\mathbf{y}_i = \left(y_{i1},...,y_{iq}\right)' \underset{p}{\sim} N_q(\boldsymbol{\mu},\mathbf{V}), i = 1,...,N$ are independent. Let $\mathbf{y}_1,...,\mathbf{y}_n$ be a sample of size n selected by informative sampling.

1. Under the exponential sampling – equation (4): the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\mathbf{V}$ are given by:

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{y}} - \hat{\mathbf{V}}\tilde{\mathbf{a}} \tag{20a}$$

and

$$\hat{\mathbf{V}} = n^{-1}\sum_{i=1}^{n}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})' = \{\hat{\mathbf{V}}_{ij}\} = \{s_{ij}\} \tag{20b}$$

where $\tilde{\mathbf{a}}$ is the least square estimator under the model: $E_s(w_i \mid \mathbf{y}_i) = \exp(-a_0 - \mathbf{a}'\mathbf{y}_i)$, $\overline{\mathbf{y}} = (\overline{y}_1,...,\overline{y}_q)'$, $\overline{y}_i = n^{-1}\sum_{j=1}^{n} y_{ij}$

and $\hat{V}_{ij} = s_{ij} = n^{-1}\sum_{k=}^{n}(y_{ik} - \overline{y}_i)(y_{jk} - \overline{y}_j)$.

2. Under the linear sampling – equation (6): the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\mathbf{V}$ are defined by the equations:

$$(\overline{\mathbf{y}} - \hat{\boldsymbol{\mu}})(\tilde{b}_0 + \tilde{\mathbf{b}}'\hat{\boldsymbol{\mu}}) = \hat{\mathbf{V}}\tilde{\mathbf{b}} \tag{21a}$$

and

$$n\hat{\mathbf{V}} = \sum_{i=1}^{n}(\overline{\mathbf{y}} - \hat{\boldsymbol{\mu}})(\overline{\mathbf{y}} - \hat{\boldsymbol{\mu}})' \tag{21b}$$

where $\tilde{b}_0$ and $\tilde{\mathbf{b}}$ are the least squares estimators under the model: $E_s(w_i \mid \mathbf{y}_i) = 1/(-b_0 - \mathbf{b}'\mathbf{y}_i)$. Solve (21a) and (21b) iteratively. Start with classical maximum likelihood estimators.

Proofs:

1. Exponential sampling. Using the two-step method of estimation.

Step-one. Estimation of informativeness parameters a_0 and $\mathbf{a}$ via the relationship (2).

Step-two. After substituting $\tilde{\mathbf{a}}$ in (5), the resulting sample log-likelihood is given by:

$$l_{rs}(\boldsymbol{\mu},\mathbf{V}) = -0.5nq\log 2\pi - 0.5n\log|\mathbf{V}| - 0.5\sum_{i=1}^{n}\left(\mathbf{y}_i-\boldsymbol{\mu}^*\right)'\mathbf{V}^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}^*\right) \quad (22)$$

where $\boldsymbol{\mu}^* = \boldsymbol{\mu} + \mathbf{V}\tilde{\mathbf{a}}$.

According to Johnson and Wichern (1998), page 182, the maximum likelihood estimators of $\boldsymbol{\mu}^*$ and $\mathbf{V}$ are given by:

$$\hat{\boldsymbol{\mu}}^* = \bar{\mathbf{y}}$$

and

$$\hat{\mathbf{V}} = n^{-1}\sum_{i=1}^{n}\left(\mathbf{y}_i-\bar{\mathbf{y}}\right)\left(\mathbf{y}_i-\bar{\mathbf{y}}\right)'.$$

Hence the result – equation (20).

2. Linear model. Using the two-step method of estimation.

Step-one. Estimation of informativeness parameters, b_0 and $\mathbf{b}$ via the relationship (2).

Step-two. After substituting $\tilde{b}_0$ and $\tilde{\mathbf{b}}$ in (7), the resulting sample log-likelihood is given by:

$$l_{rs}(\boldsymbol{\mu},\mathbf{V}) = -\frac{1}{2}nq\log 2\pi - \frac{1}{2}n\log|\mathbf{V}| - \frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)'\mathbf{V}^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}\right)$$

$$-n\log\left(\tilde{b}_0+\tilde{\mathbf{b}}'\boldsymbol{\mu}\right) \quad (23)$$

Now, differentiating $l_{rs}(\boldsymbol{\mu},\mathbf{V})$ with respect to $\boldsymbol{\mu}$ and $\mathbf{V}$, we can show that the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\mathbf{V}$ are defined by equation (21).

The following corollary provides the maximum likelihood estimators of the mean μ and the variance σ^2 when the population model is univariate normal and the sampling process is exponential and linear, which is a particular case of Theorem 2 with $q=1$.

Corollary 4. Assume $y_i \underset{p}{\sim} N\left(\mu,\sigma^2\right), i=1,\dots,N$ are independent. Let $y_1,\dots,y_n$ be a sample of size n selected under the following sampling schemes.

1.Exponential sampling. The maximum likelihood estimators of μ and σ^2 are given by:

$$\hat{\mu} = \bar{y} - \tilde{a}_1\hat{\sigma}^2 \quad (24a)$$

and

$$\hat{\sigma}^2 = s^2 = n^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{24b}$$

where $\tilde{a}_1$ is the least square estimator of the informativeness parameter a_1.

2. Linear sampling. The maximum likelihood estimators of μ and σ^2 are given by:

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{\mu}) - n\tilde{b}_1\left(\tilde{b}_0 + \tilde{b}_1\hat{\mu}\right)^{-1} = 0 \tag{25a}$$

and

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 \tag{25b}$$

where $\tilde{b}_0$ and $\tilde{b}_1$ are the estimators of the informativeness parameters b_0 and b_1.

Corollary 5. Assume $\mathbf{y}_i = (y_{i1}, y_{i2})' \underset{p}{\sim} N_2(\boldsymbol{\mu}, \mathbf{V}), i = 1,..., N$. Under the exponential sampling – equation (4): the maximum likelihood estimators of μ_1, μ_2, v_{11}, v_{22}, v_{11} and are given by:

$$\hat{\mu}_1 = \bar{y}_1 - \tilde{a}_1 s_{11} - \tilde{a}_2 s_{12} \tag{26a}$$

$$\hat{\mu}_2 = \bar{y}_2 - \tilde{a}_2 s_{22} - \tilde{a}_1 s_{12} \tag{26b}$$

$$\hat{\mathbf{V}} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \{s_{ij}\}, i, j = 1,2 \tag{27}$$

4. Application – fitting general linear model for longitudinal survey data under informative sampling

4.1.Population model

Let $y_{it}, i = 1,..., N; t = 1,..., T$ be the measurement on the *i-th* subject at time $t = 1,..., T$. Associated with each y_{it} are the (known) values, $x_{itk}, k = 1,...p$, of p explanatory variables. We assume that the y_{it} follow the regression model:

$$y_{it} = \beta_1 x_{it1} + ... + \beta_p x_{itp} + \varepsilon_{it} \tag{28}$$

where ε_{it} are random sequence of length T associated with each of the N subjects. In our context, the longitudinal structure of the data means that we expect the ε_{it} to be correlated within subjects.

Let $\mathbf{y}_i = \left(y_{i1},...,y_{iT}\right)'$, $\mathbf{x}_{it} = \left(x_{it1},...,x_{itp}\right)'$ and let $\boldsymbol{\beta} = \left(\beta_1,...,\beta_p\right)'$ be the vector of unknown regression coefficients. The general linear model for longitudinal survey data treats the random vectors $\mathbf{y}_i, i = 1,...,N$ as independent multivariate normal variables, that is

$$\mathbf{y}_i \underset{p}{\sim} N_q\left(\mathbf{x}_i\boldsymbol{\beta},\mathbf{V}\right) \tag{29}$$

where $\mathbf{x}_i$ is the matrix of size T by p of explanatory variables for subject i, and $\mathbf{V}$ has $(jk)-th$ element, $v_{jk} = \text{cov}_p\left(y_{ij}, y_{ik}\right) j,k = 1,...,T$; see Diggle, Liang and Zeger (1994).

4.2.Covariance structure of $\mathbf{y}_i$

It is useful at this stage to consider what form the matrix **V** might take. We consider three cases: the exponential correlation model, the uniform correlation model; see Diggle, Liang and Zeger (1994), and the random effect model; see Skinner and Holmes (2003).

Case1: The exponential correlation model:

In this model, the $(t,s)-\text{th}$ element of **V** has the form:

$$v_{ts} = \text{cov}_p\left(y_{it}, y_{is}\right) = \sigma^2\rho^{|t-s|}, t,s = 1,...,T \tag{30}$$

Note that the correlation, v_{ts}/σ^2, between a pair of measurements on the same unit decays toward zero as the time separation between the measurements increases.

Case2: The uniform correlation model:

In this model, we assume that there is a positive correlation, ρ, between any two measurements on the same subject. So that the $(t,s)-th$ element of **V** has the form:

$$v_{ts} = \text{cov}_p\left(y_{it}, y_{is}\right) = \sigma^2\rho, t \neq s = 1,...,T \;\; ; \;\; v_{tt} = \sigma^2, t = 1,...,T \tag{31}$$

Case3: Random effects models:

Under this model the multivariate outcomes $\mathbf{y}_i = \left(y_{i1},...,y_{iT}\right)', i = 1,...,N$ are independent with mean vector and covariance matrix given respectively by:

$$E_p(\mathbf{y}_i) = (\beta_1, ..., \beta_T) = \boldsymbol{\mu} \qquad (32a)$$

$$\mathrm{cov}_p(\mathbf{y}_i) = \sigma_u^2 \mathbf{J}_T + \sigma^2 \mathbf{V}_T = \Sigma \qquad (32b)$$

where $\mathbf{J}_T$ denotes the T by T matrix all of whose elements are one, and the $(t, t') - th$ element of $\mathbf{V}_T$ is $\rho^{|t-t'|}; t, t' = .,...,T$.

4.3.Sampling design

We assume a single-stage informative sampling design, where the sample is a panel sample selected at time $t = 1$ and all units remain in the sample till time $t = T$. Examples of longitudinal surveys, some of which are based on complex sample designs, and of the issues involved in their design and analysis can be found in Herriot and Kasprzyk (1984), and Nathan (1999). In many of the cases described in these papers, a sample is selected for the first round and continues to serve for several rounds. Then, it is intuitively reasonable to assume that the first order inclusion probabilities, π_i , depend on the population values of the response variable at the first occasion only, the values y_{i1}, and on $\mathbf{x}_{i1} = (x_{i11}, \ldots, x_{i1p})'$, and the values of known design variable, $\mathbf{z} = \{z_1, ..., z_N\}$, used for the sample selection, but not included in the working model under consideration.

4.4.Sample distribution

Under exponential inclusion probability model:

$$E_p(\pi_i \mid y_{i1}, \mathbf{x}_{i1}) = \exp(a_0^* + a_0 y_{i1} + a_1 x_{i11} + a_2 x_{i12} + ... + a_p x_{i1p}) \qquad (33)$$

Using (29), (33) and Theorem 1, we have:

$$\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}, a_0 \underset{s}{\sim} N_q(\boldsymbol{\mu}^*, \mathbf{V}) \qquad (34)$$

where,

$$\boldsymbol{\mu}^* = [\mathbf{x}'_{i1}\boldsymbol{\beta} + a_0 v_{11}, \mathbf{x}'_{i2}\boldsymbol{\beta} + a_0 v_{12}, ..., \mathbf{x}'_{iT}\boldsymbol{\beta} + a_0 v_{1T}]'$$

Note that $\mathbf{y}_{i,T-1} = (y_{i2}, y_{i3}, \ldots, y_{iT})'$. Alternatively, the sample probability density function of $\mathbf{y}_i$ can be written as:

$$f_s(\mathbf{y}_i | \mathbf{x}_i) = f_s(y_{i1} | \mathbf{x}_{i1}) f_p(\mathbf{y}_{i,T-1} \mid y_{i1}, \mathbf{x}_i) \qquad (35)$$

where

$$f_s\left(y_{i1}\middle|\mathbf{x}_{i1},\boldsymbol{\theta},\boldsymbol{\gamma}\right)=\frac{1}{\sqrt{2\pi v_{11}}}\exp\left[-\frac{1}{2v_{11}}\left(y_{i1}-\mathbf{x}'_{i1}\boldsymbol{\beta}-a_0v_{11}\right)^2\right] \tag{36}$$

$$f_p\left(\mathbf{y}_{i,T-1}\mid y_{i1},\mathbf{x}_i\right)=\frac{1}{\left(2\pi\right)^{T-1}\left|V^*_{0,T-1}\right|^{1/2}}$$

$$\mathrm{e\ x}\left[\mathrm{p}-\frac{1}{2}\left(\mathbf{y}_{i,T-1}-\boldsymbol{\mu}_{T-1}\right)'\left(V^*_{0,T-1}\right)^{-1}\left(\mathbf{y}_{i,T-1}-\boldsymbol{\mu}_{T-1}\right)\right] \tag{37}$$

$$\begin{aligned}\boldsymbol{\mu}_{T-1}&=E_p\left\lfloor\mathbf{y}_{i,T-1}\mid y_{i1},\mathbf{x}_i\right\rfloor\\&=\left[\mathbf{x}'_{i2}\boldsymbol{\beta}+\frac{v^*_{21}}{v^*_{11}}\left(y_{i1}-\mathbf{x}'_{i1}\boldsymbol{\beta}\right),\ldots,\mathbf{x}'_{iT}\boldsymbol{\beta}+\frac{v^*_{T1}}{v^*_{11}}\left(y_{i1}-\mathbf{x}'_{i1}\boldsymbol{\beta}\right)\right]'\end{aligned}$$

with general term:

$$v_{tt'}=v_{t+1,t'+1}-\left(v_{11}\right)^{-1}v_{t+1,1}v_{1,t'+1};\ t,t'=1,\ldots,T-1.$$

So that we have the following sample model:

$$\begin{aligned}y_{it}&=\beta_0+\beta_1x_{it1}+\ldots+\beta_px_{itp}+\varepsilon_{it}\\&=\mathbf{x}^*_i\boldsymbol{\beta}^*+\varepsilon_{it};i=1,2,\ldots,n.\end{aligned}, \tag{38}$$

where $\beta_0=a_0v_{11}$, $\boldsymbol{\beta}^*=\left(\beta_0,\beta_1,\ldots,\beta_p\right)'$, $\mathbf{x}^*_i=\left(\mathbf{1},\mathbf{x}_i\right)$ and the ε_{it} are a random sequence correlated within subjects.

Note that if $a_0=0$, that is the sampling design is noninformative, then the population and sample models are the same.

4.5.Estimation

We consider three method of estimation, namely: unweighted maximum likelihood, pseudo maximum likelihood, and two-step estimation based on the sample distribution.

4.5.1. Unweighted maximum likelihood

Maximum likelihood for the case where the sampling design is ignorable: the value of $\boldsymbol{\theta}=\left(\boldsymbol{\beta},\mathbf{V}\right)$ that satisfy:

$$\frac{\partial}{\partial \boldsymbol{\theta}} l_{srs}(\boldsymbol{\mu}, \mathbf{V}) = -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\theta}}\left\{ nq \log 2\pi + n \log|\mathbf{V}| + \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu})' \mathbf{V}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}) \right\} \quad (39)$$

4.5.2. Pseudo maximum likelihood

The pseudo maximum likelihood estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{V})$ is defined as the solution of:

$$\hat{U}_w(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i \in s} w_i \frac{\partial}{\partial \boldsymbol{\theta}}\Big[q \ln 2\pi + \ln|\mathbf{V}|$$

$$+ \{\mathbf{y}_i - (\boldsymbol{\mu} + \mathbf{V}\ \mathbf{a})\}' \mathbf{V}^{-1} \{\mathbf{y}_i - (\boldsymbol{\mu} + \mathbf{V}\ \mathbf{a})\}\Big] = 0 \quad (40)$$

$$\hat{U}_{ws}(\boldsymbol{\theta}) = \sum_{i \in s} w_i \frac{\partial}{\partial \boldsymbol{\theta}} \ln\{ f_s (y_{i1} | \mathbf{x}_{i1})\} + \frac{N}{n} \sum_{i \in s} \frac{\partial}{\partial \boldsymbol{\theta}} \ln\{ f_p (\mathbf{y}_{i,T-1} \mid y_{i1}, \mathbf{x}_i)\} = 0 \quad (41)$$

For more discussion, see Eideh and Nathan (2006).

4.5.3. Two-step method

Step one. Estimate a_0 via the model: $E_s(w_i \mid y_{i1}, \mathbf{x}_{i1}) = \exp(- a_0^* - a_0 y_{i1} - \mathbf{a}'\mathbf{x}_{i1})$.

Step two. Using Theorem 2, the maximum likelihood estimators of $\boldsymbol{\beta}^*$ and $\mathbf{V}$ are defined as the solution of:

$$\frac{\partial}{\partial \boldsymbol{\theta}} l_{rs}(\boldsymbol{\beta}^*, \mathbf{V}) = -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\theta}}\left\{ nq \log 2\pi + n \log|\mathbf{V}| + \sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}^*)' \mathbf{V}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}^*) \right\} = 0 \quad (42)$$

where $\boldsymbol{\beta}^* = (\beta_0, \beta_1, ..., \beta_p)'$

and $\boldsymbol{\mu}^* = [\mathbf{x}'_{i1}\boldsymbol{\beta} + a_0 v_{11}, \mathbf{x}'_{i2}\boldsymbol{\beta} + a_0 v_{12}, ..., \mathbf{x}'_{iT}\boldsymbol{\beta} + a_0 v_{1T}]'$.

4.6. Variance estimation

For variance estimation we use the inverse of Fisher information matrix and the bootstrap approach for variance, see Pfeffermann and Sverchkov (1999, 2003) and Eideh and Nathan (2006).

(a) Fisher information matrix approach:

The inverse of the observed Fisher information matrix evaluated at $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}}\right)$ is given by:

$$\begin{aligned}\hat{V}_s\left(\hat{\boldsymbol{\theta}}\right) &= \left[I_s\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1} \\ &= \left\{-\frac{1}{n}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'\partial\boldsymbol{\theta}}\right]\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right\}^{-1}\end{aligned} \tag{43}$$

(b) Bootstrap approach:

Let $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}}\right)$ be the sample maximum likelihood estimator of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{V})$ based on any of the equations (39–42) and $\hat{\boldsymbol{\theta}}_b = \left(\hat{\boldsymbol{\beta}}_b, \hat{\mathbf{V}}_b\right)$ be the ML estimator computed from the bootstrap sample $b = 1,..., B$, with the same sample size, drawn by simple random sampling with replacement from the original sample – the sample drawn under informative sampling design. The bootstrap variance estimator of $\hat{\boldsymbol{\theta}} = \left(\hat{\boldsymbol{\beta}}, \hat{\mathbf{V}}\right)$ is defined as:

$$\hat{V}_{boot}\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\theta}}_b - \bar{\hat{\boldsymbol{\theta}}}_{boot}\right)\left(\hat{\boldsymbol{\theta}}_b - \bar{\hat{\boldsymbol{\theta}}}_{boot}\right)' \tag{44}$$

where

$$\bar{\hat{\boldsymbol{\theta}}}_{boot} = \frac{1}{B}\sum_{b}^{B}\hat{\boldsymbol{\theta}}_b \, .$$

5. Conclusions

In this paper, we extend the definition of univariate sample distribution into multivariate random variables. Also, we consider a new method of estimating the parameters of the superpopulation model for analyzing multivariate normal observations from finite population when the sampling design is informative. Furthermore, the general linear model for longitudinal survey data under informative sampling using different covariance structures: the exponential correlation model, the uniform correlation model, and the random effect model, was fitted under informative sampling.

The main feature of the present estimators is their behaviours in terms of the informativeness parameters.

The paper is purely mathematical. The role of informativeness of sampling mechanism in adjusting various estimators for bias reduction, based on simulation study, under different population models and different modeling of conditional expectations of first order inclusion probabilities given response variable and covariates, can be found in Pfeffermann and Sverchkov (1999, 2003), Nathan and Eideh (2004), Eideh (2008), and Eideh and Nathan (2006, 2009).

I hope that the new mathematical results obtained will encourage further theoretical, empirical and practical research in these directions.

REFERENCES

BIRNBAUM Z.W., PAULSON E., and ANDREWS F.C. (1950). On the Effect of Selection Performed on Some Coordinates of a Multi-Dimensional Population. Psychometrika, 15, pp 191–204.

CHAMBERS, R. and SKINNER, C. (2003). *Analysis of Survey Data*. New York: John Wiley.

DIGGLE, P. J., LIANG, K. Y, and ZEGER, S. L. (1994). *Analysis of Longitudinal Data*. Oxford: Science Publication.

EIDEH A.H. (2008). Estimation and Prediction of Random Effects Models for Longitudinal Survey Data under Informative Sampling. *Statistics in Transition – New Series .Volume 9, Number 3, December 2008*, pp 485–502.

EIDEH, A. H. and NATHAN, G. (2009). Two-Stage Informative Cluster Sampling with application in Small Area Estimation. *Journal of Statistical Planning and Inference.139*, pp 3088–3101.

EIDEH, A. H. and NATHAN, G. (2006) Fitting Time Series Models for Longitudinal Survey Data under Informative Sampling. *Journal of Statistical Planning and Inference*, 136, 3052–3069.

FAHRMEIR, L. and TUTZ, G. (2001). *Multivariate statistical modeling based on generalized linear models*, 2nd edn. New York: Springer.

HERRIOT, R.A., and KASPRZYK, D. (1984). The survey of income and program participation. *American Statistical Association, Proceedings of the Social Statistics Section*, pp.107–116.

HOEM, J.M. (1989). The issue of weights in panel surveys of individual behaviors. In Panel Surveys, (Eds.), Kasprzyk, D., Duncan, G.J., Kalton, G., and Singh, M.P., New York: Wiley, pp. 539–565.

LOHNSON, R.A., and WICHERN, D.W. (1998). *Applied Multivariate Statistical Analysis, 4th edn*. New Jersey: Prentice Hall.

KASPRZYK, D., DUNCAN, G.J, KALTON, G., and SINGH, M.P. (Eds.) (1989). Panel Surveys. New York: Wiley.

KRIEGER, A.M, and PFEFFERMANN, D. (1997) Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*. 13: 123–142.

MARDIA, K.V., KENT, T.J. and BIBBY, J.M. (1979). *Multivariate analysis*, New York: Academic Press.

NATHAN, G. (1999). A review of sample attrition and representativeness in three longitudinal surveys. GSS methodology Series No. 13. London: Office of National Statistics.

NATHAN, G. and EIDEH, A. H. (2004). L'analyse des données issues des enquêtes longitudinales sous un plan de sondage informatif. in**:** *Échantillonage et Méthodes d'Enquêtes*. (P. Ardilly – Ed.) Paris: Dunod, pp 227-240.

PFEFFERMANN, D. (1993). The role of sampling weight when modeling survey data. International Statistical Review 61: 317–337.

PFEFFERMANN, D. (1996). The use of sampling weights for survey data analysis. Statistical Methods in Medical Research, V.5: 239–261.

PFEFFERMANN, D., KRIEGER, A. M, and RINOTT, Y. (1998). Parametric Distributions of Complex Survey Data under Informative Probability Sampling. *Statistica Sinica*, 8, 1087 1114.

PFEFFERMANN, D. and SVERCHKOV, M. (1999). Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data. *Sankhya***,** 61, Ser. B, 66–186.

PFEFFERMANN, D. and SVERCHKOV, M. (2003). Fitting Generalized Linear Models under Informative Probability Sampling. *Analysis of Survey Data*. (eds. R. Chambers and C. J. Skinner), pp. 175–195. New York: Wiley.

SKINNER, C.J., HOLT, D., and SMITH, T.M.F (Eds.) (1989). Analysis of Complex Surveys, New York: Wiley.

SKINNER, C.J., and HOLMES, D. (2003). Random Effects Models for Longitudinal Data. *Analysis of Survey Data***.** (eds. R. Chambers and C. Skinner), pp. 175–195. New York: Wiley.

SMITH, T.M.F. and HOLMES, D.J. (1989). Multivariate analysis. In *Analysis of Complex Surve*ys, eds. C.J. Skinner, D. Holt and T.M.F. Smith, New York: Wiley, pp. 165–190.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 539—554

MODELS IN SURVEY SAMPLING

Carl-Erik Särndal[1]

ABSTRACT

Models, especially in the form of assumed relationships between study variables and auxiliary variables, have influenced survey sampling theory and practice over the last four decades. Some of the early debates between the design-based school and the model-based school are revisited. In their pure forms, they offer two fundamentally different outlooks and approaches to inference in sample surveys. Complete reconciliation and agreement cannot be expected. But the tendency today is that each of the two approaches recognizes and profits from important elements in the other. We see an often fruitful interaction, as discussed in this article.

1. A polarization occurs

The objective in this article is to reflect on a topic that has received much attention, in discussions at conferences and seminars and, more formally, in articles and books. I am referring to the role of models in survey sampling theory and practice. My remarks do not attempt to paint a complete picture. They represent a few personal impressions and conclusions about a development in which I participated. Many important developments go unmentioned in the text that follows.

Behind the topic (in its modern aspect) lies a roughly forty year old split, or scientific conflict if one prefers. To assign a time span of forty years is of course rather arbitrary, but not without some good justification.

Among opinions that have been expressed: Why should survey samplers be different? Just about every other branch of statistics is built around modeling, why should survey samplers resist? Those who persist in the established tradition of survey sampling (the design-based framework), are they not keeping in step with the times? But recent years have brought a change: models are now well engrained in the design-based philosophy and practice as well.

Thus, a polarization occurred around four decades ago: design-based inference became contrasted with model-based (or model dependent) inference. These terms were not in common use before 1970. Today they are standard usage,

[1] Statistics Sweden, S-70189 Örebro, Sweden; carl.sarndal@scb.se.

not only among specialists in survey sampling, but among other categories of scientists as well. After all these years, the issue is not settled; neither side is a winner.

Once an awareness had been created, and the basic differences between the two approaches had been made clear, many were attracted to the topic. There was a period of intellectual curiosity. What are, more precisely, the differences? Which estimators are favored (have better accuracy) under one or the other approach? Some took a categorical stand in favour of one approach, convinced that the other was wrong. Others were more neutral, content to understand and appreciate each approach for what it is, without taking any demonstrative position in favour of one or the other.

The early debates in the 1970's and 1980's took place for the most part in the arena of those "pure survey conditions" that I mention later. The survey practitioner, who seldom encounters pure conditions, feels uneasy at times about the incapacity of the standard theory (the design-based one) to address all the contingencies – nonresponse and many others – that affect surveys today.

Much has been written on the concept of models, and on their function, by authors with a philosophy of science perspective on economics, sociology or other disciplines. Serious model building is a maturing process, involving, over a long time, a thorough interplay between a theoretical content and empirical content. But in survey sampling, "the model" is often "a default statistical model", a specification, suitable for the moment, of a formula for a hypothetical relationship between a study variable *y* (one of the usually many in a survey) and those other variables called auxiliary, because something is known about them at a level beyond the sample itself. When the survey methodologist says "should we model it", he/she is wondering whether to bring in some assumptions of relationship, unverifiable but not entirely out of place, to save survey resources or to bypass other practical difficulties.

2. The early debates

Two papers that made an impression on me as a young man, and on many others, were Brewer (1963) and Royall (1970). The first is titled *Ratio estimation and finite population: some results deductible from the assumption of an underlying stochastic process*. The second is titled *On finite population sampling under certain linear regression models*. (I cite full titles, because they in themselves convey a message.) While the second is traditionally cited, I personally assign much value also to Brewer's 1963 contribution, a feeling underscored also by Brewer's relentless and insightful efforts later to combine, and to seek the positive side of, both approaches, as I discuss later.

In those articles, Brewer and Royall examine a basic situation: The estimation of a population total $Y = \sum_U y_k$, where y_k is observed for all units $k \in s$, where s is a probability sample from $U = \{1, 2, ..., N\}$, and where population totals are known for one or more auxiliary variables. (I focus in this article on the estimation of finite population parameters called descriptive, such as totals, means and functions of totals. Inference about super-population parameters, called analytic, is a different story. Also, I do not address issues arising in the important area of longitudinal surveys.)

The design-based approach (although not referred to by that name until later) had become the ruling methodology, following Neyman (1934) and contributions in its footsteps during the 1940's and 1950's. In this approach, the probability structure for inferences comes from the randomization distribution, from the probabilities with which different samples are potentially drawn (although one and only one is realized in a survey). The statistical properties (mean, variance and so on) of an estimate are evaluated by averaging over all possible samples under the given sampling design. It is an unconditional distribution.

By contrast, Royall (1970) and his followers present a serious attempt to construct inferences from an alternative source, the model alone, conditionally on the set, *s*, of units sampled from *U* in some way, not necessarily by probability sampling. Although Royall did not use the term, "model-based inference" quickly became a recognized concept and part of the standard terminology. This approach revived the old controversy dating back to the beginning of the 20th century. Forms of purposive (non-probability) sampling and balanced sampling had already been practiced in the 1890's; the Norwegian experience attributed to Kiaer is frequently cited.

3. The literature in a context of pure conditions

A sometimes heated debate took place 1970-1990, principally among sampling theoreticians. The arena was one of "pure conditions", or one may call it "debate on the foundations".

Pure conditions address (with minor variations) the following situation: A probability sample s is drawn from the finite population $U = \{1, 2, ..., k, ..., N\}$. The known design weight for unit *k* is $d_k = 1/\pi_k$, where $\pi_k = \Pr(k \in s) > 0$ is the inclusion probability (whether to use the π_k or not in inference is a divisive question). The value y_k of the study variable *y* is recorded for all $k \in s$. The objective is to estimate the population total $Y = \sum_U y_k$. Auxiliary information will be used, usually so that $\mathbf{x}_k$ is an auxiliary vector value known for $k \in U$ (or, at a minimum, so that the total $\sum_U \mathbf{x}_k$ is known, imported from an accurate source). Nonresponse, measurement error, frame error and other non-sampling errors are absent. There is only *one* study variable *y*, although in practice most surveys have many. No particular country, no particular survey is addressed; we are concerned with bare basics, the foundations.

Following the early debates in the 1970's, many articles have been written on some aspect of estimation under pure conditions. They are valuable contributions to the literature, especially in the light of recent advances such as multilevel modeling, non-parametric regression modeling and others. But pure conditions have little to do with today's harsh survey conditions with high nonresponse, frame errors and other imperfections. Still it is appropriate today to place a piece of research within the context of the pure conditions, and develop one's topic within either of the two paradigms, the design-based one or the model-based (model dependent) one, and to get one's theoretical article accepted in the best of journals.

In the decades following 1970 one took great interest in comparing estimators generated by one or the other approach, and, more importantly, in comparing the properties (bias, variance, and so on) under one or the other of the two frameworks. As a typical example of that period, I was concerned in Särndal (1978), *Design-based and model-based inference in survey sampling*, with exploring the possibilities under the two modes. I was not alone in this type of endeavor.

4. Passionate stands, and modeling as an act of taking responsibility

In their article *An evaluation of model-dependent and probability sampling inferences in sample surveys*, Hansen, Madow and Tepping (1983) took a spirited stand in favour of probability sampling inference (that is, design-based inference), as opposed to a model-dependent (model-based) inference. It was an attack perhaps not so much on the use of models as a concern about the lack of robustness of the model-based estimation. (Hansen and collaborators were not opposed to the idea of "giving models their just and appropriate place"; they had shown the rich possibilities of model oriented reasoning in their work on total survey error models.) As a discussant of that paper, Royall passionately defended his model-based view.

An influential participant in the early debates, Smith (1976a) also strongly favoured the model-based outlook: "The basic question to ask is why should finite population inference be different from inferences made in the rest of statistics? … My view is that survey statisticians should accept their responsibility for providing stochastic models for finite populations in the same way as statisticians in the experimental sciences". Later, Smith's position was to change dramatically, no doubt a result of careful inspection.

As Brewer (1999) notes, "Royall argued that survey sampling was out of step with statistics as a whole. Statisticians working in other fields used their data to build models and analyzed them in those terms … but survey statisticians were using an entirely irrelevant source of probability structure not related to the data themselves but only to the manner in which they had been collected."

The thought that design-based survey statisticians might fail to "take responsibility" for their design-based estimates is somewhat of an antithesis. Morris Hansen and other influential proponents of the design-based mode, from the 1940's and on, were the opposite of "not responsible", conscientious as they were to provide policy makers and authorities in government with impartial, defendable statistics. In fact, why should survey statisticians not be different from the experimentalists? Experiments are typically small, with relatively few units undergoing treatment, and focus is on hypothesis testing of treatment differences. By contrast, government surveys are large, multipurpose, and they cater to a different type of user.

5. Is reconciliation possible?

Now, if indeed differences exist, are they of crucial importance? Did the debate exaggerate them? Can the two approaches be reconciled? The verb "to reconcile" means both (a) to return to friendly relations and harmony, and (b) to make consistent or congruous or compatible. Both meanings are relevant here. Smith (1994) titles his Hansen lecture *Sample surveys 1975-1990; An age of reconciliation*? A hint of "not reconciled but coming fairly close" lies in the title of Brewer, Hanif and Tam (1988), *How nearly can model-based prediction and design-based estimation be reconciled*?

Smith (1994) notes "In the absence of models for the underlying social processes which are generally held to be true, model-based inferences lose all their desirable properties." A discussant at that occasion also, Royall (1994) persisted: "Professor Smith's paper … is an announcement of a dramatic change in his own thinking. … I will simply try to show that there *must* be errors in his reasoning, because his conclusion is wrong … He goes to the shocking extreme of advising … what is clearly *wrong*, namely quoting the unconditional standard error."

The underlying statistical principles differ fundamentally; to make the two thought processes consistent and compatible is simply not possible: Design-based inference is "unconditional, referring to all possible probability samples under the given sampling design"; model-based inference "conditional, under the model, for the one and only realized sample". That is an irreconcilable difference. Now, in a given situation, under comparable conditions, the estimators delivered by the two approaches may agree. But the variances and the mean squared errors do not generally agree. They refer to different conceptions of "long run repetitions", over all possible samples in the design-based case, over all finite populations generated under the model in the other case. Nevertheless one can, in some situations at least, take action to also make measures of precision acceptable from both vantage points, as for example Särndal, Swensson and Wretman (1989) show in

The weighted residual technique for estimating the variance of the general regression estimator of the finite population total.

Not surprisingly, the literature has examined the compromise of “let us average over both”, that is, over both the distribution determined by the assumed model and the distribution determined by the randomized sample selection. One is led to consider the concept of anticipated variance (model-expected design-expected mean squared error).

Brewer in particular has contributed much to a vision of making the inferences palatable from both angles, the design-based one and the model-based one; Brewer (1999) argues “there is something substantial to be gained by using them in combination”, in a single estimator defendable from both sides. This theme is also developed in Brewer (1995), *Combining design-based and model-based inference* and in his book, Brewer (2002), *Combined survey sampling inference; weighing Basu’s elephants*.

The survey sampling literature has thus come to harbour two streams with fundamentally different starting points. But how does an individual statistician react? Must he/she choose sides? Can a self-respecting survey sampling statistician embrace both lines of thought, defending sometimes one, sometimes the other approach? I have observed colleagues and friends adopt different attitudes in regard to those questions. One option is to switch sides, depending on the practical problem at hand, as for example when one defends model-dependent inference in regard to the small area problem, while advocating, in other instances, a design-based approach. As for myself, I experience some difficulty with subscribing to a “double-natured ethic”. An adaptable attitude is however suggested in Smith’s (1994) remark: “My overall conclusion is that there is no single paradigm for statistical inference and that different classes of problems require different solutions. Instead of looking for unity we should concentrate on identifying the differences and enjoy the diversity of our subject. Complete reconciliation is neither possible nor desirable. Vive la différence.” But Smith’s own preference is stated in these words: “The case I am making is for procedural inference, and this refers to the unconditional randomization distribution. I now find the case for hard-line randomization inference based on the unconditional distribution to be acceptable … I now think that the framework for descriptive inference should be the unconditional distribution relating to the original sampling procedure.” This corresponds well with my own thinking since the early 1980’s.

6. Awareness not without risk of confusion.

The opposition design-based vs. model based may be perfectly clear and simple to experts in survey sampling, those who have regularly monitored the theoretical literature over the past 30 years. But have the discussions led by survey theoreticians created undesirable confusion in other fields, among scientists who rely on statistics and understand a great deal about survey design and estimation, without being specialists? Some signs suggest that this has indeed occurred. Design-based inference is unconditional, model-based inference is conditional, that sounds simple, but explaining the implications to specialists in applied fields may not be a trivial task. In any case, it is interesting to note that the debate on the foundations among survey sampling specialists has reached beyond the science of statistics itself.

Forest research is an area where sampling has a long tradition, for example for the estimation of tree volume. A thorough review destined to researchers in that discipline is Gregoire (1998), *Design-based and model-based inference in survey sampling: appreciating the difference.* Although the article is essentially on the side of the design-based tradition, it is noted that recent times have led to a debate also in that discipline: "… current literature in forestry and ecology indicate much ongoing confusion about the distinction between these two modes of inference. A failure to appreciate the underpinnings of one or the other mode of inference could lead to needless abandonment of a survey design that might otherwise be ideal for purposes of scientific inquiry."

Other examples of "uncertainty and confusion" could be mentioned. From the field of neuromorphology comes an article by Geuna (2000) titled *Appreciating the difference between design-based and model-based sampling strategies in quantitative morphology of the nervous system.* This author states "New technical procedures have been devised and applied successfully to neuromorphological research. However … one element of confusion is related to uncertainty about the meaning, implications, and advantages of the design-based sampling strategy that characterize the new techniques." In this scientific discipline, the design-based alternative appears to be the novelty, a challenger to an earlier established model-based view. This confirms the impression that in any given science, one of the two views is the well entrenched one, the other being an intruder that must fight a battle to be recognized.

To explain (to readers other than the experts) the basic difference between the unconditional (design-based) and the conditional (model-based) mode may be comparatively easy. However, the task of explaining is further complicated by the rise in the last 25 years of mixture modes: Design-based inference can be (and is usually) model assisted, while, on the other side, the model-based inference can or should account for the randomized sample selection. The next few sections examine those proliferations.

7. They cannot live without one another

How did models influence the classical design-based survey statisticians, those who built the design-based tradition, by books and influential articles, in the 1940's to 1960's? By all signs, models had a place in their thinking, but they were "not explicit". These writers did not abhor models, but apparently they abstained, except in rare occasions, from stating them, as if there was no need to be explicit. For example, models were used in a variety of ways to design samples, in modeling (or, less pretentiously, just guessing) stratum variances to determine strata sampling fractions.

Now today, given that the two fundamentally different lines of thought exist, can the proponents of the design-based vision proceed effectively without reference to models? Can proponents of the model-based approach work without any reference to the features of the randomized sample selection? The answer is "no" in both cases. It is very hard to maintain that "elements of the other" should not be present; only the most "hard-core defendants" (if they still exist at this time) of either approach would pretend otherwise.

In each approach, the need is felt to "account for the other", to integrate aspects of the other. For design-based people, the challenge was, from the 1980's and on, to make models explicit in the formal presentation. For the model-based people the challenge was, and still is, to make the formal presentation (notably the model statements) reflect and incorporate the randomized sample selection.

Brewer (1999) describes how design-based theory underwent a change from implicit to explicit usage of models: "The ratio estimator … provides a good example of the way in which models of the population were long used in an implicit fashion by design-oriented survey statisticians. The essential difference between ordinary design-based inference and model assisted survey sampling is that the latter brings such implicit assumptions into the open … Early examples may be found in Cochran (1953, 1963, 1977), Brewer (1963) and Foreman and Brewer (1971), but it was definitely established as the dominant version of the design-based approach with the publication of Särndal, Swensson and Wretman (1992)." It was apparent to my co-authors and me, in the mid-1980's when essential parts of *Model Assisted Survey Sampling* were written, that "the marriage had to take place". We opted for the alternative "design-based assisted by models". A broader perspective opened up for models inside the design-based framework; "the dominant version" it became, perhaps, but I like to see it as a "necessary version", considering how those of us in design-based sampling theory experienced the field in the 1980's.

Another example of design-based accommodation to model features becomes apparent when we look at the sample selection stage. The appealing idea of balanced sampling resided in the tradition of purposive, non-randomized selection, that is, the idea that sample means of auxiliary variables should, roughly at least, equal their known population counterparts. The challenge is: Find a proper probability sampling device (thus placed within randomization theory) that is certain to deliver a balanced sample. An answer is the cube method of Deville and Tillé (2004); by ingenious randomization, we select only among samples guaranteed to be (almost) balanced; the randomization device seeks only among "good, balanced samples" and selects one of those.

So it was almost by necessity that the scientific evolution brought, as one possibility, the model assisted design-based perspective and, as a second possibility, a model-based (model dependent) design-influenced perspective. The latter calls for integrating the features of the randomized sample selection (stratified sampling, two-stage sampling and so on) in the model statement. For example, Smith (1976b), in his model-based period, challenges the design-based model assisted GREG (for generalized regression estimator) in Cassel, Särndal and Wretman (1976): "Why should the selection probabilities, *p*, take any precedence over the model ξ? … The design *p* is at the choice of the statistician and would usually be based on prior information about the population which should already be embodied in ξ".

The second possibility is also explicit in Kott (2005), *Randomization-assisted model-based survey sampling*. His position, "long espoused in public", is that "the dominant model-assisted (randomization-based) survey sampling paradigm, although fruitful in many ways, should be supplanted by a randomization-assisted model-based one. That is because inference should be based on the sample actually observed rather than averaged over all potential samples."

Also in a model-based vein, Little (2003) states, in *To model or not to model? Competing modes of inference for finite population sampling*: "Models need to properly reflect features of the sample design such as weighing, stratification and clustering, or (model-based) inferences are likely to be distorted" . He points out that models of high complexity can now be entertained: "Computational power has expanded dramatically since the days of early model versus randomization debates, and much can be accomplished using software for mixed models in the major statistical packages." But needless to say, these advanced models are now also incorporated in the design-based model assisted framework.

8. Models as assistants

The model assisted design-based approach is thus a means of bringing model features into the open without upsetting the design-based basis for inference. The characteristics of the model are not crucial to the validity of the design-based inferences.

In the randomization/model marriage that produced the model assisted outlook, the models play an obedient role. A relationship between *y* and **x** is taken into account, but "its truth" is not essential; it remains in the background, lacking the influential role it would have in model-based inference.

Science thrives by arguments and counterarguments. As one can expect, criticism has been levied of this seemingly subservient role of the model: The design-based inferences are recognized as "robust", but may be less efficient than they could be; it is argued that the full potential of "a correct model" is not realized, that the model denied its justified, more influential role, and so on.

Smith (1994) points out that "randomizers should not make concessions towards predictive inference". The advent of the model assisted design-based approach granted randomizers this freedom. Nevertheless, this approach still resides essentially within “pure conditions”. It requires, ideally at least, an absence of nonresponse and other non-sampling errors. I return later to the question of survey nonresponse.

9. Looking beyond the pure conditions: The practice of survey sampling

Three articles by Kalton (1981, 1983, 2002) are titled *Models in the practice of survey sampling*, the third one with the addition “revisited”. I examine these contributions with particular interest; they are highly relevant in a discussion on the role of models in survey sampling, because they examine the confrontation of theory with thorny practical matters, leaving aside the theoretician’s predilection for “pure conditions” and “foundations”.

Kalton (2002) notes that: “Models are widely used within the design-based mode of inference, both in sample design and in estimation, but in a “model-assisted” manner so that the validity of the survey estimates does not depend on the validity of the model assumptions.” However, as this author goes on to say, “Valid design-based inferences require that nonresponse and other non-sampling errors be of next-to-negligible extent … The design-based approach … for descriptive analysis of large scale surveys … cannot fully address all problems of making inferences … Although design-based inference is the standard form of inference with large-scale sample surveys, in practice some reliance on model-dependent inference is necessary.” These phrases reflect a regret that the standard theory (the design-based one), although preferred and held in high regard by practitioners, does not provide answers for all circumstances. The words “cannot fully assess address all the problems” and “some reliance on models” are crucial. What role, more precisely, should then be attributed to models?

10. Models as crutches

Kalton (2002) states: “My general approach to the use of model-dependent methods for descriptive estimation is to treat the model as a crutch, to be used only to the extent that the survey data cannot fully support the desired estimates. If the sample is strong enough, and if there is no weakness from missing data, then design-based inferences alone will serve well.”

“To treat the model as a crutch” is a colorful image. It is not entirely misplaced to view the design-based theory as a handicapped guide for practice, one that is in need of support from artificial devices (models) to meet the needs. A certain gap exists between the practical reality and the pure design-based theory that is supposed to back up that practice.

This suggests an approach, rather informal and unstructured, in which models are brought to bear, whenever needed to supplement or mend the ailing theory. On the other hand, it might imply a somewhat uncontrolled “model interference” to overcome survey imperfections. Also, the pure concepts of design-based unbiasedness and design-based variance become compromised. For example, what value is there in computing the design-based variance of survey estimates when the unknown squared bias (arising for example when missing data are imputed) is likely to be the dominating component of the mean squared error?

The weakness "sample not strong enough" is invoked in practically all the literature on model dependent small area estimation, where it breeds the idea of "borrowing strength" from data outside the domain, something which necessarily brings a presence of models in the construction of a small area estimator.

11. Critical areas: Small area estimation and nonresponse

In design-based practice, the predicament of "sample not strong enough" occurs often. Kalton (2002) continues: "… many of the developments in survey sampling in the past quarter century have been concerned with the application of model-dependent methods to address such problems as missing data and small area estimation". To my mind also, missing data and small area estimation, have clearly stood out (and still do) as two significant and critical areas (or testing grounds) for the reliance on models in survey sampling. They do so for different reasons.

Design-based inference, the standard in practice, grew out of a theory for essentially pure conditions; it "cannot fully address" estimation with missing data or for small areas. The reasons for failure are different in the two cases: For very small areas or sub-populations, the total sample size is insufficient to deliver acceptable design-based precision. For the missing data (or nonresponse) problem, it is the unknown response mechanism and the unknown response probabilities that are the root of the problem.

In both cases, taking measures to make the standard theory functional is viewed not as a scientific necessity, but rather as being too expensive. Increasing the response rate, by follow-ups and other measures, to reach "decent levels of response" is deemed prohibitively expensive, as is an increase in total sample size up to a point where well-supported design-based estimates become possible not only for large sub-populations but also for small ones. But the statistician is expected to cope with the poor prospects.

Kalton (2002) notes: "From the design-based perspective, the approach to estimation for small areas is first to seek direct estimators of adequate precision, making full use of auxiliary information in a model-assisted way. When that approach fails, it becomes necessary to resort to indirect estimators that depend on statistical models." The approach fails, because in our time pressure is imminent to produce numbers on any tiny group in society, without a willingness to supply the funds needed to do that with adequate precision. Hence, model-dependent small area estimation theory becomes one of the crutches, to use that colorful term; that theory drops from the start any ambition to build inferences on the randomization distribution.

The missing data syndrome is one from which practically all surveys suffer today. Nonresponse rates are very high and all the time increasing. The dominating line of thought is to mend, or repair, the design-based inferences through imputation and/or nonresponse weighting adjustment. Both are viewed as forms of “model intervention”. I return to the nonresponse question in Section 13.

12. The calibration perspective

Another angle of the questions of “models or not?” and “models to what extent?” becomes apparent with the increased popularity in the last fifteen years of calibration theory and practice. The procedure of attaching weights to the observed values of a study variable, and to sum the weighted values, has always appealed to survey samplers, especially the practitioners. The idea to make the weight system respect “control totals” is also old. Weighting by poststratification is among the simplest and oldest examples of calibration. Deming (1943), in the book *Statistical Adjustment of Data*, focused on making the weights conform to marginal counts of cross-classified tables; this became known as the raking ratio method. Long tradition and expertise in “controlled weighting” are evident at the US Bureau of the Census, as seen for example in Alexander (1987), and in other important survey organizations, such as I.N.S.E.E. in France. The modern term is *calibration* (on auxiliary information); Deville and Särndal (1992), in *Calibration estimators in survey sampling*, contributed a more general and more complete view of that methodology.

To survey practitioners, the prime motivation for calibration is not as much the potential for increased accuracy as rather the desire to “achieve consistency”, within a design-based perspective, with known or estimated statistics from other sources or survey occasions. In that frame of mind, a modeling of y-to-**x** relationships is secondary. Nevertheless, the aims of calibration are twofold: (i) realizing consistency and (ii) improving accuracy, reducing both bias (due to nonresponse for example) and variance.

The practitioner understands perfectly well the motivation behind calibration and its results. The theoretician on the other hand is looking for models behind the procedure and when he/she has difficulty identifying them, a natural reaction is that “models are there somehow, but they are skillfully hidden”. Thus estimation by calibration brings the question of “the role of models” to a critical test.

The model assisted design-based GREG estimator of the 1980’s was an offspring of the idea that a relationship between a study variable y and the auxiliary vector $\mathbf{x}$ can be exploited to improve precision; that was the prime motivation. It was my view, too, and I arrived at calibration through the back door, so to speak. I recall my surprise, almost stupefaction, at my insight in the early 1980’s that the model assisted design-based GREG has an expression as a weighted sum, $\sum_s w_k y_k$, with a weight system that has the intriguing property $\sum_s w_k \mathbf{x}_k = \mathbf{X}$, stating that the weights w_k for $k \in s$ are consistent with the auxiliary information $\mathbf{X}$ for the survey, where for example $\mathbf{X} = \sum_U \mathbf{x}_k$, a known vector total for the population. Today this calibration property is a self-evident and trivial consequence. I was surprised because that property seemed so remote from, or incongruent with, my vision at the time of the relationship *y*-to-**x** (and the prospect of a reduced variance) as the sole driving force. But one's vision can change, as exemplified earlier in this article.

I find the essence of calibration theory particularly striking when applied to estimation for surveys with nonresponse. There the question of "use of models or not" receives a different twist, contrasting with the common view, noted earlier, that nonresponse necessitates a reliance on modeling, notably a modeling of the unknown response mechanism.

13. Calibration for nonresponse adjustment

A probability sample s is drawn from the population $U = \{1, 2, ..., k, ..., N\}$. The known design weight is $d_k = 1/\pi_k$. Nonresponse occurs; the study variable value y_k is observed only for the response set r, where $r \subset s \subset U$. How r was generated is unknown. With suitably specified weights w_k, an estimator $\hat{Y} = \sum_r w_k y_k$ of $Y = \sum_U y_k$ is computed on the data y_k observed only for $k \in r$. In particular, we can use calibrated weights, made to satisfy $\sum_r w_k \mathbf{x}_k = \mathbf{X}$, where $\mathbf{X}$ summarizes the appropriate information for the auxiliary vector $\mathbf{x}_k$; such weights are $w_k = d_k\{1 + (\mathbf{X} - \sum_r d_k \mathbf{x}_k)'(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k)^{-1}\mathbf{x}_k\}$. As presented in for example Särndal and Lundström (2005), *Estimation in surveys with nonresponse,* the information may be of two kinds: At the population level, transmitted by the vector $\mathbf{x}^*_k$, and at the sample level, transmitted by $\mathbf{x}^\circ_k$, both vector values known for all $k \in s$, that is, for respondents and for nonrespondents. The population total $\sum_U \mathbf{x}^*_k$ is known, consisting for example of (updated) census counts on groups based on age, sex and region. The total $\sum_U \mathbf{x}^\circ_k$ is unknown but is estimated without bias by $\sum_s d_k \mathbf{x}^\circ_k$, where $\mathbf{x}^\circ_k$ may express features of the data collection and other survey operations. Then we compute the calibrated weights so that $\sum_r w_k \mathbf{x}_k = \mathbf{X}$ holds with $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$ and $\mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$, implying $\sum_r w_k \mathbf{x}_k^* = \sum_U \mathbf{x}_k^*$ and $\sum_r w_k \mathbf{x}_k^\circ = \sum_s d_k \mathbf{x}_k^\circ$. In this procedure $\mathbf{X}$ represents the “clearly stated information” that one decides to use. That clear message is not enough for the modeler. He/she deplores a lack of a “clearly stated model assumptions”, and may conclude that “calibration is a black box”. In this and other applications, the calibration approach, with its emphasis on the information on which to calibrate rather than on modeled relationships, is received with suspicion by some.

14. Conclusion

The design-based perspective on survey sampling was, at least in its earlier years, hailed as “the scientific approach” to inquiries about human or other finite populations. The period 1940 to 1960 was a break-through, the heyday of the clean, trustworthy design-based theory. Today, judicious survey statisticians feel, with some justification, that survey sampling can be redeemed as a scientific field for future generations only by making it fully appreciated by “the other branches of statistics”. There is a belief that this must necessarily happen, through a use of ever more advanced models. I am not convinced. Advances in modeling have taken place in the past few decades. Still, my impression is that while the models serve a useful purpose in the design-based model assisted theory, they must become much stronger to be “good enough” for a model-based theory, stronger not only in their mathematical formulation, but also in their capacity to grasp the real nature (the underlying social processes) of a relationship y-to-$\mathbf{x}$, and this not only for one isolated y-variable, but for every one of those many that partake in a large survey.

REFERENCES

ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183–188.

BREWER, K.R.W. (1963). Ratio estimation and finite population: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, 5, 93–105.

BREWER, K.R.W. (1995). Combining design-based and model-based inference. In Cox, B.G. et al. (editors), *Business Survey Methods*. New York: Wiley, 589–606.

BREWER, K.R.W. (1999). Design-based or prediction inference? Stratified random vs stratified balanced sampling. *International Statistical Review,* 67, 35–47.

BREWER, K.R.W. (2002) *Combined Survey Sampling Inference.Weighing Basu's Elephants*. London: Arnold.

BREWER, K.R.W., HANIF, M. and TAM, S.M. (1988). How nearly can model-based prediction and design-based estimation be reconciled? *Journal of the American Statistical Association* 83, 128–32.

CASSEL, C.M, SÄRNDAL, C.E. and WRETMAN, J.H. (1976) Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* **63**, 615–620.

COCHRAN, W.G. (1953, 1963, 1977). *Sampling Techniques*. New York: Wiley; 1st, 2nd, 3rd edition.

DEMING, W.E. (1943). *Statistical Adjustment of Data*. New York: Wiley.

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

DEVILLE, J.C. and TILLÉ, Y. (2004). Efficient balanced sampling: The Cube Method. *Biometrika,* 91, 893–912.

FOREMAN, E.K. and BREWER, K.R.W. (1971). The efficient use of supplementary information in standard sampling procedures. *Journal of the Royal Statistical Society, series B,* 33, 391–400.

GEUNA, S. (2000). Appreciating the difference between design-based and model-based sampling strategies in quantitative morphology of the nervous system. *The Journal of Comparative Neurology,* 427, 333–339.

GREGOIRE, T.G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research,* 28, 1429–1447.

HANSEN, M.H., MADOW, W.G. and TEPPING, B.J. (1983). An evaluation of model-dependent an probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association,* 78, 776-807.

KALTON, G. (1981). Models in the practice of survey sampling. Invited paper, 43rd session of the International Statistical Institute, Buenos Aires.

KALTON, G. (1983). Models in the practice of survey sampling. *International Statistical Review,* 51, 175–188.

KALTON, G. (2002). Models in the practice of survey sampling (revisited). *Journal of Official Statistics,* 18, 129–154.

KOTT, P.S. (2005). Randomization-assisted model-based survey sampling. *Journal of Statistical Planning and Inference*, 129, 263–277.

LITTLE, R.J. (2003). To model or not to model? Competing modes of inference for finite population sampling. University of Michigan, Department of Biostatistics, working paper 2003:4.

NEYMAN, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.

ROYALL, R.M. (1970). On finite population sampling under certain linear regression models. *Biometrika*, 57, 377–387.

ROYALL, R.M. (1994). Discussion of Smith (1994). *International Statistical Review*, 62, 19–20.

SÄRNDAL, C.E. (1978). Design-based and model-based inference in survey sampling. *Scandinavian Journal of Statistics*, 5, 27–52.

SÄRNDAL, C.E., SWENSSON, B. and WRETMAN, J. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76, 527–537.

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

SÄRNDAL, C.E. AND LUNDSTRÖM, S. (2005). *Estimation in surveys with nonresponse*. New York: Wiley.

SMITH, T.M.F. (1976a). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society A*, 139, 183–204.

SMITH, T.M.F. (1976b). Discussion of Cassel, Särndal and Wretman (1976). *Biometrika* **63**, 620.

SMITH, T.M.F. (1994). Sample surveys 1975–1990; An age of reconciliation? *International Statistical Review*, 62, 5–19.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 555—562

ON EFFICIENT DIFFERENCE TYPE ESTIMATORS

A. K. P. C. Swain[1]

ABSTRACT

This paper considers a more efficient difference type estimator in a finite population set-up. Ratio type and regression type estimators are derived as special cases. Further efficiencies of these estimators are compared with classical ratio and regression estimators. Numerical illustrations are provided to compare efficiencies of different competitive estimators.

Key words: Difference type, Ratio type and Regression type estimators, Efficiencies of estimators, Auxiliary information, Simple random sampling without replacement.

1. Introduction

In sample surveys, ratio and regression estimators are often used to make use of auxiliary information to produce more efficient estimators compared to estimator based on simple mean per unit of the study variable.

The literature on ratio and regression methods is quite extensive. Many researchers, such as Bedi and Hajela (1984), Gupta (1978), Pandey (1980), Prasad (1986), Ray and Singh (1981), Rao (1978), Rao (1993), Ray and Sahai (1980), Srivastava (1967), Tailor and Sharma (2009) and others, have come forward to suggest more methods of improving precision of conventional ratio and regression estimators under certain conditions.

In the following we shall consider some more efficient difference type, ratio type and regression type estimators to estimate the population mean of the study variable in the presence of auxiliary information.

[1] Visiting Professor, Vivekananda Institute of Social Works and Social Sciences, Bhubaneswar-751001, India. E-mail: akpcs@rediffmail.com.

2. Generalized Difference type estimator

Let there be a finite population U consisting of N identifiable units $U_1\,, U_2\,, ..., U_N$. On the i^{th} unit a paired value $(Y_i\,,\, X_i)$, $(i = 1, 2,..., \mathrm{N})$ is attached where Y_i and X_i $(i = 1, 2,..., \mathrm{N})$ are realized values of the study variable y and auxiliary variable x respectively.

Define $\overline{Y}$ and $\overline{X}$ as the population means of y and x respectively where $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ and $\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$. For a simple random sample 's' of size n drawn without replacement from the finite population U, define the sample means of y and x as $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ respectively.

In order to estimate the population mean $\overline{Y}$, define the generalized difference type estimator

$$\hat{\overline{Y}}_d^* = \alpha\overline{y} + d\left(\overline{X} - \overline{x}\right), \tag{2.1}$$

where α is a suitably chosen scalar and d is a known constant or a random variable converging in probability to a constant.

$\hat{\overline{Y}}_d^*$ is a biased estimator with bias equal to

$$Bias\left(\hat{\overline{Y}}_d^*\right) = E\left(\hat{\overline{Y}}_d^*\right) - \overline{Y}$$

$$= \alpha\overline{Y} - \overline{Y} = (\alpha - 1)\overline{Y}, \tag{2.2}$$

The mean square error of $\hat{\overline{Y}}_d^*$ is given by

$$MSE\left(\hat{\overline{Y}}_d^*\right) = \alpha^2 V(\overline{y}) + d^2 V(\overline{x}) - 2\alpha d\ Cov\left(\overline{y}\,,\,\overline{x}\right) + (\alpha - 1)^2\,\overline{Y}^2 \tag{2.3}$$

The value of α which minimizes $MSE\left(\hat{\overline{Y}}_d^*\right)$ is

$$\alpha_{opt} = \frac{d\ Cov\left(\overline{y}\,,\,\overline{x}\right) + \overline{Y}^2}{V(\overline{y}) + \overline{Y}^2}, \tag{2.4}$$

Thus, the minimum MSE of $\hat{\overline{Y}}_d^*$ to $O\left(\frac{1}{n}\right)$ is given by

$$MSE\left(\hat{\bar{Y}}_d^*\right)_{\min} \cong \frac{\theta\left(S_y^2+d^2S_x^2-2dS_{xy}\right)}{\left(1+\theta C_y^2\right)^2}, \tag{2.5}$$

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

$$S_{xy}=\frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)\left(X_i-\bar{X}\right),\ C_y^2=\frac{S_y^2}{\bar{Y}^2},$$

and $\theta=\left(\frac{1}{n}-\frac{1}{N}\right)$.

3. Ratio type estimator

Substituting $d=r=\frac{\bar{y}}{\bar{x}}$ in (2.1), the ratio type estimator is given by

$$\hat{\bar{Y}}_R^*=\alpha\bar{y}+r(\bar{X}-\bar{x}). \tag{3.1}$$

Assuming that the bias in r is negligible for large samples, the values of α, which minimizes $MSE\left(\hat{\bar{Y}}_R^*\right)$ is

$$\alpha_{opt}=\frac{1+\theta\rho\ C_yC_x}{1+\theta\ C_y^2}, \tag{3.2}$$

where C_y and C_x are the coefficients of variation of y and x respectively and ρ is the correlation coefficient between y and x.

Thus, the minimum MSE of $\hat{\bar{Y}}_R^*$ to $O\left(\frac{1}{n}\right)$ is,

$$MSE\left(\hat{\bar{Y}}_R^*\right)=\frac{\bar{Y}^2\left[\theta\left(C_y^2+C_x^2-2\rho C_yC_x\right)\right]}{\left(1+\theta C_y^2\right)^2} \tag{3.3}$$

The classical ratio estimator $\hat{\bar{Y}}_R=\left(\frac{\bar{y}}{\bar{x}}\right)\bar{X}$ has the mean square error to $O\left(\frac{1}{n}\right)$, given by

$$MSE\left(\hat{\bar{Y}}_R\right)=\theta\bar{Y}^2\left(C_y^2+C_x^2-2\rho C_yC_x\right). \tag{3.4}$$

It may be verified that $MSE\left(\hat{\bar{Y}}_R^*\right)\leq MSE\left(\hat{\bar{Y}}_R\right)$, showing thereby that $\hat{\bar{Y}}_R^*$ is more efficient than $\hat{\bar{Y}}_R$.

4. Regression type estimators

Putting $d=\hat{\beta}$, where $\hat{\beta}$ is all estimate of regression coefficient of y on x, we have the regression type estimator (Kaur, 1985), given by

$$\hat{\bar{Y}}_{\mathrm{Re}g}^*=\alpha\bar{y}+\hat{\beta}\left(\bar{X}-\bar{x}\right). \tag{4.1}$$

Neglecting bias in $\hat{\beta}$ for large sample, we have

$$Bias\left(\hat{\bar{Y}}_{\mathrm{Re}g}^*\right)=\left(\alpha-1\right)\bar{Y}\ . \tag{4.2}$$

The value of α which minimizes the $MSE\left(\hat{\bar{Y}}_{\mathrm{Re}g}^*\right)$ is

$$\alpha_{opt}=\frac{1+\theta\rho^2C_y^2}{1+\theta C_y^2}\ . \tag{4.3}$$

Thus, the minimum MSE of $\hat{\bar{Y}}_{\mathrm{Re}g}^*$ is

$$MSE\left(\hat{\bar{Y}}_{\mathrm{Re}g}^*\right)=\theta\bar{Y}^2C_y^2\left(1-\rho^2\right)\left(1+\theta C_y^2\right)^{-2}, \tag{4.4}$$

which is less than the approximate MSE of the conventional linear regression estimator $\hat{\bar{Y}}_{\mathrm{Re}g}=\bar{y}+\hat{\beta}\left(\bar{X}-\bar{x}\right)$, given by

$$MSE\left(\hat{\bar{Y}}_{\mathrm{Re}g}\right)=\theta\bar{Y}^2C_y^2\left(1-\rho^2\right). \tag{4.5}$$

However, the estimator $\hat{\bar{Y}}_{\mathrm{Re}g}^*$ happens to be a special case of a generalized estimator

$$\hat{\bar{Y}}_G^*=K_1\bar{y}+K_2\left(\bar{X}-\bar{x}\right), \tag{4.6}$$

where K_1 and K_2 are suitably chosen constants to be determined by minimizing $MSE\left(\hat{\bar{Y}}_G^*\right)$.

Thus,

$$MSE\left(\hat{\bar{Y}}_G^*\right) = K_1^2\ V(\bar{y}) - 2K_1K_2\ Cov(\bar{y}\ ,\ \bar{x}) + K_2^2\ V(\bar{x}) + (K_1 - 1)^2\,\bar{Y}^2 . \quad (4.7)$$

The optimum values of K_1 and K_2 are

$$K_2 = K_1\beta \quad \text{and} \quad K_1 = \left[1 + \theta C_y^2\left(1 - \rho^2\right)\right]^{-1}, \quad (4.8)$$

where β is the population regression coefficient of y on x. As such, the minimum mean square error of $\hat{\bar{Y}}_G^*$ to $O\left(\frac{1}{n}\right)$ is given by

$$MSE\left(\hat{\bar{Y}}_G^*\right)_{\min} = \theta\bar{Y}^2 C_y^2\left(1 - \rho^2\right)\left[1 + \theta C_y^2\left(1 - \rho^2\right)\right]^{-2}, \quad (4.9)$$

which is less than the approximate mean square error of $\hat{\bar{Y}}_{\mathrm{Re}g}$.

Bedi and Hajela (1984) proposed a modified regression estimator

$$\hat{\bar{Y}}_{BH} = w\left(\bar{y} + \beta\left(\bar{X} - \bar{x}\right)\right) \quad (4.10)$$

where w is a suitably chosen constant so as to minimize the mean square error of $\hat{\bar{Y}}_{BH}$. It may be verified that $MSE\left(\hat{\bar{Y}}_{BH}\right)_{\min} = MSE\left(\hat{\bar{Y}}_G^*\right)$

Rao (1978) suggested a regression type estimator

$$\hat{\bar{Y}}_{\mathrm{R}ao}^* = K_1\bar{y} + K_2\bar{x}\,, \quad (4.11)$$

where K_1 and K_2 are constants, chosen optimally subject to condition of unbiasedness i.e. $E\left(\hat{\bar{Y}}_{\mathrm{R}ao}^*\right) = \bar{Y}$, which implies $(K_1 - 1)\bar{Y} + K_2\bar{X} = 0$.

Minimizing $MSE\left(\hat{\bar{Y}}_{\mathrm{R}ao}^*\right)$ subject to condition for unbiasedness gives

$$K_{1(opt)} = \frac{C_x^2 - \rho C_y C_x}{C_y^2 + C_x^2 - 2\rho C_y C_x} \quad \text{and} \quad K_{2(opt)} = R\cdot\frac{C_y^2 - \rho C_y C_x}{C_y^2 + C_x^2 - 2\rho C_y C_x}, \quad (4.12)$$

where $R = \dfrac{\overline{Y}}{\overline{X}}$.

Substituting the optimum values of K_1 and K_2 in the expression for mean square error of $\hat{\overline{Y}}^{*}_{\mathrm{R}ao}$, we have $O\left(\dfrac{1}{n}\right)$,

$$MSE\left(\hat{\overline{Y}}^{*}_{\mathrm{R}ao}\right) = \theta \cdot \overline{Y}^2 \cdot \frac{\left(1-\rho^2\right)C_y^2 C_x^2}{C_y^2 + C_x^2 - 2\rho C_y C_x}. \qquad (4.13)$$

$\hat{\overline{Y}}^{*}_{\mathrm{R}ao}$ is more efficient than the conventional linear regression estimator $\hat{\overline{Y}}_{\mathrm{Re}g}$ if $\rho < \dfrac{1}{2}\dfrac{C_y}{C_x}$, and is more efficient than $\hat{\overline{Y}}^{*}_{\mathrm{Re}g}$ if $\rho < \dfrac{1}{2}\left(\dfrac{C_y}{C_x} - \theta C_y C_x\right)$.

Further, $\hat{\overline{Y}}^{*}_{\mathrm{R}ao}$ is more efficient than the more efficient regression type estimator $\hat{\overline{Y}}^{*}_{G}$, if $\rho + \theta C_y C_x\left(1-\rho^2\right) < \dfrac{1}{2}\dfrac{C_y}{C_x}$.

5. Numerical illustration

Consider the following hypothetical populations to illustrate the performance of different estimators considered above.

Population 1 (Kaur, 1985)

$S_y^2 = 620$, $S_x^2 = 7619$, $S_{xy} = 1453$

$\overline{Y} = 26.30$, $\overline{X} = 117.28$

$\rho = 0.6685$, $n = 100$

Population 2 (Menendez and Reyes, 1998)

$S_y^2 = 8.40$, $S_x^2 = 3.31$, $\rho = 0.20$

$\overline{Y} = 12.0$, $\overline{X} = 16.6$, $N = 5000$, $n = 100$

Table 1. Percent Relative Efficiencies

Estimators	Population – 1	Population - 2
	Relative Efficiency	Relative Efficiency
$\bar{y}$	100	100
$\hat{\bar{Y}}_R$	176.39	97.62
$\hat{\bar{Y}}_R^*$	179.56	97.72
$\hat{\bar{Y}}_{\mathrm{Re}g}$	180.80	104.17
$\hat{\bar{Y}}_{\mathrm{Re}g}^*$	184.05	104.29
$\hat{\bar{Y}}_G^*$	182.60	104.28
$\hat{\bar{Y}}_{Rao}^*$	165.87	518.21

6. Conclusion

1. Although theoretically subject to approximations, $\hat{\bar{Y}}_R^*$ and $\hat{\bar{Y}}_{\mathrm{Re}g}^*$ are more efficient than $\hat{\bar{Y}}_R$ and $\hat{\bar{Y}}_{\mathrm{Re}g}$ respectively, the increase in efficiency for large sample sizes is only marginal.
2. When comparing $\hat{\bar{Y}}_{\mathrm{Re}g}^*$ with $\hat{\bar{Y}}_G^*$, it may be seen that $\hat{\bar{Y}}_{\mathrm{Re}g}^*$ is always more efficient than $\hat{\bar{Y}}_G^*$.
3. Under certain conditions, $\hat{\bar{Y}}_{Rao}^*$ is more efficient than $\hat{\bar{Y}}_{\mathrm{Re}g}$ and $\hat{\bar{Y}}_{\mathrm{Re}g}^*$ and $\hat{\bar{Y}}_G^*$. The empirical studies through hypothetical populations show that $\hat{\bar{Y}}_{Rao}^*$ is inferior to all estimators under consideration for the population–1 and there is large gain in precision in case of $\hat{\bar{Y}}_{Rao}^*$ to other estimators considered for illustration for population-2.

REFERENCES

BEDI, P.K. and HAJELA, D. (1984). An estimator for mean utilizing known coefficient of variation and auxiliary variable, J. Stat. Res., 18, 29–33.

COCHRAN, W.G. (1977). Sampling Techniques, 3rd ed., Wiley.

GUPTA, P.C. (1978). On some quadratic and higher degree ratio and product estimators, J. Ind. Soc. Ag. Stat., 30, 71–80.

KAUR, P. (1985). An efficient regression type estimator in survey sampling, Biom. Journal, 27, 107–110.

MENENDEZ, E. and REYES, A (1998). On an efficient Regression Type estimator, Biom. J., 40, 1, 79–84.

PANDEY, G.S. (1980). Product – Cum – Power estimators, Cal. Stat. Assoc. Bull. 29, 103–108.

PRASAD, B. (1986). Some unbiased estimators Versus mean per unit and ratio estimators in finite population sample surveys, Comm. in Stat. Th. and Meth., 15, 3647–3657.

RAO, I.S. (1978). On a method of using auxiliary information, contributions to statistics, edited by Department of Statistics, Utkal University, Bhubaneswar (India), pp. 65–69.

RAO, T.J. (1993). Auxiliary information in sample surveys: unpublished manuscript.

RAY, S.K. and SAHAI, A. (1980). Efficient families of ratio and product type estimators, Biometrika, 67, 211–215.

RAY, S.K. and SINGH, R.K. (1981). Difference – Cum – ratio type estimators. J. Ind. Stat. Assoc., 19, 147–151.

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

TAILOR, R and SHARMA, B (2009). A modified Ratio-cum-Product Estimator of finite population mean using known coefficient of variation and coefficient of kurtosis, Statistics in Transition-New Series, Vol – 10, PP. 15–24.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 563—584

THE ANALYSIS OF MORTALITY CHANGES IN SELECTED EUROPEAN COUNTRIES IN THE PERIOD 1960–2006

Sabina Denkowska,[1] Monika Papież,[2]

ABSTRACT

Demographic changes which took place in the 20th century clearly reveal progressive ageing of whole societies. This phenomenon influences the risk connected with calculating the products of insurance companies and pension funds, where calculated mortality is one of the most important factors. The paper presents the analysis of mortality changes in male and female populations in selected countries from Central Europe (the Czech Republic, Hungary, Poland and Slovakia) and from Western Europe (France, Italy, Spain and Sweden) in the period 1960–2006. The analysis of the mortality changes has been carried out with the use of variables proposed in 2007 by J. P. Morgan in his work *Life Metrics – A toolkit for measuring and managing longevity and mortality risks.* The data used for the analysis have been obtained from www.mortality.org. The van Broekhoven algorithm has been applied for smoothing crude mortality rates across different ages. The analysis of mortality changes in selected European countries in the period 1960–2006 has shown considerable differences in the changes of initial mortality rate. Apart from obvious differences in male and female mortality, significant differences in the dynamics of mortality between Western and Central European countries were revealed. The most significant differences in the change of graduated initial mortality rate have been observed for people above 40–45 from Central and Western European countries. The period of most striking disproportions in the change of graduated initial mortality rate were the years 1970–1990, which seems to be the result of the socio-economic policy in Central European countries.

Key words*:* mortality, graduated initial rate of mortality, cluster analysis.

[1] Cracow University of Economics, Department of Statistics, 27 Rakowicka St., 31-510 Cracow, Poland. Phone: (+48 12) 293-52-10. Email: sabina.denkowska@uek.krakow.pl.

[2] Cracow University of Economics, Department of Statistics, 27 Rakowicka St., 31-510 Cracow, Poland. Phone: (+48 12) 293-52-10. Email: papiezm@uek.krakow.pl.

1. Introduction

Demographic changes which took place in the 20th century clearly reveal progressive ageing of whole societies. This phenomenon is connected with economic growth, increased life standards resulting from an improved financial situation, advancements in medical sciences, as well as decrease in the number of children in a new family model. Ageing of populations influences the risk connected with calculating the products of insurance companies and pension funds, where calculated mortality is one of the most important factors.

In management of pension funds and insurance companies two risk sources play a fundamental role: the investment (financial) risk and the demographic risk. The demographic risk is divided into two components: the insurance risk and the longevity risk. The longevity risk is particularly important – it derives from the improvements in mortality trends, which determine systematic deviations of the number of deaths from its expected values. This risk should be assuaged by replacing traditional mortality tables used in the assessment of insurance products with projected mortality tables including a forecast of future trends in mortality.

Thus, estimating future changes in mortality trends and accurately identifying the laws governing mortality seems crucial. Mortality can be analysed from two perspectives: a statistical one and a biological one. The statistical approach takes into account historical data and, using them, analyses mortality trends in the past as a means of predicting future trends. The biological approach is based on life standards, advancements in medicine, environment, and the lifestyle of the population.

The paper presents the analysis of mortality changes in Poland and selected European countries carried out on the basis of historical data. The following countries have been chosen for the analysis: from Central Europe: the Czech Republic, Hungary, Poland and Slovakia, from Western Europe: France, Italy, Spain and Sweden. The analysis of the mortality changes has been carried out with the use of the following variables: m_x – *crude central rate of mortality*; q_x – *graduated initial rate of mortality*; e_x – *life expectancy*. These variables were proposed in 2007 by J. P. Morgan in his work *Life Metrics – A toolkit for measuring and managing longevity and mortality risks*, which has since become an international toolkit used for measuring and managing the mortality risk and the longevity risk. These tools measure both risks using standardised methods.

Life Metrics toolkit was applied to analyse mortality changes in male and female populations in selected countries in the period 1960–2006. The data used for the analysis have been obtained from www.mortality.org. The van Broekhoven algorithm1 was applied for smoothing crude mortality rates across different ages.

[1] See: van Broekhoven, H. (2002). Market Value of Liabilities Mortality Risk: A Practical Model. *North American Actuarial Journal* 6(2), 95–106.

2. Life Expectancy and the GDP

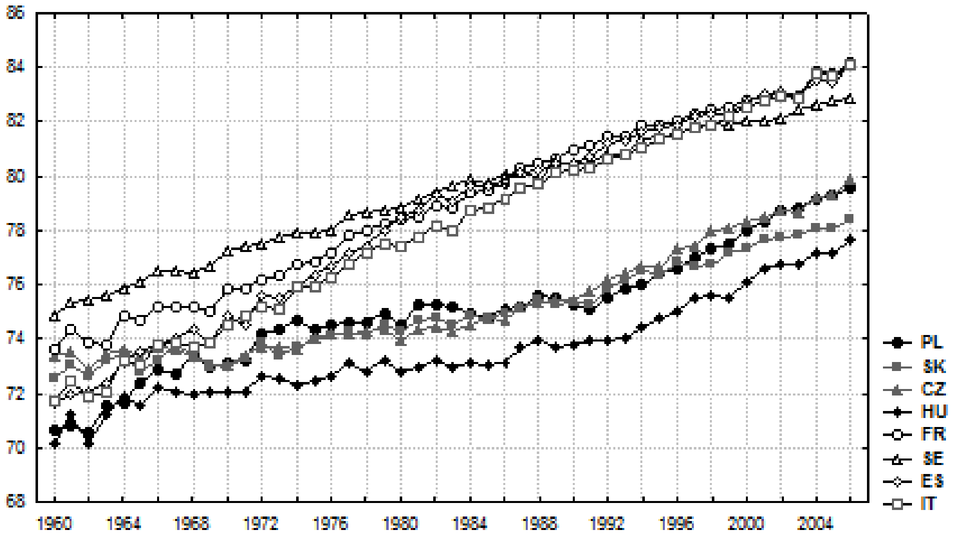
The analysis of life expectancy at the age 0 both in male (Fig. 1) and female (Fig. 2) populations has revealed a significant difference between two groups of countries. Since 1970's life expectancy at the age 0 has been substantially longer for Western European countries: France, Italy, Spain and Sweden than for Central European countries: the Czech Republic, Hungary, Poland and Slovakia. Table 1 presents the dynamics of the life expectancy growth in 2006 relative to 1960 for women and men in selected countries.

Table 1. The dynamics of the life expectancy growth in 2006 relative to 1960 for women and men in selected countries

Country	CZ	HU	PL	SK	FR	ES	IT	SE
	Czech Republic	Hungary	Poland	Slovakia	France	Spain	Italy	Sweden
Female	9%	11%	13%	8%	14%	17%	17%	11%
Male	9%	5%	10%	3%	15%	16%	18%	10%

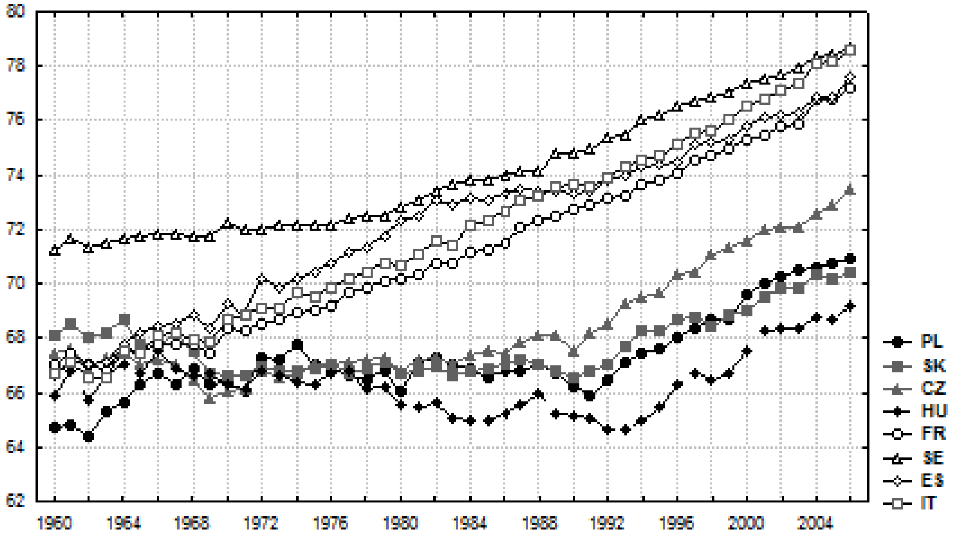
Source: Own calculations.

Figure 1. Life expectancy for men at the age 0 in the period 1960–2006



Source: Data obtained from www.mortality.org.

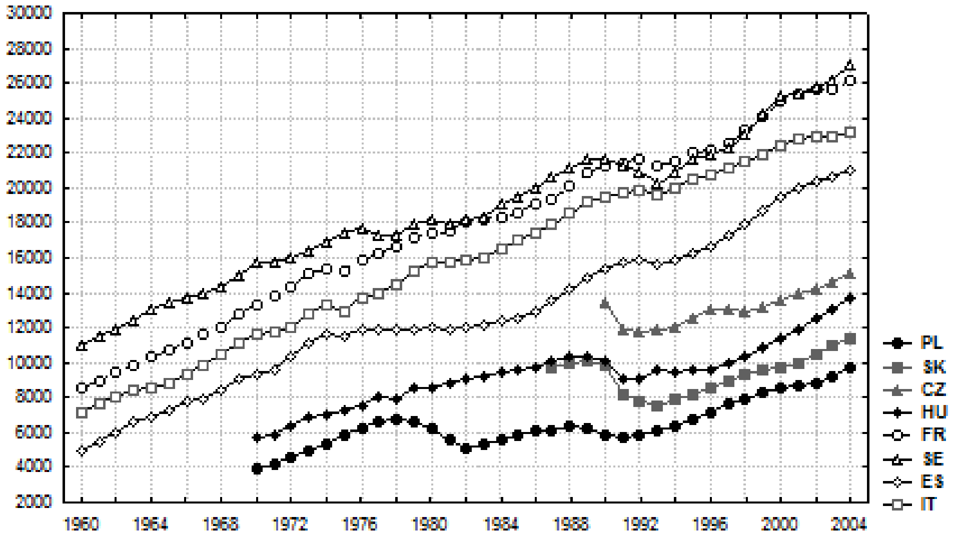
Figure 2. Life expectancy for women at the age 0 in the period 1960–2006



Source: Data obtained from www.mortality.org.

Fig. 3 presents the values of GDP per capita in the analysed countries. The analysis of these values indicates that, similarly to the case of life expectancy, two distinct groups can be observed – one with higher values of GDP (Western European countries) and the other with lower values (Central European countries). This conclusion seems to confirm the statement of a Polish demographer, Edward Rosset, that life tables are "a barometer of social progress". Thus, the differences in life expectancy and the changes in mortality, i.e. lengthening of life expectancy, are a reflection of life conditions which consist of work, eating habits, natural environment, health care and education.

Figure 3. Real GDP per capita (Constant Prices: Laspeyres) (I$ in 2000 **Constant Prices)**



Source: Data obtained from http://pwt.econ.upenn.edu/php_site/pwt62/pwt62_form.php.

The countries selected for the analysis differ in such aspects as geography, climate or culture, yet the preliminary analysis (Fig. 1 and Fig. 2) shows that they do not have a significant impact on the values of life expectancy. It seems (Fig. 3) that the socio-economic development, which happens as a result of long-term socio-economic policy of a given country, has the greatest influence on life expectancy and mortality.

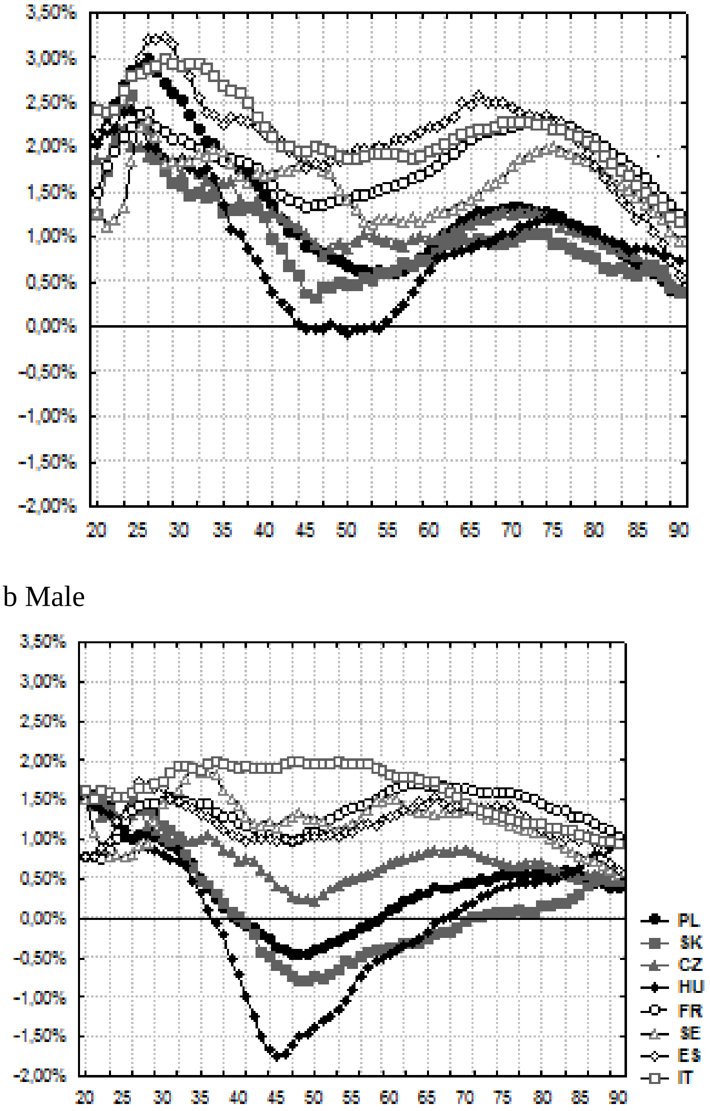
Thus, it might be useful to check from what age the changes in the course and the dynamics of mortality begin – the changes resulting from different political systems and the socio-economic development of particular countries.

3. Average Rate and Volatility of Changes of Graduated Initial Mortality Rate

To analyse mortality first graduated initial rate of mortality (q_x) was calculated for women and men at the age 20–90, and next the average rate of change of graduated initial mortality rate for the period 1960–2005 for women and men at the age 20–90. Figure 4 shows the average rate of the change of graduated initial mortality rate for the period 1960–2005 for women and men at the age 20–90.

Figure 4. Average rate of changes of graduated initial mortality rate in the period 1960–2005 at the age 20–90.

a Female



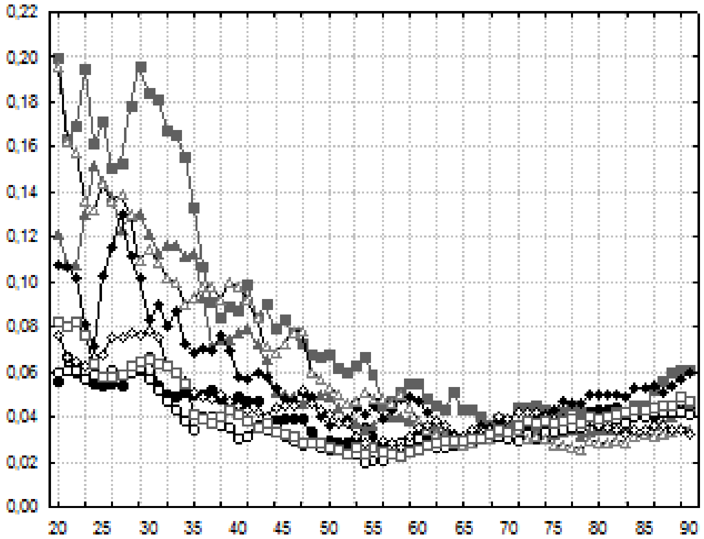
b Male

Source: Own calculations.

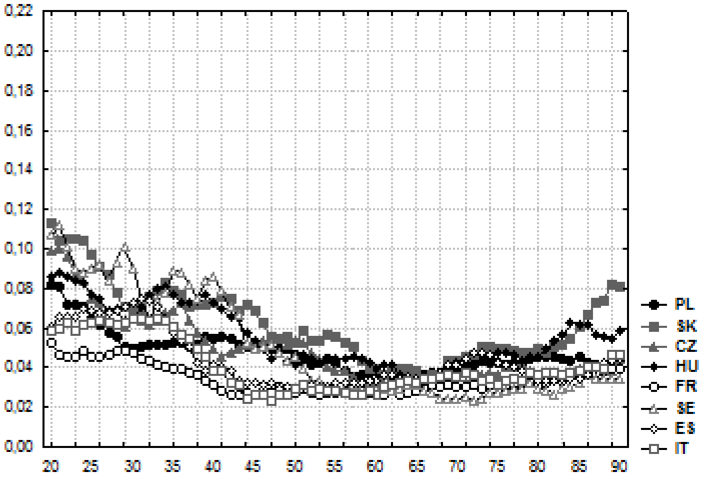
The analysis of Figure 4 reveals a considerably better situation in Western European countries as far as the average changes in the mortality rate for both men and women are concerned. It is especially true for men over 30 and for women over 40. The differences are less conspicuous for very young persons and people reaching 90. For example, in Italy, in the years 1960–2005 mortality rate for men at the age 35–57 was systematically decreasing by almost 2% every year. At the same period in Hungary mortality rate among men aged 45 increased by on average 1,73%. Such increasing mortality rate among middle-aged men was also visible in Slovakia and Poland in the period between 1960–2005.

Figure 5. Volatilities of graduated initial mortality rate changes for selected countries for males and females at the age 20–90.

a Female



b Male



Source: Own calculations.

Figure 5 presents volatility values calculated using standard deviation for selected countries for men and women at the age 20–90 in the period 1960–2005. It is worth paying attention to the curves representing women at the age 20–40, especially the high values of standard deviation for Slovakia, Sweden, the Czech Republic and Hungary.

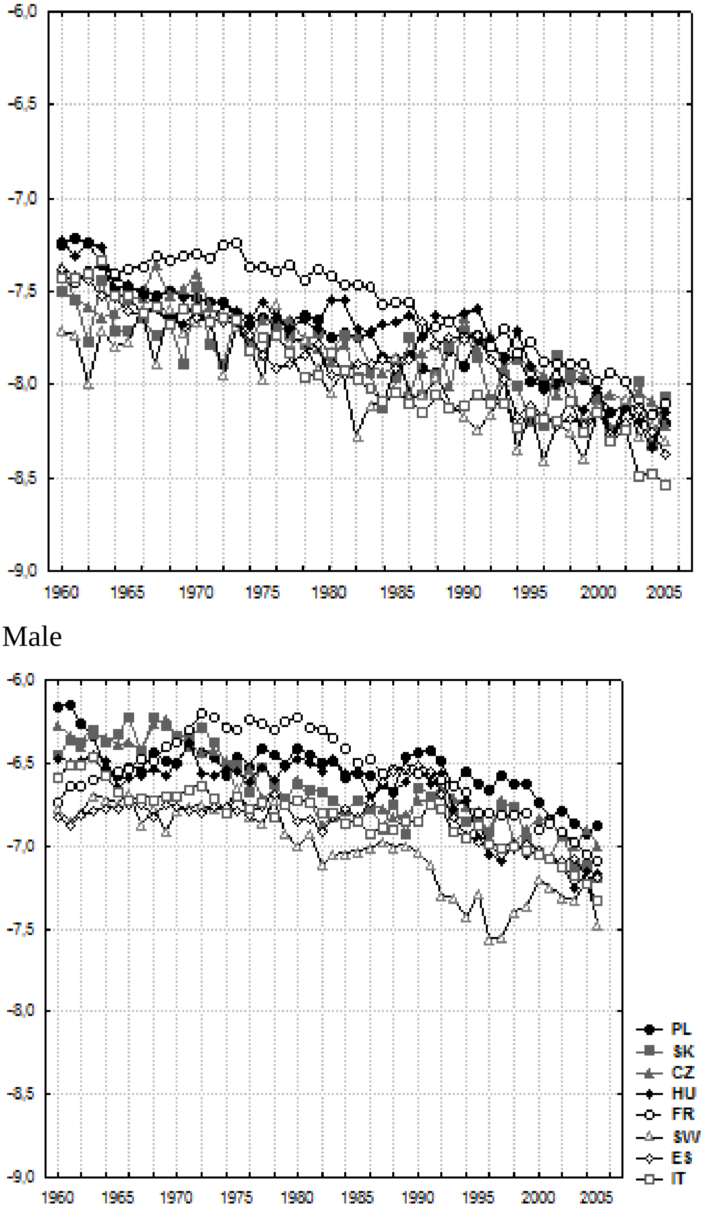
4. Cluster Analysis

Figures 6–9 present the values of q_x on a logarithmic scale at the age of 20, 45, 55 and 65 in the period 1960–2005. They show that both in male and female populations the curves in selected countries of Central and Western Europe differ. To find out at what age a significant difference in q_x can be observed, cluster analysis has been applied. After clustering, significantly "different" observations should be found in different clusters. To cluster countries into sets with "similar" q_x the Euclidean distance and Ward's method have been applied. The Ward procedure is one of the most efficient agglomerative hierarchical clustering methods. It uses an analysis of variance approach to evaluate the distances between clusters. Figures 6b–9b present the d*endrograms* of the clustering of the logarithm q_x at the age of 20, 45, 55 and 65 in the period 1960–2005 which have been done in the same scale. The analyses of particular dendrograms lead to distinct groups of countries. Tables 2–5 present groups of countries at the age of 20, 45, 55 and 65 in the period 1960–2005, which have been obtained for the arbitrarily accepted linkage distance d=3.

The objects of the analysis were the selected countries and the variables were the values of logarithms q_x in the years 1960–2005. So, the observation matrix consists of 8 objects and 45 variables. The calculations were obtained using the STATISTICA program. The results are presented in Figures 6–9.

Figure 6a. The values of q_x on a logarithmic scale at the age of 20 in the period 1960–2005

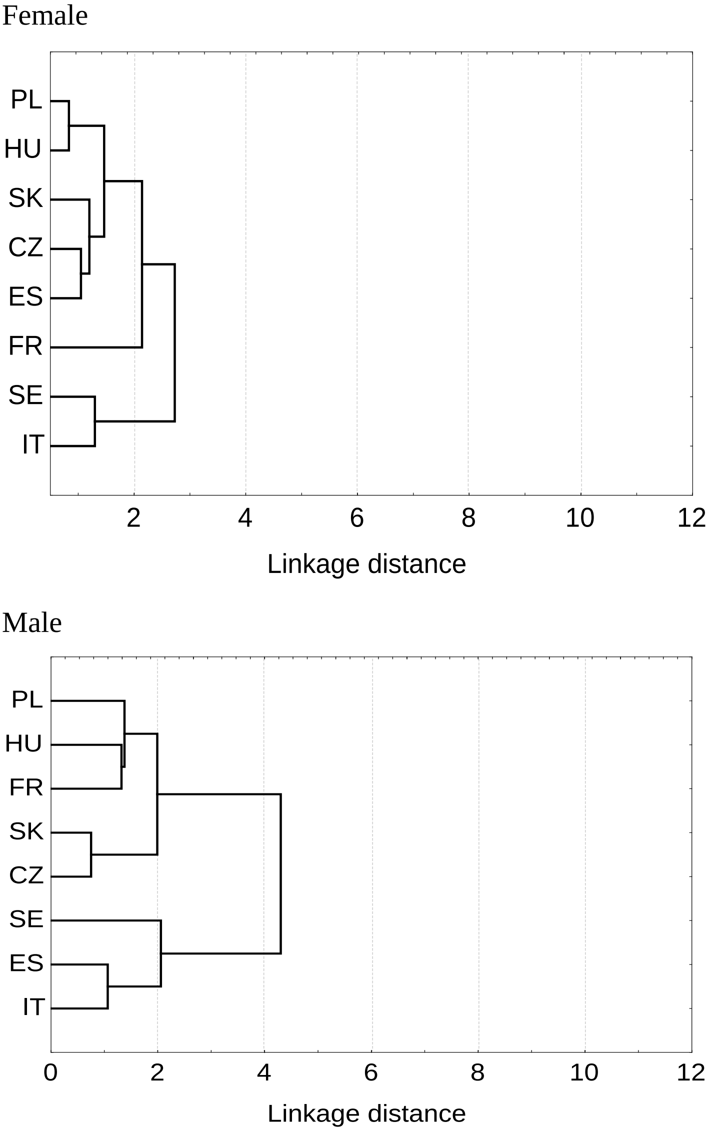
Female



Male

Source: Own calculations.

Figure 6b. The d*endrograms* of the clustering of the logarithm q_x at the age of 20 in the period 1960–2005 obtained using the Euclidean distance and *Ward's method*



Source: Own calculations in STATISTICA 8.0.

Table 2. Groups of countries obtained from d*endrograms* of the clustering of the logarithm q_x at the age of 20 in the period 1960–2005 (see: fig.6b) at the linkage distance d=3

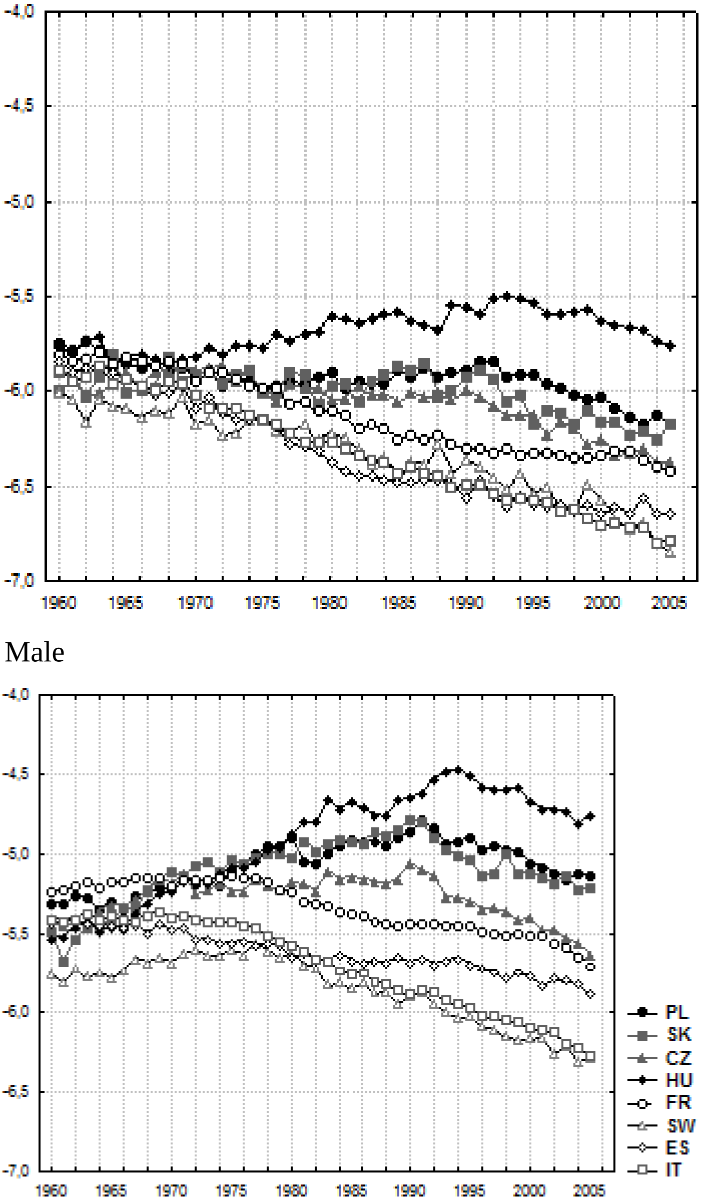
Female		Male	
Gr. 1	PL, HU, SK, CZ, ES, FR, SE, IT	Gr. 1	PL, HU, FR, SK, CZ
		Gr. 2	SE, ES, IT

Source: Own elaboration.

For younger populations (under 40–45) the differences in q_x (measured on a logarithmic scale) for particular countries are not very significant. For older cohorts the linkage distances are increasing, which indicates more significant differences. From 1970's lnq_x for both male and female populations in Western Europe became smaller and the improvement in mortality was faster than in Central European countries. It is most clearly visible from the age of 45 (Fig.7–9).

Figure 7a. The values of q_x on a logarithmic scale at the age of 45 in the period 1960–2005

Female

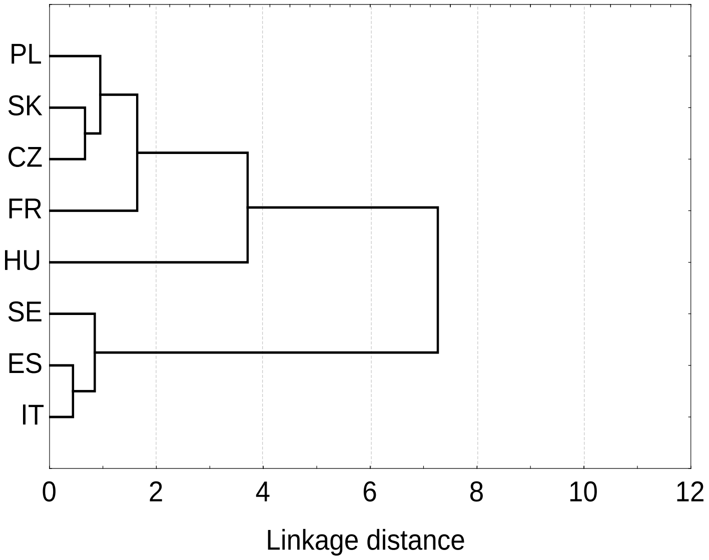


Male

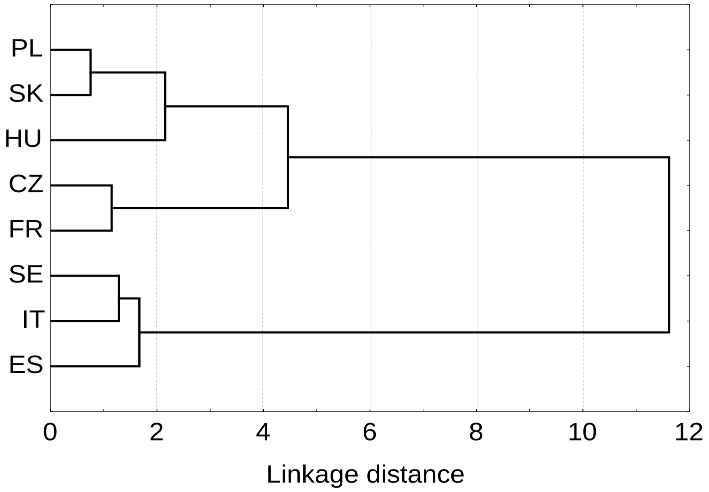
Source: Own calculations.

Figure 7b. The d*endrograms* of the clustering of the logarithm q_x at the age of 45 in the period 1960–2005 obtained using the Euclidean distance and *Ward's method*

Female



Male



Source: Own calculations in STATISTICA 8.0.

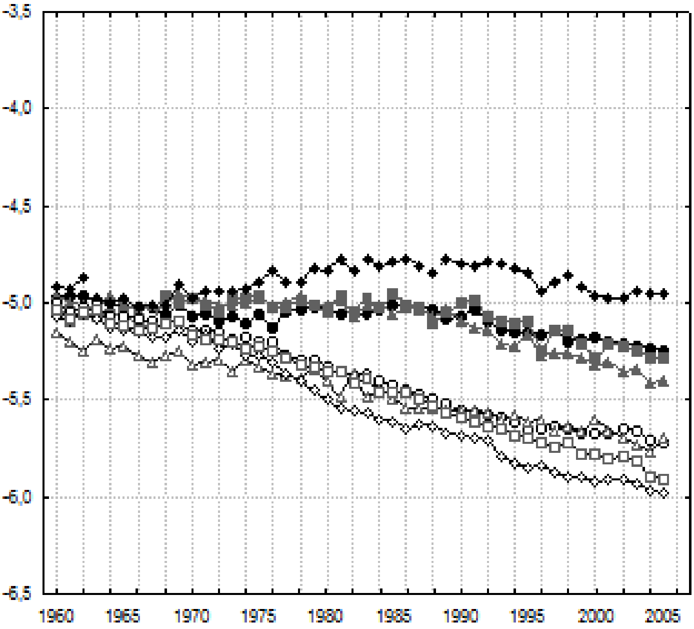
Table 3. Groups of countries obtained from d*endrograms* of the clustering of the logarithm q_x at the age of 45 in the period 1960–2005 (see: fig.7b) at the linkage distance d=3.

Female		Male	
Gr. 1	PL, SK, CZ, FR	Gr. 1	PL, SK, HU
Gr. 2	HU	Gr. 2	CZ, FR
Gr. 3	SE, ES, IT	Gr. 3	SE, IT, ES

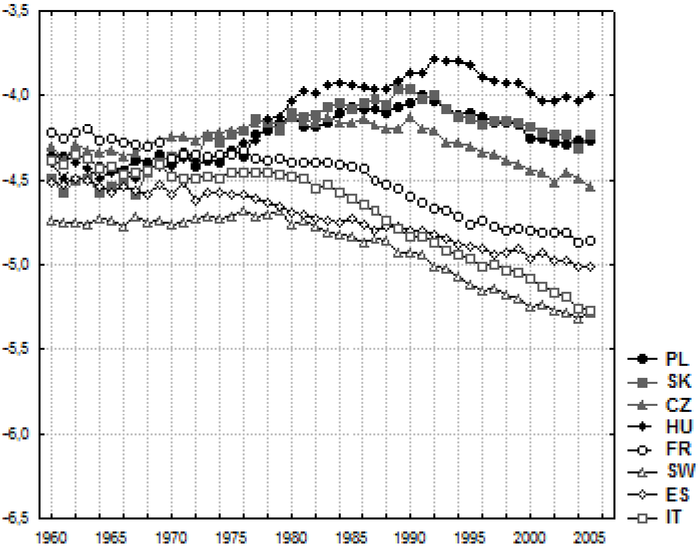
Source: Own elaboration.

Figure 8a. The values of q_x on a logarithmic scale at the age of 55 in the period 1960–2005

Female



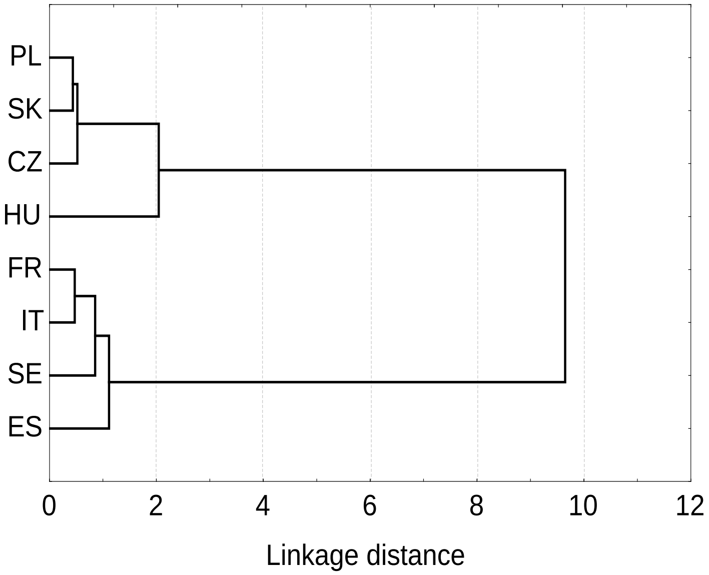
Male



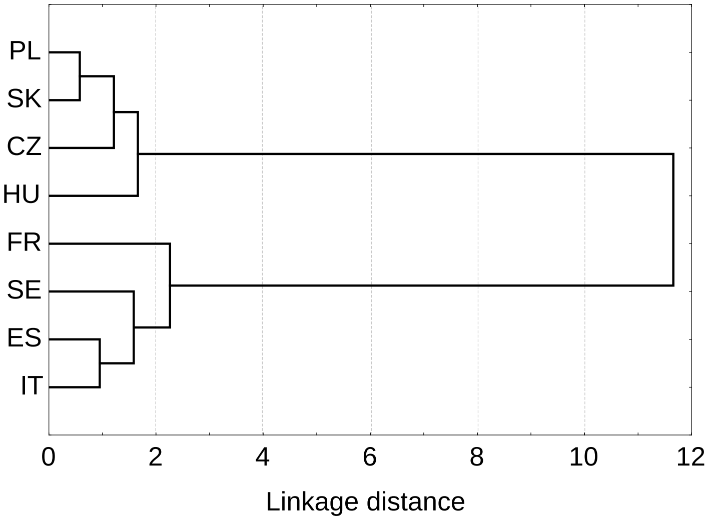
Source: Own calculations.

Figure 8b. The d*endrograms* of the clustering of the logarithm q_x at the age of 55 in the period 1960–2005 obtained using the Euclidean distance and *Ward's method*

Female



Male



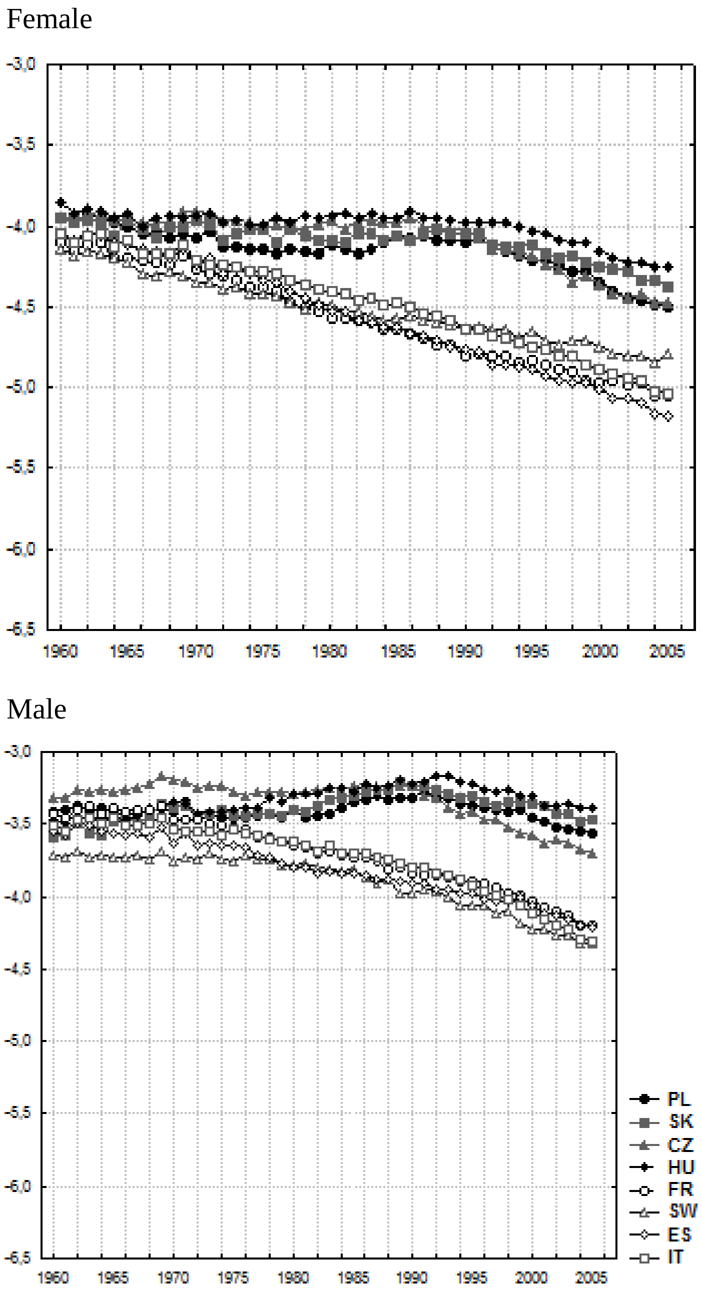
Source: Own calculations in STATISTICA 8.0.

Table 4. Groups of countries obtained from d*endrograms* of the clustering of the logarithm q_x at the age of 55 in the period 1960–2005 (see: fig.8b) at the linkage distance d=3

Female		Male	
Gr. 1	PL, SK, CZ, HU	Gr. 1	PL, SK, CZ, HU
Gr. 2	FR, IT, SE, ES	Gr. 2	FR, IT, SE, ES

Source: Own elaboration.

Figure 9a. The values of q_x on a logarithmic scale at the age of 65 in the period 1960–2005



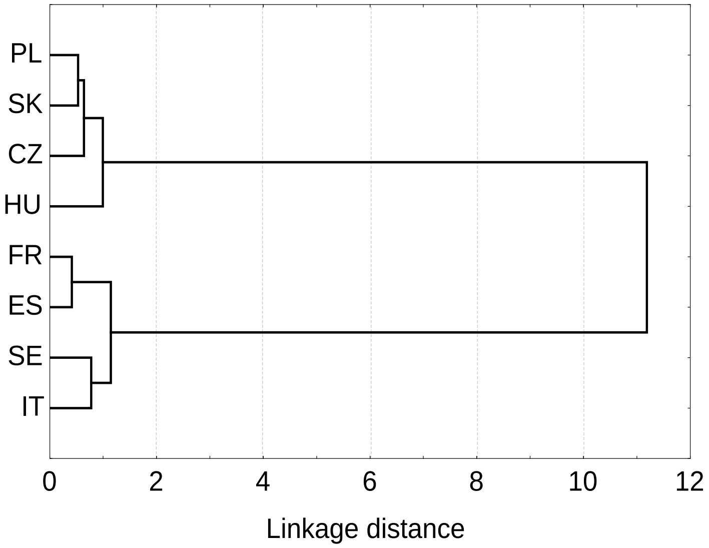
Female

Male

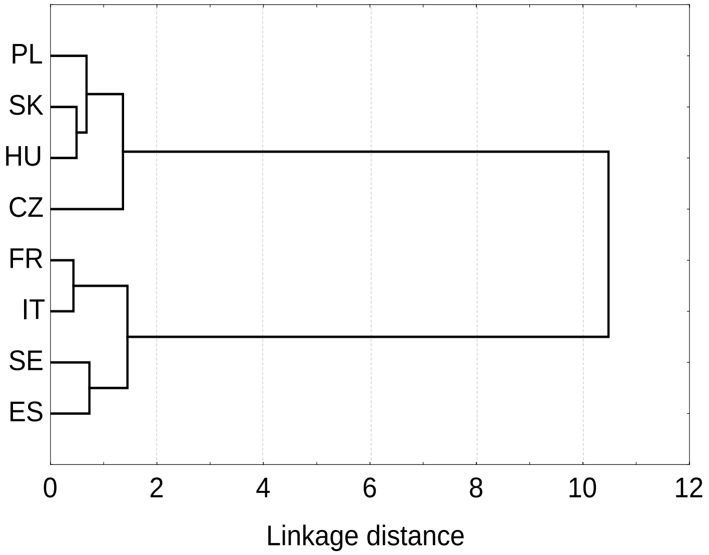
Source: Own calculations.

Figure 9b. The d*endrograms* of the clustering of the logarithm q_x at the age of 65 in the period 1960–2005 obtained using the Euclidean distance and *Ward's method.*

Female



Male



Source: Own calculations in STATISTICA 8.0.

Table 5. Groups of countries obtained from d*endrograms* of the clustering of the logarithm q_x at the age of 65 in the period 1960–2005 (see: fig.9b) at the linkage distance d=3

Female		Male	
Gr. 1	PL, SK, CZ, HU	Gr. 1	PL, SK, CZ, HU
Gr. 2	FR, IT, SE, ES	Gr. 2	FR, IT, SE, ES

Source: Own elaboration.

5. Conclusions

The analysis of mortality changes in selected European countries in the period 1960–2007 has shown considerable differences in the changes of initial mortality rate. Four distinct groups have been analysed: female populations in selected Western European countries, female populations in selected Central European countries, male populations in selected Western European countries and male populations in selected Central European countries. Apart from obvious differences in male and female mortality, significant differences in the dynamics of mortality between Western and Central European countries have been revealed. In spite of their geographic, climatic and cultural differences, the countries constituting both groups seemed quite homogenous.

The most significant differences in the change of graduated initial mortality rate have been observed for people above 40–45 from Central and Western European countries. The period of most striking disproportions in the change of graduated initial mortality rate was the years 1970–1990, which seems to be the result of the socio-economic policy in Central European countries. The transformations which took place in 1990's helped to reduce the differences in the dynamics (decrease) of mortality, yet the differences in life expectancy between the Western and Central European countries will probably still be visible in the next years (i.e. life expectancy in the Western countries will continue to be longer).

Another important phenomenon is connected with the fact that, irrespective of geographic and political factors, in the 1990's the process of aging of the whole populations began. In the future ageing can constitute a serious obstacle in the socio-economic development. Firstly, labour resources will grow old, i.e. there will be fewer and fewer young workers on the job market which will become dominated by less mobile older workers. Secondly, the number of people in the productive age will decrease, which means that maintaining a fast rate of economic growth will be possible only if productivity increases and a bigger percentage of people in their productive age continue working. One of the ways of assuaging the problem of ageing is lengthening the productive periods, also called lifting the actual retirement age, which will slow down the process of diminishing the number of professionally active workers. Such policy is also in accordance with demographic tendencies regarding lengthening life span. Lengthening the productive periods means resigning from a contemporary trend of early retirement and shortening the duration of the period of a 'well-deserved rest'. As a result, changes in the area of employment and social security are needed.

REFERENCES

ALHO, J. M., SPENCER, B. D., (2005), *Statistical Demography and Forecasting*. Springer Series in Statistics. New York: Springer.

ANDREEV, K. F., VAUPEL, J. W., (2005), Patterns of Mortality Improvement over Age and Time: Estimation, Presentation and Implications, http://paa2005.princeton.edu/

BENJAMIN, B., POLLARD J. H., (1993), *The Analysis of Mortality and Other Actuarial Statistics*. London: Institute and Faculty of Actuaries.

BONGAARTS, J. (2005), Long-range trends in adult mortality: Models and projection methods. *Demography*, 42 (1), pp. 23–49.

BOOTH, H., (2006), Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, 22 (3), pp. 547–581.

FRĄCKIEWICZ, J. (ED.)., (2002), *Poland versus Europe. Demographic processes at the beginning of 21st century. The process of ageing in Poland and its social consequences*, Papers from The First Demographic Congress in Poland 2001–2002, Polskie Towarzystwo Polityki Społecznej, Katowice.

GRABIŃSKI, T., WYDYMUS, S., ZELIAŚ, A. (ED.), (1989), *Methods on numeric taxonomy in modelling socio-economic phenomena*, PWN, Warszawa.

HOUGAARD, P., (2000), *Analysis of Multivariate Survival Data*, Springer, New York.

Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany): www.mortality.org.

JÓŹWIAK, J., (2004), *Methodological introduction to modern demographic analysis [in:] Poland's demographic problems before joining the European Union*, ed. Strzelecki Z., PWE, Warszawa.

KOWALESKI, J., (2003), *Methodological issues in research on the process of ageing* [in:] Papers from The First Demographic Congress in Poland 2001–2002, *Demographic processes at the beginning of 21st century. Poland’s* *versus Europe* (ed. Strzelecki Z.), Rządowa Rada Ludnościowa Volume XVII, Warszawa.

Life Metrics: *A toolkit for measuring and managing longevity and mortality risks Technical Document*, Pension Advisory Group, JPMorgan Chase Bank, N.A., 2007,

http://www.jpmorgan.com/pages/jpmorgan/investbk/solutions/lifemetrics/library.

KURKIEWICZ, J., (1992), *Basic methods of demographic analysis*, PWN Warszawa.

ROSSET, E., (1959), The process of ageing. *Demographic Studies*. Warszawa.

TABEAU, E., VAN DEN BERG JETHS, A., HEATHCOTE, C. (Eds.), (2001), *Forecasting mortality in developed countries*. Insights from a Statistical, Demographic and Epidemiological Perspective, Kluwer Academic Publishers.

VAN BROEKHOVEN, H., (2002), Market Value of Liabilities Mortality Risk: A Practical Model. *North American Actuarial Journal* 6(2), pp. 95–106.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 585—596

DISTRIBUTION APPROXIMATIONS FOR CUSUM AND CUSUMSQ STATISTICS

Reza Habibi[1]

ABSTRACT

The cumulative sum (cusum) is an important statistics in testing for a change point. This paper is concerned with the distribution approximations to the cusum statistic under the null and alternative hypotheses. We also consider distribution approximations for the cumulative sum of squares (cusumsq) test statistics. Finally, a discussion section is given.

Key words: Beta approximated; Change point; Cumulative sum; Cumulative sum of squares; Multivariate normal; Response surface regression.

1. Introduction

It is very important for economic policy to identify change points in economic and financial series. For example, Hsu (1979) tested the existence of changes in stock market data. Kim *et al*. (2000) considered the problem of multiple change point in GARCH models. Hillebrand and Schnabl (2003) studied change point detection in volatility of Japanese foreign exchange intervention under GARCH modeling. Halunga *et al*. (2009) detected changes in the order of integration of US and UK inflation. In finance, the portfolio's volatility may increase as well as risk premium rises (see e.g. Mikosch and Starica, 2003).

During the last four decades, different methods are employed for detecting change points. Page (1954) studied change point analysis in the context of quality control. Chernoff and Zacks (1964), using a quasi-Baysian approach, modeled the change points. Hinkely (1970) derived the maximum likelihood estimation of change point. Worsley (1988) constructed confidence intervals for change point in the exponential family distributions. Habibi *et al*. (2005) considered the change point detection in a general class of distributions. An excellent reference in change point problems is Csorgo and Horvath (1997). The distribution theory used for these tests is typically asymptotic. So deriving the exact distribution of

[1] Department of Statistics, Central Bank of Iran, Ferdowsi Ave., Postal Code: 1135931496, Tehran, Iran, e-mail: habibi1356@gmail.com.

test statistics is very important. One of the most useful approach to detect shift in means of observations is cumulative sum (cusum) which is described as follows. In this paper, we study the finite distributions of cusum test statistic. We also consider the shift in variance case. The cusum statistic given by

$$M_n = n^{-1/2} \max_{1 \le k \le m} |s_k|,$$

is an important statistic in testing for a change point, at which

$$s_k = \sum_{i=1}^{k} e_i,$$

$e_i = x_i - \overline{x}, \overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i, i, k = 1, ..., m, m = n-1$. Here, $x_1, ..., x_n$ is a sequence of independent normal random variables whose means are θ_i, $i = 1, ..., n$, where

$$\theta_i = \begin{cases} \theta & i = 1, 2, ..., k_0 \\ \theta + \delta & i = k_0 + 1, ..., n, \end{cases}$$

with a common known variance σ^2. It is interesting to test if θ_i are changed at unknown time point k_0, that is: $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$. The large values of M_n rejects H_0. The exact null distribution of M_n is complicated. Note that the null distribution of M_n does not depend on θ. The limiting null distribution of M_n is given by $\sigma Sup\, B(t)$ where $B(t)$ is the standard Brownian bridge on $(0,1)$ and the supremum is taken over $(0,1)$. Conniffe and Spencer (2000) proposed a central chi-squared approximation for the null distribution of M_+^2, where

$$M_+ = n^{-1/2} \max_{1 \le k \le m} s_k.$$

Their approximation method works well. However, in this note, the exact null and alternative distribution of M_n is studied. We also consider the exact distribution of change point estimator. The change point in variance is considered. These problems are not considered by Conniffe and Spencer (2000). The problem is to find the quantile c_α such that $P_{H_0}(M_n > c_\alpha) = \alpha$ for finite sample sizes *n*. We note that

$$P_{H_0}(M_n > c_\alpha) = 1 - P_{H_0}(|s_k| \le \sqrt{n} c_\alpha, \textit{ for all } k = 1, ..., m) = \alpha,$$

and

$$P_H(M_n < x) = 1 - P_H(|s_k| \le \sqrt{n}c_\alpha, \textit{ for all } k = 1,...,m), H = H_0 \textit{ and } H_1.$$

As follows, we show that $s = (s_1,...,s_m)^T$ has multivariate normal distribution under the null and alternative hypotheses. That is, M_n is the maximum of absolute of a multivariate normal distribution and that $\sqrt{n}c_\alpha$ is a two sided equi-coordinate $1-\alpha$ percent quantile of multivariate normal distribution. The percentage points of M_+ is denoted by c_α^+. Then $\sqrt{n}c_\alpha^+$ is the one sided equi-coordinate quantile of multivariate normal. Genz (1992) proposed some numerical approaches to calculate the cumulative probabilities and equi-coordinate quantiles of a multivariate normal distribution with any mean vector and covariance matrix. The function *pmvnorm* in *mvtnorm* package of *R* software performs this calculations.

To prove the normality, let $x = (x_1,...,x_n)^T$ be observation vector. Under the null hypothesis of no change point, x has a *n*–variates normal with mean vector θj_n and covariance matrix $\sigma^2 I_n$ where I_n is the $n \times n$ identity matrix and j_n is $n \times 1$ vector of 1's. The vector of deviation from the mean $e = (e_1,...,e_m)^T$ equals to Ax where A is the following partitioned matrix

$$A = \left[I_m - \frac{1}{n} J_m \vdots - \frac{1}{n} j_m \right],$$

where J_m is $m \times m$ matrix of 1's. Then has a *m*–variate normal distribution with mean vector 0 and the covariance matrix

$$\sigma^2 AA^T = \sigma^2 (I_m - \frac{1}{n} J_m).$$

Let $L = (L_1,...,L_m)^T$ is the m × m matrix of vectors L_i such that $L_i = (1,...,1,0,...,0)^T$ at which the number of 1's at L_i is $i = 1,...,m$. One can see that $s = Le$, and then s is m–variate normal with mean vector 0 and the covariance matrix $\sigma^2 D$, where

$$D_{ij} = \min(i, j) - \frac{ij}{n}, i, j = 1,...,0.$$

Under the alternative hypothesis, s is again *m*–variate normal with the covariance matrix $\sigma^2 D$ but the mean vector is $\alpha = (\alpha_1,...,\alpha_m)^T$, such that

$$\alpha_k = \frac{-\delta}{n}\min(k,k_0)\{n-\max(k,k_0)\}, k=1,\dots,m.$$

In the next section, we compute one-sided and two-sided cv's $\sqrt{n}c_\alpha^+$ and $\sqrt{n}c_\alpha$ We compare our critical points (cv's) with the cv's obtained by Monte Carlo (MC) simulation. We study the power of test. We also consider distribution approximations for the cumulative sum of squares (cusumsq) test statistic for change point detection in variance in section 3.

2. Cv's and power of test

Tables 1 and 2 give one-sided and two-sided cv's. Without loss of generality, we assume that $\sigma^2 = 1$ and $\theta = 1$. It is seen our approximated cv's are close to true cv's estimated by Monte Carlo study. This fact shows our approximation is accurate. The absolute errors of our approximation is measured by

$$e_\alpha = \left| P_{H_0}(\max_{1\le k\le m}|s_k| \le q_\alpha) - \alpha \right|,$$

where q_α is the true quantile of $\max_{1\le k\le m}|s_k|$ obtained by Monte Carlo experiment and $P_{H_0}(\max_{1\le k\le m}|s_k| \le x)$ is computed by our normal approximation. Table 3 gives the maximum and median of 100 errors e_α, $\alpha = 0.9(0.001)0.999$ for each sample sizes. Table 3, we conclude that our approximation works well. Table 4 gives the normal approximated power of test for some selected sample sizes *n*. We let $k_0 = \frac{n}{2}, \delta = 1, 2, \sigma^2 = 1$ and $\theta = 1$.

Table 1. Comparison of normal and MC two-sided cv's

α	0.1	0.05	0.025	0.01
Normal cv, n = 5	2.166470	2.466221	2.741670	3.074471
MC cv, n = 5	2.161882	2.467639	2.739131	3.074245
Normal cv, n = 6	2.422699	2.753462	3.057174	3.418402
MC cv, n = 6	2.421451	2.753106	3.057837	3.415915
Normal cv, n = 10	3.291676	3.717723	4.107556	4.569724
MC cv, n = 10	3.293115	3.710154	4.110897	4.57586
Normal cv, n = 15	4.159806	4.680369	5.159716	5.729828
MC cv, n = 15	4.160063	4.686651	5.148775	5.726891
Normal cv, n = 20	4.891475	5.494822	6.042609	6.70567
MC cv, n = 20	4.915969	5.505689	6.045103	6.689938

Table 2. Comparison of normal and MC one-sided cv's

α	0.1	0.05	0.025	0.01
Normal cv, n = 5	1.837014	2.173590	2.456540	2.826679
MC cv, n = 5	1.837081	2.172501	2.456957	2.825337
Normal cv, n = 6	2.062609	2.423182	2.753357	3.144838
MC cv, n = 6	2.060721	2.423146	2.757961	3.150792
Normal cv, n = 10	2.812123	3.291219	3.718938	4.225157
MC cv, n = 10	2.813812	3.292955	3.729271	4.237521
Normal cv, n = 15	3.563842	4.154548	4.670302	5.298982
MC cv, n = 15	3.565857	4.154409	4.659146	5.298102
Normal cv, n = 20	4.228066	4.892632	5.493382	6.229083
MC cv, n = 20	4.233962	4.903461	5.496865	6.233975

Table 3. Errors of normal approximated probabilities

n	Maximum error	Median error
5	0.0017	0.0005
6	0.0009	0.0007
10	0.0012	0.0008
15	0.0024	0.0004
20	0.0011	0.0002

Table 4. Normal approximated power, α =0.05

n/θ	1	2
5	0.16263	0.51562
6	0.19832	0.62041
10	0.29361	0.83136
15	0.40851	0.94663
20	0.52157	0.98595

Remark 1

Ploberger and Kramer (1992) used the cusum statistic replacing $e_i = x_i - \overline{x}$, by regression residuals. Our results can be extended to regression residuals. Consider the regression model

$$Y = X\beta + \varepsilon \text{ ,}$$

at which $Y_{n\times 1}$, $X_{n\times p}$, $\beta_{p\times 1}$ and $\varepsilon_{n\times 1}$ are observation vector, design matrix, unknown coefficients vector and residual vector, respectively. Suppose that X is of full rank and define $Q = I_n - P$ where $P = X(X^T X)^{-1} X^T$

Then the vector of estimated residuals $\in = (\in_1, ..., \in_m)^T$ is given by $\in = Q\varepsilon$ and $\sum_{i=1}^{n} \varepsilon_i = 0$.

Let $s_j^{\in} = \sum_{i=1}^{n} \varepsilon_i$ for $i = 1, ..., m$ .Then $s_j^{\in} = L_i^T \in$. Then

$$\operatorname{cov}(s_i^{\in}, s_j^{\in}) = \sigma_{ij}^{\in} = L_i^T Q L_j = L_i^T (I_n - P) L_j = \min(i, j) - L_i^T Q L_j .$$

Following section 2, it can be shown that the null distribution of $s_j^{\in}, j = 1, ..., m$ are multivariate normal with zero mean and covariance matrix $\sum \in = (\sigma_{ij}^{\in})$. The similar results can be extended to alternative distributions. The above results relates to known variance case. In most practical situations, variance will have to be estimated. This case is considered in section 4.

3. Change in Variance

In the previous section, we studied the change point in the means of observation. To detect change point in variance using the cusum statistic, let $y_i = x_i^2$. Here, under H_1, $x_1, ..., x_n$ is a sequence of independent normal observations such that $E(x_i) = 0$ and

$$\operatorname{var}(x_i) = \begin{cases} \sigma^2 & i = 1, 2, ..., k_0 \\ \xi^2 & i = k_0 + 1, ..., n, \end{cases}$$

where $\sigma^2 \neq \xi^2$. The means of y_i are changed under H_1 . The cusum test statistic is

$$M_3 = \frac{n^{-1/2} \max_{1\le k\le m} \left| \sum_{i=1}^{k} (y_i - \bar{y}) \right|}{\sigma^2}.$$

The limiting null distribution of M_3 is given by $\sqrt{2}\sup|B(t)|$. The Table 5 gives 5% quantile of M_3 . A response surface regression for 5% quantile of M3 is estimated as follows. The adjusted R^2 of regression is 99.1.

$$q_n = 2.76 - 2.59 n^{-1} .$$

Table 5. 5% quantile of M_3

n	q_n	n	q_n	n	q_n	n	q_n
5	2.228	10	2.514	35	2.693	55	2.706
6	2.364	15	2.596	40	2.697	60	2.715
7	2.399	20	2.606	45	2.705	65	2.723
9	2.446	25	2.671	50	2.705	70	2.737

Following Conniffe and Spencer (2000), we propose a chi-squared approximation in the form of $a_n \chi^2_{df_n}$ $(a_n > 0\, and\, df_n > 0)$ for M^2_{3+}. The moment estimates of a_n and df_n are

$$a_n = \frac{\lambda_n^2}{2\mu_n} and\ df_n = \frac{2\mu_n^2}{\lambda_n^2}$$

where μ_n and λ_n^2 are the mean and variance of M^2_{3+} under H_0. Table 6 gives μ_n , λ_n^2, a_n and df_n for $\sigma^2 = 1$. Following Conniffe and Spencer (2000), we let n = 10, 20, 30, 40, 60, 80, 100. Table 7 gives the median and maximum of absolute errors.

Table 6. Values of μ_n, λ_n^2, a_n and df_n

n	10	20	30	40	60	80	100
μ_n	6.01634	13.7955	21.8967	30.5796	48.3933	65.6576	84.3329
λ_n^2	137.622	458.696	960.258	1711.04	3663.53	6216.98	9710.17
a_n	11.4374	16.6248	21.9269	27.9768	37.8516	47.3439	57.5704
df_n	0.52602	0.82981	0.99862	1.09303	1.27851	1.38682	1.46486

Table 7. Comparisons chi-squared with MonteCarlo probabilities

n	Maximum of error	Median of errors
10	0.0015	0.0005
20	0.0015	0.0003
30	0.0014	0.0004
40	0.0012	0.0006
60	0.0016	0.0004
80	0.0014	0.0005
100	0.0015	0.0005

Remark 2

When σ^2 is unknown it is replaced by its estimate under the null hypothesis, i.e., $\overline{y}$ and the test is proposed by

$$T_n = n^{\frac{1}{2}} \max_{1 \le k \le m} \left|D_k\right|, D_k = \frac{\sum_{i=1}^{k} X_i^2}{\sum_{i=1}^{n} X_i^2} - \frac{k}{n}, k = 1,\dots,n.$$

This statistic is the cumulative sum of square (cumsumsq) proposed by Inclan and Tiao (1994). The limiting null distribution of Tn is given by $\sqrt{2}\sup\left|B(t)\right|$. For moderate sample sizes, Sanso *et al.* (2004) estimated a response surface regression for 5% quantile of T_n as follows

$$q_{0.05} = 1.359167 - 0.737020 n^{-1/2} - 0.69155 n^{-1}.$$

Remark 3

When $E(x_i) = \mu$ (known), we let $y_i = (x_i - \mu)^2$. When μ is unknown, we can use M_3 again, since means of x_i^2 are changed. We can also use M_3 with letting $y_i = (x_i - \overline{x})^2$. Call this statistic by $\hat{M}_3$. The critical values (cv) are given in Table 8 for $\alpha = 0.05$. Parameter μ is chosen by computer. As we expect, the critical values of $\hat{M}_3$ does not depend on μ. In Table 9, we compare the power of tests for M_3 and $\hat{M}_3$ cases, $\sigma^2 = 1$, $\xi^2 = 3$ and $k_0 = \frac{n}{2}$. It seems test procedure, based on $\hat{M}_3$ works much better. We are working on null and alternative distributions of $\hat{M}_3$. For other possibilities, see discussion section.

Table 8. Critical values

n	10	20	30	40	60	80	100
μ	0.791	1.375	0.163	1.101	0.785	1.111	0.723
cv of M_3	2.770	3.935	1.918	3.364	2.844	3.457	2.664
cv of $\hat{M}_3$	1.739	1.797	1.808	1.830	1.874	1.877	1.866

Table 9. Power of tests M_3 and $\hat{M}_3$

n	Power of M_3	Power of $\hat{M}_3$
10	0.344	0.417
20	0.404	0.646
30	0.795	0.788
40	0.622	0.866
60	0.810	0.950
80	0.807	0.980
100	0.956	0.997

4. Discussion

Following referee comments, this section is added to paper to present a list of future research topics. We are working on them and they will be completed in future.

4.1. Change in mean: unknown variance

Conniffe and Spencer also developed tests for the unknown variance case. In most practical situations, variance will have to be estimated and unless sample size is very large the estimate cannot safely be treated as a known value. This is why, we considered this part. I guess, I can approximate the distribution of test statistic by multivariate t distribution.

4.2. Change in mean and variance

Sometimes when variance changes so does the mean. For example, finance theory says that if a portfolio’s volatility increases the risk premium will rise and change the return. The cusum test statistics and their distributions are interesting.

4.3. Test procedures based on regression residuals

Similar to Remark 1, it is interesting to detect change point in variance using regression residuals. We are working on this topic.

4.4.Distribution of $\hat{M}_3$

The null and alternative distribution of $\hat{M}_3$ is interesting.

Acknowledgment

I thank the referee for his valuable comments on my manuscript. These comments helped me to improve the paper.

REFERENCES

CHERNOFF, H. and ZACKS, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *Ann. Math. Statist.* 35, 999–1018.

CONNIFFE, D. and SPENCER, J. E. (2000). Approximating the distribution of the maximum partial sum of normal deviates. *Journal of Statistical Planning and Inference*, 88, 19–27.

CSORGO, M., and HORVATH, L., (1997). *Limit Theorems in Change Point Analysis,* Wiley. UK.

GENZ, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1, 141–150

HABIBI, R., SADOOGHI-ALVANDI, S. M. and NEMATOLLAHI, A. R. (2005). Change point detection in a general class of distributions. *Communications in Statistics, Theory and Methods* 34, 1935–1938.

HALUNGA, A. G., OSBORN, D. R. and M. SENSIER (2009). Changes in the order of integration of US and UK infation. *Economics Letters* 102, 30–32.

HILLEBRAND, E. and SCHNABL, G. (2003). The effects of Japanese foreign exchange intervention: GARCH estimation and change point detection. *Discussion Paper* No.6. Japan Bank for International Cooperation (JBIC).

HINKLEY, D., (1970). Inference about the change-point in a sequence of random variables. *Biometrika* 57, 1–17.

HSU, D.A., (1979). Detecting shifts of parameter in gamma sequences, with applications to stock price and air traffic flow analysis. *J. Amer. Statist. Assoc.* 74, 31–40.

INCLAN, C. and TIAO, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variances. *J. Amer. Statist. Assoc*. 89, 913–923.

KIM S., S. CHO and S. LEE (2000), On the cusum test for parameter changes in GARCH(1,1) models, *Communications in Statistics, Theory and Methods* 29, 445–462.

MIKOSCH, T. and STARICA, C. (2003). Change of structure in financial time series, long range dependence and the GARCH modeld. *Review of Economics and Statistics*, forthcoming.

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* 41, 100–115.

PLOBERGER, W. and KRAMER, W. (1992). The cusum test with ols residuals. *Econometrica* 60, 271–285.

SANSO, A., ARAGO, V. and CARRION, J. L. (2004). Testing for changes in the unconditional variance of financial time series, *University of Barcelona Working Paper.*

WORSLEY, K. J. (1986). Confidence regions and test for a change point in a sequence of exponential family random variables. *Biometrika* 73, 91–104.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 597—614

ON THE LINK BETWEEN SILBER AND DAGUM DECOMPOSITIONS OF THE GINI INDEX

Mauro Mussini[1]

ABSTRACT

In this paper, we combine two alternative procedures for decomposing the Gini index: the Silber matrix approach and the Dagum technique based on Gini mean difference. This combined decomposition can be used for overlapping as well as no overlapping population subgroups. The presented decomposition procedure derives the within-group, the between-group, and the overlapping inequalities by exploiting all of the information on income distribution contained in one matrix accounting for pair wise disparities between per capita income shares. Moreover, this matrix approach can serve the scope of counting the frequency of overlapping, providing new insights for a more complete analysis of overlapping. The suggested methodology is applied to survey data on Italian employee incomes in 2000 and 2008.

Key words: Inequality measure; Gini decomposition; overlapping.

1. Introduction

The literature offers various approaches for calculating the Gini index and decomposing it by subgroup (Bhattacharya and Mahalanobis, 1967; Giorgi, 1999; Shorrocks, 1984; Pyatt, 1976). Among those procedures using matrix algebra, there is Silber (1989) method. Otherwise, Dagum (1997) decomposed the Gini index starting from its expression in terms of relative mean difference. This paper combines Silber and Dagum approaches in a matrix form useful for the subgroup decomposition of the Gini index.

Introducing an algebraic linear operator, called *G*-matrix, Berrebi and Silber (1987) proposed a matrix compact form for computing the Gini index. Silber (1989) derived a decomposition of this matrix expression either by subgroup or by income source. Considering per capita income shares, Silber technique captures within and between group inequalities, the latter accounting for inequality between the average incomes of the population subgroups. Then,

[1] University of Milan Bicocca. Email: mauro.mussini1@unimib.it.

subtracting these two measures of inequality from the overall Gini index, one obtains the overlapping component of the overall inequality index.

Dagum decomposition yields a breakdown of the Gini index on the basis of Gini pair wise income difference criterion, reckoning the inequality within subgroups and the inequality across subgroups. While the within-group inequality is equivalent to the within-group component of total inequality measured by Silber technique, the across-group component measures the inequality due to the disparities between subgroup income distributions instead of the differences between subgroup average incomes (Mussard, 2004). By using the concept of relative economic affluence (Dagum, 1987), Dagum further decomposed the across-group inequality into the net across-group inequality and the inequality contribution attributable to transvariation. Monti (2008) pointed out that the net across-group and the transvariation components are equivalent to the well known between-group and overlapping inequalities, respectively. Costa (2008) thoroughly discussed Dagum methodology focusing on the role of overlapping component and introducing simplified expressions for Dagum decomposition when a population is partitioned in two subgroups.

This work links Silber and Dagum ways of decomposition, thereby providing a decomposable matrix expression for the Gini index relying on Gini pair wise difference criterion and *G*-matrix operator. This matrix decomposition has the appealing feature of deriving the overlapping (transvariation) component as an intuitive and straightforward contribution of the inequality across subgroup income distributions. A further attractiveness of this procedure is that it provides a matrix compact form for counting the absolute, relative frequency of overlapping phenomenon.

The rest of the paper is organized as follows. Section 2 outlines Silber decomposition of Gini index. Section 3 briefly recalls Dagum decomposition approach and introduces a new matrix-based decomposition technique combining Dagum and Silber procedures. Section 4 illustrates a matrix compact form for counting transvariation cases on the basis of the matrix approach introduced in Section 3, paying attention to its possible application when dealing with the aim of measuring stratification; in Subsection 4.1, we apply the proposed decomposition procedure to sample data providing information on Italian employee incomes in 2000 and 2008. Section 5 concludes.

2. Silber matrix form decomposition

In the literature, the use of matrix algebra for computing the Gini index was introduced by Pyatt (1976) who proposed a matrix decomposition approach on the basis of game theory. Silber (1989) suggested another matrix form for calculating and decomposing the Gini index. Given a population of size n and mean income μ, let $\mathbf{y}$ stand for the n by 1 vector whose elements $y_1, y_2, \ldots, y_n$ are arranged in decreasing order. Being s_i the proportion of total income earned by the income receiver whose income has the i-th rank in vector $\mathbf{y}$, Silber formula is as follows

$$G = \mathbf{e'Gs}\,, \tag{1}$$

where $\mathbf{s}$ is the column vector of the per capita income shares s_1, s_2,…, s_n sorted in decreasing order, $\mathbf{e}$ is an n by 1 column vector with elements equal to 1/n, and $\mathbf{G}$ is an n by n matrix whose elements g_{ij} are equal to -1 when $j > i$, to +1 when $j < i$, to 0 when $j = i$ (called G-matrix). Equation (1) can be decomposed by subgroup when income receivers are grouped according to some criterion (e.g. income class, socio-demographic characteristics). Suppose the n income earners are partitioned into h subgroups, with $n = \sum_{k=1}^{h} n_k$. Let us introduce the ordering criterion A arranging incomes by subgroup mean income in decreasing order (first key of ordering) and by per capita income in decreasing order within each subgroup (second key of ordering). Let $\mathbf{s}_A$ denote the vector of income shares arranged by the ordering A. By partitioning the n by n G-matrix into h^2 sub matrices, one obtains

$$G = \begin{bmatrix} G(n_1, n_1) & \cdots & G(n_1, n_h) \\ \vdots & \ddots & \vdots \\ G(n_h, n_1) & \cdots & G(n_h, n_h) \end{bmatrix}, \tag{2}$$

where the main diagonal block matrix $G(n_k, n_k)$ is an n_k by n_k G-matrix, whereas the n_k by n_l off-diagonal block matrix $G(n_k, n_l)$ has all elements equal to -1 when $k < l$, to +1 when $k > l$. Now, the within-group component of the Gini index is

$$G^W = \sum_{k=1}^{h} \mathbf{e'}(n_k)\mathbf{G}(n_k, n_k)\mathbf{s}_A(n_k), \tag{3}$$

where $\mathbf{e}(n_k)$ is an n_k by 1 vector with elements equal to 1/n and $\mathbf{s}_A(n_k)$ is an n_k by 1 vector whose elements are the elements of $\mathbf{s}_A$ belonging to subgroup k arranged in decreasing order. Indeed, the between-group Gini index can be written as

$$G^B = \sum_{k=1}^{h} \sum_{l \neq k}^{h} \mathbf{e'}(n_k)\mathbf{G}(n_k, n_l)\mathbf{s}_A(n_l). \tag{4}$$

Next, subtracting expressions in (3) and (4) from the Gini expression for ungrouped data in (1), one obtains

$$\begin{aligned} G^T &= \mathbf{e'Gs} - \sum_{k=1}^{h} \mathbf{e'}(n_k)\mathbf{G}(n_k,n_k)\mathbf{s}_A(n_k) - \sum_{k=1}^{h}\sum_{l\neq k}^{h} \mathbf{e'}(n_k)\mathbf{G}(n_k,n_l)\mathbf{s}_A(n_l) \\ &= \mathbf{e'Gs} - \mathbf{e'Gs}_A, \end{aligned} \qquad (5)$$

which represents the transvariation (overlapping) component of the Gini index. Given two subgroups of income receivers, any pair formed by two income receivers belonging to different subgroups is said to transvary if the sign of the difference between their incomes is opposite to the sign of the difference between the corresponding subgroup mean incomes. If there is at least one transvarying pair, the two subgroups are said to transvary. When dealing with income, this definition of transvariation is equivalent to the definition of overlapping. In fact, two subgroups of income receivers are said to overlap if at least one income belonging to the subgroup with lower mean income is higher than at least one income of the subgroup with higher mean income. Thus, we use, throughout the paper, transvariation and overlapping as synonymous terms.

3. A matrix decomposition linking Dagum and Silber approaches

It is well known that the Gini index can be expressed as a function of Gini relative mean difference (Gini, 1912)

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|y_i - y_j\right|}{2\mu n^2}. \qquad (6)$$

Starting from equation (6) based on the Gini pair wise difference criterion, Dagum (1997) decomposition yields a breakdown of the Gini index into a within-group contribution to the overall inequality

$$G^W = \frac{\sum_{k=1}^{h}\left(\sum_{i=1}^{n_k}\sum_{j=1}^{n_k}\left|y_{ki} - y_{kj}\right|\right)}{2\mu n^2}, \qquad (7)$$

and an across-group contribution to the total inequality due to the disparities between subgroup income distributions

$$G^{AG} = \frac{\sum_{k=1}^{h}\sum_{l\neq k}^{h}\left(\sum_{i=1}^{n_k}\sum_{j=1}^{n_l}\left|y_{k,i} - y_{l,j}\right|\right)}{2\mu n^2}. \qquad (8)$$

In general, equations (4) and (8) are not equivalent, the former accounts for inequality between subgroup mean incomes, the latter considers all possible pair wise disparities between incomes belonging to different subgroups. The between-group and the across-group components coincide only if subgroup income distributions do not overlap (Mookherjee and Shorrocks, 1982).

Now, we combine Silber matrix technique and Dagum approach by providing a new matrix compact form for the Gini index. Let us introduce the *n* by 1 vector $\mathbf{1}_n$ with elements equal to 1, then $\mathbf{D}=\mathbf{1}_n\mathbf{y}'-\mathbf{y}\mathbf{1}_n'$ represents the skew symmetric matrix containing the n^2 pair wise differences between incomes of vector $\mathbf{y}$ defined in the previous section. Using the Hadamard product operator (see Vernizzi (2009) and Vernizzi *et al.* (2010) for further details on the use of Hadamard product for Gini computation purpose), denoted by $\circ$, it can be verified that

$$G=\frac{1}{2\mu}\mathbf{e}'\left(\mathbf{G}\circ\mathbf{D}\right)\mathbf{e}, \tag{9}$$

where $\mathbf{G}$ is an *n* by *n* *G*-matrix and $\mathbf{e}$ has dimension equal to *n* by 1. In (9), the entry wise product between $\mathbf{G}$ and $\mathbf{D}$ ensures that all non-positive pair wise differences in the upper diagonal entries of $\mathbf{D}$ change their sign. Thus, the matrix $\mathbf{G}\circ\mathbf{D}$ contains the n^2 non-negative pair wise income differences considered in (6). Since $\frac{1}{\mu n}\mathbf{y}=\mathbf{s}$, expression (9) can be written in terms of per capita income shares,

$$G=\frac{n}{2}\mathbf{e}'\left(\mathbf{G}\circ\mathbf{S}\right)\mathbf{e}. \tag{10}$$

where $\mathbf{S}=\mathbf{1}_n\mathbf{s}'-\mathbf{s}\mathbf{1}_n'$ is the skew symmetric matrix containing the n^2 pair wise differences between per capita income shares. Expression (10) can be re-arranged by using the Hadamard product properties yielding the following matrix compact formula (see Appendix A)

$$G=\frac{1}{2n}tr\left(\mathbf{G}\mathbf{S}'\right). \tag{11}$$

Next, we introduce the *n* by 1 vector $\mathbf{w}_k$ with non-zero elements equal to 1 in the corresponding positions filled by income shares belonging to subgroup *k* in the vector $\mathbf{s}$. Then, the square matrix $\mathbf{W}=\sum_{k=1}^{h}\mathbf{w}_k\mathbf{w}_k'$ has 1 element in the entries corresponding to pair wise differences involving two income shares of a same subgroup in $\mathbf{S}$. By inserting $\mathbf{W}$ in equation (10), one obtains

$$\begin{aligned}G^W&=\frac{n}{2}\mathbf{e}'\left(\mathbf{G}\circ\mathbf{W}\circ\mathbf{S}\right)\mathbf{e}\\&=\frac{1}{2n}tr\left(\mathbf{G}\mathbf{S}^{W}{}'\right),\end{aligned} \tag{12}$$

where $\mathbf{S}^{W} = \mathbf{W} \circ \mathbf{S}$. Expression (12) is equivalent to equations (3) and (7), yielding the within-group contribution to the overall inequality due to the disparities between income shares belonging to a same subgroup. By setting $\mathbf{J} = \mathbf{1}_n \mathbf{1}_n'$, the square matrix $\mathbf{J} - \mathbf{W} = \sum_{k=1}^{h} \sum_{l \neq k}^{h} \mathbf{w}_k \mathbf{w}_l'$ stands for the matrix selecting pair wise differences between income shares belonging to different subgroups from $\mathbf{S}$. Thus, the expression for the across-group component of the total inequality is

$$\begin{aligned} G^{AG} &= \frac{n}{2} \mathbf{e}' \left[\mathbf{G} \circ \left(\mathbf{J} - \mathbf{W} \right) \circ \mathbf{S} \right] \mathbf{e} \\ &= \frac{1}{2n} tr \left(\mathbf{G} \mathbf{S}^{AG}{}' \right), \end{aligned} \tag{13}$$

where $\mathbf{S}^{AG} = \left(\mathbf{J} - \mathbf{W} \right) \circ \mathbf{S}$.

Now, let $\mathbf{A}$ stand for the *n* by *n* permutation matrix re-arranging the elements of $\mathbf{s}$ in accordance with the ordering criterion *A* defined in the previous section, $\mathbf{s}_A = \mathbf{A}\mathbf{s}$. In equation (13), by replacing $\mathbf{S}$ and $\mathbf{W}$ with $\mathbf{S}_A = \mathbf{1}_n \mathbf{s}_A' - \mathbf{s}_A \mathbf{1}_n' = \mathbf{A}\mathbf{S}\mathbf{A}'$ and $\mathbf{W}_A = \mathbf{A}\mathbf{W}\mathbf{A}'$ respectively, after some algebraic manipulations one obtains (see Appendix B)

$$\begin{aligned} G^{B} &= \frac{n}{2} \mathbf{e}' \left[\mathbf{G} \circ \left(\mathbf{J} - \mathbf{W}_A \right) \circ \mathbf{S}_A \right] \mathbf{e} \\ &= \frac{1}{2n} tr \left(\mathbf{A}'\mathbf{G}\mathbf{A}\mathbf{S}^{AG}{}' \right), \end{aligned} \tag{14}$$

which is equivalent to Silber expression for the between-group Gini component defined in (4). Then, subtracting the expression for G^B in (14) from the formula (13), we have the expression for the transvariation (overlapping) component

$$G^{T} = \frac{1}{2n} tr \left[\left(\mathbf{G} - \mathbf{A}'\mathbf{G}\mathbf{A} \right) \mathbf{S}^{AG}{}' \right] \tag{15}$$

which is equivalent to (5). The matrix $\mathbf{G} - \mathbf{A}'\mathbf{G}\mathbf{A}$ detects the transvarying pairs from the matrix $\mathbf{S}$: it has non-zero elements equal to -2 (+2) in its upper (lower) diagonal entries corresponding to transvarying pairs in $\mathbf{S}$. The matrix $\mathbf{A}'\mathbf{G}\mathbf{A}$ identifies violations of the decreasing ordering of income shares in $\mathbf{s}_A$ by showing a value opposite to that of $\mathbf{G}$ in each entry corresponding to a difference in $\mathbf{S}$ between two incomes representing a transvarying pair, while $\mathbf{A}'\mathbf{G}\mathbf{A}$ and $\mathbf{G}$ have the same elements in the entries corresponding to non-transvarying pairs. It follows immediately that subtracting $\mathbf{A}'\mathbf{G}\mathbf{A}$ from $\mathbf{G}$, the non-zero elements (equal to -2 in the upper-diagonal entries, to +2 in the lower-diagonal ones) of $\mathbf{G} - \mathbf{A}'\mathbf{G}\mathbf{A}$ fill the entries corresponding to transvarying pairs in the matrix $\mathbf{S}$. A simple numerical example is shown to provide a practical explanation of the role of $\mathbf{A'GA}$. Consider four income receivers partitioned in two subgroups each containing two individuals and let the income shares of subgroup 1 be given by 0.4 and 0.2, and those of subgroup 2 by 0.3 and 0.1. The vector of income shares sorted by decreasing order is $\mathbf{s}' = \begin{pmatrix}0.4 & 0.3 & 0.2 & 0.1\end{pmatrix}$ whereas the vector of income shares sorted by the ordering A is $\mathbf{s}_A' = \begin{pmatrix}0.4 & 0.2 & 0.3 & 0.1\end{pmatrix}$. Then, the pair of income shares (s_2=0.3, s_3=0.2) represents a transvarying pair. Thus the matrix $\mathbf{A'GA}$ has a +1 element in its (2,3)the entry and a -1 in the (3,2)the entry instead of, respectively, a -1 and a +1 element as shown in $\mathbf{G}$. The remaining elements of $\mathbf{A'GA}$ coincide with the elements of $\mathbf{G}$. Then, it can easy verified that $\mathbf{G} - \mathbf{A'GA}$ is a matrix with the (2,3)the element equal to -2 and the (3,2)the one equal to +2, while the other elements are equal to zero.

In the case of one pair of identical income shares belonging to different subgroups, this pair is denoted by non-zero elements in $\mathbf{G} - \mathbf{A'GA}$ only if the income share belonging to poorer (on average) subgroup is ranked ahead of the income share belonging to richer (on average) subgroup in vector $\mathbf{s}$. Regardless of the convention adopted for ranking identical income shares belonging to different subgroups in $\mathbf{s}$, one pair of identical shares provides a null contribution to the transvariation component because the corresponding elements of $\mathbf{S}$ are equal to zero. In the following, we shall come back to the choice of the ordering criterion for ranking identical income shares of different subgroups in order to obtain matrix expressions counting transvarying pairs.

If transvariation does not exist, the vector of income shares arranged by the ordering criterion A coincides with the vector of income shares sorted in decreasing order, and then $\mathbf{A} = \mathbf{I}_n$ yielding $G^T = 0$.

Given a unique matrix containing all the information on the disparities concerning the income distribution, denoted by $\mathbf{S}$, the presented decomposition technique has the appealing feature of deriving the various components of the overall inequality by separating the pair wise inequality contributions contained in $\mathbf{S}$ by means of permutation and selection matrices defined according to a population partition into exclusive and exhaustive subgroups. Starting from the matrix expression for the across-group inequality, the transvariation inequality is achieved as a clear and straightforward contribution of disparities between subgroup distributions, while the between-group inequality is calculated without assuming that all income receivers of a subgroup have the same income. As noted in Ebert (2010), G^{AG} and G^B represent two extreme ways of measuring disparity between subgroup distributions, one accounting for inequality between all pairs of incomes belonging to different subgroups (G^{AG}), the other reducing disparity between subgroups to inequality between subgroup average incomes (G^B). Utilizing the full information available, the proposed decomposition technique reconciles these extreme possibilities of determining inequality between subgroup distributions by showing that both G^{AG} and G^{B} can be derived from the same standpoint, one matrix containing full information, and without employing average incomes.

4. A new matrix expression for counting transvariation and its application

In this section, we show that the proposed technique can be used to derive a convenient matrix form for counting the absolute, relative frequency of transvarying pairs.

By considering per capita income shares, the total number of transvarying pairs can be expressed as follows

$$T=\sum_{k=2}^{h}\sum_{l=1}^{k-1}\left(\sum_{i=1}^{n_k}\sum_{j=1}^{n_l}\eta\left(s_{ki},s_{lj}\right)\right), \tag{16}$$

where

$$\eta\left(s_{ki},s_{lj}\right)=\begin{cases}1 & if \quad \left(s_{ki}-s_{lj}\right)\left(\overline{s}_k-\overline{s}_l\right)<0\\ 0 & if \quad \left(s_{ki}-s_{lj}\right)\left(\overline{s}_k-\overline{s}_l\right)>0\\ \dfrac{1}{2} & if \qquad\qquad s_{ki}=s_{lj}\end{cases},$$

with the adopted convention of counting one half each pair having identical members (Calò, 2006) and the mean income share of subgroup k denoted by $\overline{s}_k$. In order to rewrite equation (16) in matrix compact form we introduce the following notation. Let $\mathbf{s}_U$ stand for the vector of income shares arranged in decreasing order, where the identical income shares are sorted by subgroup mean in decreasing order. Let $\mathbf{s}_L$ stand for the vector of income shares arranged in decreasing order, where the identical income shares are sorted by subgroup mean in increasing order. Then, $\mathbf{A}_U$ and $\mathbf{A}_L$ denote the permutation matrices satisfying the equalities $\mathbf{s}_A=\mathbf{As}=\mathbf{A}_U\mathbf{s}_U=\mathbf{A}_L\mathbf{s}_L$. It is worth emphasizing that by replacing $\mathbf{A}$ in (15) with $\mathbf{A}_U$ or $\mathbf{A}_L$, one has expressions equivalent to (15) as the ordering of elements in $\mathbf{s}$ does not pay attention to the reciprocal positions filled by identical income shares. However, the attempt of counting transvarying pairs in matrix form needs to specify the ordering criterion for identical income shares when sorting income shares in decreasing order.

By using $\mathbf{A}_U$, $\mathbf{A}_L$ and $\mathbf{G}$, a compact matrix form for counting the number of transvarying pairs can be written as (see appendix C)

$$
\begin{aligned}
T &= \frac{1}{4} tr\left[\mathbf{GG'} - \mathbf{A}_U\mathbf{'GA}_U\mathbf{G'}\right] + \frac{1}{8} tr\left[\mathbf{A}_U\mathbf{'GA}_U\mathbf{G'} - \mathbf{A}_L\mathbf{'GA}_L\mathbf{G'}\right] \\
&= T^A + T^C.
\end{aligned} \qquad (17)
$$

In (17), T^A represents the number of pairs whose members actually transvary, while T^C accounts for pairs having identical members (conventional transvariation). Expression (17) can be re-arranged without the separation between actual and conventional transvariations, yielding

$$
T = \frac{n(n-1)}{4} - \frac{1}{8} tr\left[\mathbf{A}_U\mathbf{'GA}_U\mathbf{G'} + \mathbf{A}_L\mathbf{'GA}_L\mathbf{G'}\right]. \qquad (18)
$$

Given a population of size *n*, the total number of combinations by taking two members at a time out of population is equal to $n(n-1)/2$. Then, by dividing *T* by $n(n-1)/2$ one obtains the ratio between the number of transvarying pairs and the number of all possible pairs of income shares,

$$
PT = \frac{1}{2} - \frac{1}{4n(n-1)} tr\left[\mathbf{A}_U\mathbf{'GA}_U\mathbf{G'} + \mathbf{A}_L\mathbf{'GA}_L\mathbf{G'}\right] \quad .(19)
$$

The presented matrix approach enables us to achieve the transvariation component as an intuitive and precise contribution to the overall inequality and, at the same time, to compute the absolute, relative frequency of the transvariation phenomenon. This can represent an advantage when dealing with the measurement of subgroup separation, such as the definition of stratification indices that requires an in-depth analysis of overlapping. As pointed out in the literature (Yitzhaki and Lerman, 1991; Yitzhaki, 1994; Frick *et al.*, 2006), overlapping can be seen as the inverse of stratification. Since a raise in overlapping diminishes the separation of subgroup distributions, the greater the degree of overlapping between subgroup distributions the greater the distance from the perfect stratification between subgroup distributions. By interpreting overlapping as non-stratification between population subgroups, we notice that, while the overlapping concept is linked to between-group inequality (it is defined with respect to inequality between subgroup average incomes), the well known between-group inequality measure only accounts for differences in subgroup average incomes, neglecting overlapping degree between subgroup distributions. Recent studies (Liao, 2008; Monti and Santoro, 2009) argued that stratification measurement is closely related to between-group inequality measurement. Monti and Santoro (2009) provided a Gini decomposition in which the between-group inequality component captures changes in stratification degree between two population subgroups by counting the number of transvarying pairs. Starting from Dagum decomposition of the Gini index, Liao (2008) developed stratification indices based on the between-group inequality measure and the frequency of pair wise income comparisons involving incomes of different subgroups.

For instance, we consider the following stratification index proposed by Liao (2008) and defined in absence of transvariation,

$$I_3 = I_1 I_2 \tag{20}$$

where

$$I_1 = \frac{G - G^W}{G}$$

and

$$I_2 = \frac{\sum_{k=2}^{h} \sum_{l=1}^{k-1} \left(\sum_{i=1}^{n_k} \sum_{j=1}^{n_l} d \right)}{\left(n^2 - n \right) / 2},$$

where *d* equals 1. The index I_3 is compounded by two other stratification indices, one accounting for the proportion of total inequality ascribed to between-group inequality, denoted by I_1, another reckoning the proportion of the number of pair wise comparisons between members of different subgroups, indicated by I_2. By means of the index I_3, the proportion of the Gini index due to between-group inequality is made relative to the proportion of the pair wise comparisons between members of different population subgroups. I_3 varies from 0, where no stratification exists, to 1, where there is full stratification. In the real world, no overlapping subgroup income ranges are seldom observed, and then we attempt to extend the above considered stratification index to the case of overlapping subgroups. In order to take into account the presence of overlapping (non-stratification), we modify the indices I_1 and I_2 as follows

$$I_1^* = I_1 - \frac{G^T}{G} \tag{21}$$

and

$$I_2^* = I_2 - PT\,; \tag{22}$$

thus, we have

$$I_3^* = I_1^* I_2^*\,. \tag{23}$$

The index I_3^* ranges between 0 (no stratification) and 1 (full stratification), and it coincides with I_3 when dealing with no overlapping subgroup distributions. In (23) the proportion of the Gini ascribed to between-group inequality (measured by I_1^*) is made relative to the proportion of pair wise income differences between members belonging to different subgroups that actually contributes to subgroup separation, which is obtained by subtracting the proportion of transvarying pairs from the proportion of pair wise differences between members of different subgroups (index I_2^*).

4.1.Empirical application

We provide an empirical application of the methodology developed in this paper by using income data collected by the Survey of Household Income and Wealth (henceforth, SHIW) of the Bank of Italy in 2000 and 2008 (Banca d'Italia, 2010). By considering individual data concerning employee incomes, we decompose the Gini inequality index by gender. Furthermore, we apply the modified index of stratification (I_3^*) in order to measure the stratification between subgroups. As noticed in the literature (Sørensen, 1990; Monti and Santoro, 2009), gender discrimination on the labour market often implies that men earn more than women. This fact can be seen as an example of stratification, where income inequality between subgroups defined by gender represents an aspect linked to the existence of different strata. Since stratification is a phenomenon changing over time, it seems interesting to measure stratification variation over a given period (in this case the 2000–2008 period).

Both surveys were carried out by considering household as sample unit, so sample weights are referred to households. Even if household is the basic statistical unit, household weight can be used for each household member when considering individual incomes. In the performed analysis we applied weights provided by the central bank of Italy for each household. After deleting non-positive incomes, the samples are composed of 5911 individual incomes in 2008 and 6439 incomes in 2000. Incomes are expressed in Lire in the year 2000, whereas incomes referred to 2008 are expressed in Euro.

Tables 1 reports the results of the Gini index decomposition, the stratification indices, and some descriptive statistics for the 2000-2008 period. As expected, men average income is higher than women one in both 2000 and 2008. From Table 1, it emerges that overall inequality decreases between 2000 and 2008. This decrease is mainly attributable to the lower within-group inequality shown in 2008. By looking at the stratification index I_3^* in Table 1, we observe that the year 2000 shows an index I_3^* of 0.1111 while the year 2008

Table 1. Gini decomposition, stratification indices and descriptive statistics: Italy 2000–2008.

Sample	2008	2000
G	0.2573898	0.2663702
G^W	0.1286575	0.1368498
G^{AG}	0.1287323	0.1295204
G^B	0.07700322	0.09582038
G^T	0.05172908	0.03370006
I_1^*	0.2991697	0.3597264
I_2^*	0.3172425	0.3089238
I_3^*	0.0949034	0.1111280
N. of observations	5911 (women=2000, men=3911)	6439 (women=2598, men=3841)
Women av. income	14205.52	21594.68*
Men av. income	18031.14	27750*
Av. income	16373.34	25277.12*

** Average income is expressed in thousands of Lire.*

Source: elaborations on SHIW data.

exhibits an index I_3^* of 0.0949, suggesting that both the situations are far from full stratification. By comparing the index I_3^* of 2000 with the index I_3^* of 2008, we observe that the stratification decreased between 2000 and 2008. This decrease can be ascribed to the ascent of transvariation inequality in 2008. Even if the proportion of across-group pairwise differences contributing to subgroup separation is higher in 2008 (measured by I_2^*), its effect is counterbalanced by the fall of the between-group inequality due to the raise in G^T, as noted above. Thus, the intensity of transvariation, in terms of inequality, is more important than its relative frequency when we isolate the determinants of stratification variation moving from 2000 to 2008. Moreover, it is worth emphasizing that our empirical findings suggest that an increase in transvariation inequality component does not lead to a major role of within-group component on total inequality. In fact, the within-group component decreased between 2000 and 2008 whereas the overlapping component increased, underlining that a raise of the extent to which subgroup distributions overlap does not imply an increase of the inequality within subgroups.

5. Conclusions

The contribution of this article is two-fold. First, it has provided a matrix form decomposition of the Gini index which combines Silber matrix approach and Dagum decomposition based on the Gini pairwise difference criterion, thereby enabling to derive the within-group, the between-group and the transvariation components of total inequality by exploiting all of the information contained in one matrix accounting for disparities between per capita income shares of the overall population. Second, this matrix approach yields the quantification of transvariation in terms of inequality intensity and frequency of the event occurrence. This implies that the suggested matrix approach can be useful not only for decomposing inequality but also measuring phenomena closely linked with transvariation, such as stratification.

In conclusion, the proposed matrix approach offers a complete tool for the Gini decomposition which can be applied when dealing with no overlapping as well as overlapping subgroups.

Appendix A

This appendix proves that

$$G=\frac{n}{2}\mathbf{e}'\left(\mathbf{G}\circ\mathbf{S}\right)\mathbf{e}=\frac{1}{2n}tr\left(\mathbf{GS}'\right). \tag{A1}$$

We recall some useful properties of the Hadamard product. Given the n by n matrices $\mathbf{O}$, $\mathbf{P}$, $\mathbf{Q}$, where $\mathbf{O}$ is a permutation matrix, the n by 1 vector $\boldsymbol{\lambda}$ and the diagonal matrix $\boldsymbol{\Lambda}$ whose diagonal elements are the elements of $\boldsymbol{\lambda}$, the following properties hold:

i) $\mathbf{O}'\left(\mathbf{P}\circ\mathbf{Q}\right)\mathbf{O}=\left(\mathbf{O}'\mathbf{PO}\right)\circ\left(\mathbf{O}'\mathbf{QO}\right)$ (see Faliva, 1996, pp. 157);

ii) $\boldsymbol{\lambda}'\left(\mathbf{P}\circ\mathbf{Q}\right)\boldsymbol{\lambda}=tr\left(\mathbf{P}\boldsymbol{\Lambda}\mathbf{Q}'\boldsymbol{\Lambda}\right)$ (see Abadir and Magnus, 2005, pp. 340).

By using the property *ii*) , we have

$$G=\frac{n}{2}tr\left(\mathbf{GES}'\mathbf{E}\right), \tag{A2}$$

where $\mathbf{E}$ is a diagonal matrix with diagonal elements equal to *1/n*. Now, given that $\mathbf{E}=\frac{1}{n}\mathbf{I}_n$, it follows

$$\begin{aligned}G&=\frac{n}{2n^2}tr\left(\mathbf{GI}_n\mathbf{S}'\mathbf{I}_n\right)\\&=\frac{1}{2n}tr\left(\mathbf{GS}'\right).\end{aligned} \tag{A3}$$

Appendix B

This appendix proves that

$$G^B = \frac{n}{2}\mathbf{e}'\left[\mathbf{G}\circ\left(\mathbf{J}-\mathbf{W}_A\right)\circ\mathbf{S}_A\right]\mathbf{e} = \frac{1}{2n}tr\left(\mathbf{A}'\mathbf{GAS}^{AG}{}'\right). \qquad \text{(B1)}$$

Let recall the ordering criterion *A* defined in Section 2, for which incomes are sorted by decreasing subgroup mean income (first key) and, within each subgroup, in decreasing order (second key). Supposing that incomes are sorted by the ordering *A*, the well-known expression for the between-group inequality (Mookherjee and Shorrocks, 1982) can be written as

$$\begin{aligned} G^B &= \frac{1}{2n^2\mu}\sum_{k=1}^{h}\sum_{l\neq k}^{h} n_k n_l \left|\mu_k - \mu_l\right| \\ &= \frac{1}{2n^2\mu}\sum_{k=1}^{h}\sum_{l=k+1}^{h} 2n_k n_l \left(\mu_k - \mu_l\right) \\ &= \frac{1}{2n^2\mu}\sum_{k=1}^{h}\sum_{l=k+1}^{h}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} 2\left(y_{ki} - y_{lj}\right) \\ &= \frac{1}{2n}\sum_{k=1}^{h}\sum_{l=k+1}^{h}\sum_{i=1}^{n_k}\sum_{j=1}^{n_l} 2\left(s_{ki} - s_{lj}\right), \end{aligned} \qquad \text{(B2)}$$

where G^B is expressed as a function of the across-group pairwise share differences. Being $\mathbf{A}$ the permutation matrix re-arranging the element of $\mathbf{s}$ in accordance with the ordering *A*, it follows that $\mathbf{s}_A = \mathbf{As}$ is the vector of income shares sorted by the ordering *A* and $\mathbf{w}_{Ak} = \mathbf{Aw}_k$ denotes the vector with non-zero elements equal to 1 in the positions filled by income shares of subgroup *k* in $\mathbf{s}_A$. Thus, the matrices $\mathbf{S}_A = \mathbf{1}_n\mathbf{s}_A' - \mathbf{s}_A\mathbf{1}_n'$ and $\mathbf{W}_A = \sum_{k=1}^{h}\mathbf{w}_{Ak}\mathbf{w}_{Ak}'$ are, respectively, the matrix containing pairwise differences obtained from shares sorted by the ordering *A* and the matrix selecting within-group pairwise differences from $\mathbf{S}_A$. Then, the matrix compact form accounting for across-group pairwise differences in $\mathbf{S}_A$ is equivalent to the between-group inequality component as expressed in (B2),

$$G^B = \frac{n}{2}\mathbf{e}'\left[\mathbf{G}\circ\left(\mathbf{J}-\mathbf{W}_A\right)\circ\mathbf{S}_A\right]\mathbf{e}. \qquad \text{(B3)}$$

Then, since $\mathbf{J}-\mathbf{W}_A = \mathbf{A}\left(\mathbf{J}-\mathbf{W}\right)\mathbf{A}'$ and $\mathbf{S}_A = \mathbf{ASA}'$, using the property *i*) in Appendix A, expression (B3) can be re-written as

$$\begin{aligned} G^B &= \frac{n}{2}\mathbf{e}'\left\{\mathbf{G}\circ\mathbf{A}\left[\left(\mathbf{J}-\mathbf{W}\right)\circ\mathbf{S}\right]\mathbf{A}'\right\}\mathbf{e} \\ &= \frac{n}{2}\mathbf{e}'\left\{\mathbf{G}\circ\mathbf{A}\mathbf{S}^{AG}\mathbf{A}'\right\}\mathbf{e}. \end{aligned} \tag{B4}$$

We now use the result obtained in Appendix A, so that expression (B4) becomes

$$\begin{aligned} G^B &= \frac{1}{2n}tr\left(\mathbf{G}\mathbf{A}\mathbf{S}^{AG}{}'\mathbf{A}'\right) \\ &= \frac{1}{2n}tr\left(\mathbf{A}'\mathbf{G}\mathbf{A}\mathbf{S}^{AG}{}'\right) \end{aligned}.$$

Appendix C

This appendix proves that

$$\begin{aligned} T &= \frac{1}{4}tr\left(\mathbf{G}\mathbf{G}'-\mathbf{A}_U{}'\mathbf{G}\mathbf{A}_U\mathbf{G}'\right)+\frac{1}{8}tr\left(\mathbf{A}_U{}'\mathbf{G}\mathbf{A}_U\mathbf{G}'-\mathbf{A}_L{}'\mathbf{G}\mathbf{A}_L\mathbf{G}'\right) \\ &= T^A+T^C. \end{aligned} \tag{C1}$$

The matrix $\mathbf{G}-\mathbf{A}'\mathbf{G}\mathbf{A}$ entered in (15) detects transvarying pairs by identifying, with non-zero elements (± 2), violations of decreasing ordering of income shares when moving from $\mathbf{s}$ to $\mathbf{s}_A$. However, a further ordering criterion has to be added in order to detect pairs with identical members (income shares) belonging to different subgroups. By assuming identical shares are sorted by subgroup mean in decreasing order in $\mathbf{s}$, the permutation matrix $\mathbf{A}_U$ re-arranges them in $\mathbf{s}_A$ maintaining their reciprocal positions; then only actual transvarying pairs are detected by $\mathbf{G}-\mathbf{A}_U{}'\mathbf{G}\mathbf{A}_U$, the number of which can be expressed as follows

$$\begin{aligned} T^A &= \frac{1}{4}\mathbf{1}_n{}'\left[\left(\mathbf{G}-\mathbf{A}_U{}'\mathbf{G}\mathbf{A}_U\right)\circ\mathbf{G}\right]\mathbf{1}_n \\ &= \frac{1}{4}tr\left[\mathbf{G}\mathbf{G}'-\mathbf{A}_U{}'\mathbf{G}\mathbf{A}_U\mathbf{G}'\right]. \end{aligned} \tag{C2}$$

where $\mathbf{1}_n$ is an n by 1 vector with elements equal to 1. By sorting identical income shares by subgroup mean in increasing order in $\mathbf{s}$, their reciprocal positions in $\mathbf{s}_A$ are exchanged by means of the permutation matrix $\mathbf{A}_L$. Thus, the matrix $\mathbf{G}-\mathbf{A}_L{}'\mathbf{G}\mathbf{A}_L$ detects pairs having identical members as transvarying pairs (conventional transvarying pairs), plus the actual transvarying pairs. In order to quantify the number of conventional transvarying pairs, one can subtract the number of actual transvarying pairs expressed in (C2) from the number of transvarying pairs obtained by replacing $\mathbf{G}-\mathbf{A}_U\mathbf{'GA}_U$ in (C2) with $\mathbf{G}-\mathbf{A}_L\mathbf{'GA}_L$ accounting for actual as well as conventional transvarying pairs. Therefore, transvarying pairs with identical members can be counted in matrix formula as follows

$$\begin{aligned} T^C &= \frac{1}{2}\left\{\frac{1}{4}\mathbf{1}_n\mathbf{'}\left[\left(\mathbf{G}-\mathbf{A}_L\mathbf{'GA}_L\right)\circ\mathbf{G}\right]\mathbf{1}_n - \frac{1}{4}\mathbf{1}_n\mathbf{'}\left[\left(\mathbf{G}-\mathbf{A}_U\mathbf{'GA}_U\right)\circ\mathbf{G}\right]\mathbf{1}_n\right\} \\ &= \frac{1}{8}\mathbf{1}_n\mathbf{'}\left\{\left[\left(\mathbf{G}-\mathbf{A}_L\mathbf{'GA}_L\right)-\left(\mathbf{G}-\mathbf{A}_U\mathbf{'GA}_U\right)\right]\circ\mathbf{G}\right\}\mathbf{1}_n \qquad \text{(C3)} \\ &= \frac{1}{8}\mathbf{1}_n\mathbf{'}\left[\left(\mathbf{A}_U\mathbf{'GA}_U-\mathbf{A}_L\mathbf{'GA}_L\right)\circ\mathbf{G}\right]\mathbf{1}_n, \end{aligned}$$

where the denominator in (C3) is 8 instead of 4 since we adopt the convention of counting one half each transvarying pair with identical members, as said in the main text. Next, using the result obtained in Appendix A, we can re-arrange expression (C3) as follows

$$T^C = \frac{1}{8}tr\left(\mathbf{A}_U\mathbf{'GA}_U\mathbf{G'}-\mathbf{A}_L\mathbf{'GA}_L\mathbf{G'}\right).$$

REFERENCES

ABADIR, K. M., MAGNUS, J. R., (2005), Matrix Algebra. Cambridge University Press, New York.

BANCA D’ITALIA, (2010), Survey on Households Income and Wealth. Supplements to the Statistical Bulletin – sample survey in 2008, sample survey in 2000. Available at http://www.bancaditalia.it/statistiche/indcamp/bilfait/boll_stat;internal&action=_setlanguage.action?LANGUAGE=en

BERREBI, Z. M., SILBER J., (1987), Regional Differences and the components of Growth and Inequality Change. *Economics Letters*, **25**, 295–298.

BHATTACHARYA, N., MAHALANOBIS, B., (1967), Regional disparity in household consumption in India. *American Statistical Association Journal*, March, 143–162.

CALÒ, D. G., (2006), On a Transvariation Based Measure of Group Separability. *Journal of Classification*, **23**, 143–167.

COSTA, M., (2008), Gini Index Decomposition for the Case of Two Subgroups. *Communications in Statistics – Simulation and Computation*, **37**(4), 631–644.

DAGUM, C., (1987), Measuring the Economic Affluence between Populations of Income Receivers. *Journal of Business & Economic Statistics*, **5**(1), 5–12.

DAGUM, C., (1997), A new approach to the decomposition of Gini income inequality ratio. *Empirical Economics*, **22**, 515–531.

EBERT, U., (2010), The decomposition of inequality reconsidered: Weakly decomposable measures. *Mathematical Social Sciences*, **60**, 94–103.

FALIVA, M., (1996), Hadamard matrix product, graph and system theories: motivations and role in Econometrics. In: Camiz, S., Stefani, S. (eds.) Matrices and Graphs, Theory and Applications to Economics, pp 152–175. World Scientific, London.

FRICK, J., GOEBEL, J., SCHECHTMAN, E., WAGNER, G., YITZHAKI, S., (2006), Using Analysis of Gini (ANOGI) for Detecting Whether Two Subsamples Represent the Same Universe: The German Socio-Economic Panel Study (SOEP) Experience. *Sociological Methods & Research*, **34**, 427–468.

GINI, C., (1912), Variabilità e Mutabilità. Bologna: Tipografia Paolo Cuppini.

GIORGI, G. M., (1999), Income Inequality Measurement: the statistical approach. In: Silber, J. (eds.) Handbook on Income Inequality Measurement, pp 245–267. Kluwer, Boston.

LIAO, T. F., (2008), The Gini unbound: analyzing class inequality with model-based clustering. In: Betti, G., Lemmi, A. (eds.) Advances on income inequality and concentration measures, pp 201-221. Routledge, New York.

MONTI, M. G., (2008), A note on the residual term R in the decomposition of the Gini Index. Argumenta Oeconomica, **20**, 107–138.

MONTI, M. G., SANTORO, A., (2009), A note on between-groups inequality with an application to genders. Working Paper Series Econpubblica CRPS, n. 135 October available at
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1338054

MOOKHERJEE, D., SHORROCKS, A.F., (1982), A decomposition analysis of the trend in UK income inequality. *Economic Journal*, **92**, 886–902.

MUSSARD, S., (2004), The Bidimensional Decomposition of the Gini Ratio. A case Study: Italy. *Applied Economics Letters*, **11**(8), 503–505.

PYATT, G., (1976), On the interpretation and disaggregation of Gini Coefficient. *The Economic Journal*, **86**, 243–255.

SØRENSEN, J., (1990), Perception of women's opportunity in five industrialized nations. *European Sociological Review*, **6**(2), 151–164.

SHORROCKS, A.F., (1984), Inequality decomposition by population subgroups. *Econometrica*, **52**, 1369–1385.

SILBER, J., (1989), Factor Components, Population Subgroups and the Computation of the Gini Index of Inequality. *Review of Economics and Statistics*, **71**, 107–115.

VERNIZZI, A., (2009), Applying the Hadamard product to decompose Gini, concentration, redistribution and re-ranking indexes. *Statistics in Transition*, **10**(3), 505–525.

VERNIZZI, A., MONTI, M.G., MUSSINI, M., (2010), A Gini and Concentration Index Decomposition with an Application to the APK Reranking Measure. DEAS Working Paper n. 2010-10, Available at http://www.economia.unimi.it/uploads/wp/DEAS-2010_10wp.pdf

YITZHAKI, S., LERMAN, R., (1991), Income stratification and income inequality. *Review of Income and Wealth*, **37**(3), 313–329.

YITZHAKI, S., (1994), Economic distance and overlapping distributions. *Journal of Econometrics*, **61**, 147–159.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 615—638

A TYPOLOGY OF POLISH FARMS USING PROBABILISTIC D–CLUSTERING

Andrzej Młodak[1], Jan Kubacki[2]

ABSTRACT

The Agricultural Census conducted in Poland in 2010 was partially based on administrative sources. These data collection will be supplemented by sample survey of agricultural farm. This research is aimed at creation of an effective typology of Polish farms, which is necessary for proper sampling and reflection of many special types of agricultural activity, such as combining it with non–agricultural work. We propose some universal form of such typology constructed using data collected from administrative sources during the preliminary agricultural census conducted in autumn 2009. It is based on the especially prepared method of fuzzy clustering, i.e. probabilistic d–clustering adopted for interval data. For this reason, and because of an ambiguous impact of some key variables on classification, relevant criterions are presented as intervals. They are arbitrarily established, but also – as an alternative way – are generated endogenically, using an original optimization algorithm. For a better comparison, relevant classification for data collected "from nature" is provided.

Key words: agricultural census, probabilistic d–clustering, interval data.

1. Introduction

The Agricultural Census in Poland in 2010 was conducted according to significantly different rules than those used in its previous exercises. The main source of information gathered during the census was administrative databases. For instance, most of such data was collected from the farm registers maintained by local self – government authorities (such as Tax Register of Real Estates, Register of Lands and Buildings) or by the Agency for Restructuring and Modernization of Agriculture, which is engaged in services of applications of farmers concerning subventions from the European Union budget. The direct detailed survey (concerning mainly methods of agricultural production) was

[1] Statistical Office in Poznań, Urban Statistics Centre. E–mail: a.mlodak@stat.gov.pl.

[2] Statistical Office in Łódź, Centre of Realization of Statistical Surveys. E–mail: j.kubacki@stat.gov.pl.

planned to be conducted on 30% sample of farms with agricultural land area up to 1 ha and some farms with this area between 1 ha and 2 ha (the remaining were planned to be examined in an exhaustive way). The farms for which no or very few administrative data are available were also additionally interviewed (target survey).

Taking these expectations into account and following the fact that structure of Polish farms gradually changes (among others due to the accession of Poland to the European Union) and competitiveness on the market of agricultural products is ever-increasing, there is a necessity to construct a typology of farms which could allow these changes to be reflected also in future statistical agricultural surveys. Additionally, it is connected with the fact that the Agricultural Census will be conducted according to the rules of the Farm Structure Survey adopted within the EU (Regulation (EC) No 1166/2008 of the European Parliament and of the Council). It means that this survey shall be carried out in the form of a census, i.e. it should cover farms with area greater than 1 ha. In the Polish circumstances a necessity to examine smaller farms also occurs. They are not, however, a significant part of the overall population of farms, and their production is not especially significant, but due to the above mentioned special character of the Polish agriculture, a survey of them seems to be also required. Moreover, it is very important to reflect many special or mixed types of agricultural activity, such as combining it with non–agricultural work.

It is clear that a creation of a universal typology of agricultural holdings is very difficult. Some scientists argue even that, in practice, it is not feasible. Nevertheless, some – at least relatively efficient – typology is necessary to properly conduct the statistical surveys. We have undertaken a trial to construct such categorization on the basis of a dataset which was assumed to be available for all farms at the moment when the census started. It was burdened with some inconveniences, which we have tried to eliminate, although it was not always fully possible. However, the most serious of them seem to be minimized.

To satisfy these expectations we have constructed a basic and universal typology of farms using some fuzzy probabilistic d–clustering. It is a generalization of the proposal of A. Ben – Israel and C. Iyigun (2008) into an interval case. Because many key features determining a character of the farm are described by interval or ratio variables with continuous distribution of observation, the criterion of classification of a farm to a given class should be expressed by a set of intervals reflecting typical scope of values of relevant variables realized within this class. For any object (farm) we determine a class such that the probability of belonging of a farm to it is the greatest. The object will be assigned to this class. The criterion intervals can be established arbitrarily or determined using an endogenous optimization based on derivation of interval-valued function. Both approaches are here presented.

Our method differs significantly from many popular fuzzy classification algorithms, such as c-means Bezdek's approach (J. C. Bezdek (1973), R. J. Hathaway and J. C. Bezdek (1988)), Gustafson – Kessel method (D. E. Gustafson and W. C. Kessel (1979)), Gath – Geva probabilistic suggestion (I. Gath and, A. B. Geva (1989)) or even unsupervised FCM–NM algorithm based on normalized Mahalanobis distance (J. – M. Yih and S. – F. Huangh (2010)). All these proposals have a common feature – they are based on the fuzzy c–means procedure and are point–oriented. That is, the clusters are determined by their point centers (usually centroids). Moreover, some of them tend, for instance, to create spherical shaped clusters (Bezdek's method) or deform original diversification of objects, which is very important in multivariate analysis (FCM–NM). Our method reflects the common practical situation, when classes are defined by reference intervals of respective variables. The three area groups of farms (above 2 ha, 1–2 ha and 0–1 ha) can be the simplest and very good example in this context. Therefore, an original distance of a point from an interval was defined. The variations of diagnostic variables are kept. The computation seems to be also much faster, because in its basic part it is non–iterative.

The presented experiment is a test study using data collected during the preliminary (trial) census before the main Agriculture Census conducted in autumn 2009. It was conducted in all farms located in the following four rural gminas (Polish NUTS 5 territorial units):

- Gniezno (Wielkopolskie Voivodship – Polish NUTS 2 region),
- Kamień (Podkarpackie Voivodship),
- Kołobrzeg (Zachodniopomorskie Voivodship) ,
- Rutki (Podlaskie Voivodship).

As a result of this survey, two databases have been constructed. The first of them is called the 'master record' and contains all data collected from administrative sources. This file was the main basis of our classification, because on a similar file (but covering the whole population of farms) the sampling and other primarily census activities will have to be performed. The second ('gold record') consists of information received "from nature", i.e. directly from the farms using the modern interview techniques (such as CATI – Computer Assisted Telephone Interview, CAPI – Computer Assisted Personal Interview, etc.) and, of course, is much greater than the "master record", because the scope of information which can be gathered during individual contact with respondent is much more broader than collected in official registers. The "gold record" reflects then the information collected directly from respondents during the preliminary census whereas the "master record" – the data from administrative sources for the same respondents obtained directly during this census (and which were assumed to be collected also just before the main census). Therefore, to assess an efficiency of our construction, we have compared our results with those which can be obtained using this more detailed information contained in the "gold record" and using the same theoretical methods.

The paper is organized as follows. Firstly (chapter 2), we present our proposal of classification of farms with its justification. The chapter 3 contains a list of variables used to determine the classes of farms. Most of them are constructed especially for our investigation using the available information in the "master record" file. We explain exactly the methods of their computation. Next (chapter 4) the classification algorithm and method of endogenous optimization of criterion intervals is described. The empirical results of analysis and computation are given in chapter 5. Finally (chapter 6), main conclusions are formulated.

2. Main assumptions and proposal of a typology

The final form of the classification has been elaborated on the basis of many consultations with experts dealing with agricultural statistics. It was agreed that the classification should take into account three essential aspects of the analyzed problem. Firstly, in each of the size groups some farms conducting no agricultural activity can occur. Secondly, among farms conducting such activity it is worth to select these which specialize mainly in crop production and these in which animal production is prevailing direction of their activity. Theoretically, this division can be executable on the basis of results of Farm Structure Survey (FSS), but it is mainly sample (10% sample of individual farms). Because of this, it was assumed that the basis of full classification prepared before the census will be data coming from administrative sources. Of course, the scope of information available there is smaller that could be obtained from direct survey (such as FSS). This fact should be reflected in the main division.

Finally, we propose the following typology of farms:

1) Farms with the agricultural land area above 2 ha:
    a. Farms conducting productive agricultural activity with prevalence of the crop output (it will be denoted further as C1A),
    b. Farms conducting productive agricultural activity with prevalence of the animal output (C1B),
    c. Farms which conduct agricultural activity and conduct neither crop nor animal production, but maintain the agricultural land in the good agricultural culture (C1C)
    d. Farms conducting no agricultural activity (C1D).
2) Farms with the agricultural land area between 1 ha and 2 ha:
    a. Farms conducting productive agricultural activity with prevalence of the crop output (it will be denoted further as C2A),
    b. Farms conducting productive agricultural activity with prevalence of the animal output (C2B),
    c. Farms, which conduct agricultural activity and conduct neither crop nor animal production, but maintain the agricultural land in the good agricultural culture (C2C)
    d. Farms conducting no agricultural activity (C2D).
3) Farms with the agricultural land area below 1 ha:
    a. Farms conducting productive agricultural activity with prevalence of the crop output (it will be denoted further as C3A),

b. Farms conducting productive agricultural activity with prevalence of the animal output (C3B),
c. Farms, which conduct agricultural activity and conduct neither crop nor animal production, but maintain the agricultural land in the good agricultural culture (C3C)
d. Farms conducting no agricultural activity (C3D).

This proposal has a introductory character. That is, when more data will be available it will be further developed by adding new subclasses defined by a nominal criterion based on values of a new variable. For example, if due to some local circumstances we would like to select within the C1A class, farms planting the flax, we should find all farms belonging to C1A, for which the sown area of flax is greater than 0.

3. Classification variables

Our approach is based on variables measured on interval or ratio scale. A restriction only to the nominal variables results in significant restriction of information on the size and structure of a given phenomenon, which could be quite obvious in practice. Therefore, the interval or ratio variables should be preferred. But, as we stated in the previous paragraph, the nominal variables could be used rather to create more detailed classification levels, what seems also to be easy for potential users of the classification. This assumption is, however, followed by many additional problems. For example, we have to decide, which farms should be classified to the group of farm where area of agricultural land used to the crop production is very small in relation to their total land area. It is also worth noting that in the case of interval or ratio variables, adherence of an object to a given class is usually expressed by some tolerance set of values (understood, in general, as a real interval). The collection of presently used variables satisfies these postulates.

Taking into account the scope of information available in administrative sources, we have proposed the following set of classification variables:

1) Total agricultural land area in ha (denoted further as *Land*)
2) Coefficient of intensity of crop production (in %) (*Crop*)
It is defined as

$$Crop = \begin{cases} \frac{grc}{pzw+grc} \text{ if } pzw + grc \neq 0, \\ 0 \text{ if } pzw + grc = 0, \end{cases}$$

where *grc* is the area of land under the agricultural activity used to the crop production in a farm, and *pzw* denotes the size of stocks in farms raising animals recalculated into main forage area.

3) Coefficient of intensity of animal production (in %) (*Animal*)
It is defined as

$$Animal = \begin{cases} \frac{pzw}{pzw+grc} \text{ if } pzw + grc \neq 0, \\ 0 \text{ if } pzw + grc = 0, \end{cases}$$

where *grc* and *pzw* are as above.

4) Share of agricultural land maintained in good agricultural culture in the total land area of a farm (in %) (*Culture*)
5) Share of meadows and pastures in the total land area of a farm (%) (*Meadows*).

The *grc* will be computed as a sum of area of land under particular crops or used for agricultural production in another way (e.g. as orchards, tree and bush nurseries, fixed crops under cover – such as mushrooms, etc.).

The quantity *pzw* is computed as a number of farm animals being in a given farm recalculated per Livestock Units (LU; done by multiplication of number of particular animals by respective coefficients established in relevant EU and domestic authorities regulations – cf. e.g. A. Tonini (2007), A. Tonini and R. Jongeneel (2007) or H. Lipińska and J. Gajda (2006)) and divided by the average population of LU per 1 ha of Main Forage Area (MFA) in agricultural regions. In Poland, 4 large agricultural regions specified from the point of view of natural conditions and potential for agricultural development have been established. They have the following values of the MFA per one LU: 0.76 (Pomorze and Mazury – northern and north–western regions), 1.50 (Wielkopolska and Śląsk – western and south–western parts of Poland), 1.31 (Mazowsze and Podlasie – central and north – eastern regions) and 0.90 (Małopolska and Pogórze – southern part of the country). In the trial census each agricultural region was represented by one gmina.

The coefficients *Crop* and *Animal* show which type of production has the main importance in the agriculture. If one of them is greater than 50% then the latter must be smaller than 50%. If one of them amounts to zero and the latter does not, then the farm is regarded to be concentrated only on the production of type represented by non–zero index. If no production is conducted then both indices are equal to zero.

This set of variables is, of course, not ideal. Due to practical reasons, it had to be based, however, only on the database assumed to be collected for all farms directly before the main census (from administrative sources). As mentioned in Chapter 5, it was of non–satisfactory quality due to lack of a harmonization of various registers. Our trail to improve the quality of the "master record" has not, of course, eliminated all inconveniences, but minimized only most serious of them. On the other hand, it is commonly regarded that the specialization of production could be better assessed using the structure of marketable output or standard gross margin. These data are available only from The Farm Structure Survey (based only on a relatively small – 10% – sample of individual farms) but not from the administrative sources and therefore they cannot be used for the census purposes.

When using traditional classification method for the internal structure of a set of the farms described by a number of variables, the orders of magnitude of these variables need to be standardized to retain the uniform influence of the individual variables on the calculated distances. In our case, all the variables considered are "by definition" normalized on [0,1] (or, more precisely, on [0%, 100%]) because their values are presented in %. Therefore, no additional normalization seems to be necessary (as the basic characteristics of their distributions would remain practically unchanged after such transformation).

4. Classification algorithm

There exist many various methods of cluster analysis. Most of them generates, however, the classes in an endogenous way (e.g. by representatives, centroids or optimal thresholds of similarity), being exclusively a result of performance of the clustering process and properties of the model (cf. B. S. Everitt et al. (2001)). One can obtain then high–quality clusters, which are usually rather hard to interpret. In the analyzed case we have the classes being established arbitrarily due to some external circumstances (e.g. expectations of users). Therefore, the criterion of appurtenance also should be fixed "in advance". Due to variety of measurement methods and statistical properties of analyzed variables, the most reasonable solution seems to be a unique characterization of particular classes by intervals reflecting scopes of required, or typical realization of variables for farms belonging to these classes.

The classification method can be described as follows. Let $n \in \mathbb{N}$ denotes the number of objects (farms, in this case), and $m \in \mathbb{N}$ – the number of features (variables) characterizing these objects. Thus, we have at disposal m features $X_1, X_2, \dots, X_m$. Denote by x_{ij} a value of the feature X_j for *i*–th object, $i = 1,2,\dots,n$, $j = 1,2,\dots,m$. Due to the properties of our specific model we assume that all observations are nonnegative. The set of all analyzed objects will be denoted as Γ. Each object belonging to Γ is uniquely represented by the vector $\gamma_i = (x_{i1}, x_{i2}, \dots, x_{im}) \in \mathbb{R}^m$. Assume that $k \in \mathbb{N}$, $1 \leq k \leq n$ is the fixed number of typological classes which the set Γ we would like to divide into. Our purpose is then to obtain a sequence of subsets $\Omega_r \subseteq \Gamma$, $r = 1,2,\dots,k$, such that $\Omega_r \cap \Omega_q = \emptyset$ for every $r, q = 1,2,\dots,k$, $r \neq q$ and $\bigcup_{r=1}^{k} \Omega_r = \Gamma$.

Each of the k proposed typological classes has to be described by unique criterions for allocation of a given object to it. For the feature X_j we determine then k_j ($k_j \in \mathbb{N}$, $2 \leq k_j \leq k$) of criterion intervals $\lambda_{1j}, \lambda_{2j}, \dots, \lambda_{k_j j}$, such that $\bigcup_{q=1}^{k_j} \lambda_{qj} = \mathbb{R}_+ \cup \{0\} = [0, \infty)$, $j = 1,2,\dots,m$. The intervals are desired to be disjoint, although it is not absolutely necessary. According to these conditions, the interval λ_{rj} is of the form $\lambda_{qj} = [a_{qj}, b_{qj}) \subseteq \mathbb{R}_+$, where $a_{qj} < b_{qj}$, $q = 1,2,\dots,k_j - 1$, $j = 1,2,\dots,m$. Assume that any class Ω_r is determined by an interval vector $\Phi_r = (\varphi_{r1}, \varphi_{r2}, \dots, \varphi_{rm})$, where $\varphi_{rj} = [\alpha_{rj}, \beta_{rj}) \subseteq \mathbb{R}_+$ is the interval belonging to the set of criterion intervals of a given feature, i.e. $\varphi_{rj} \in \{\lambda_{1j}, \lambda_{2j}, \dots, \lambda_{k_j j}\}$ (and therefore $\alpha_{rj} = a_{qj}$ and $\beta_{rj} = b_{qj}$ for some $q \in \{1,2,\dots,k_j\}$), and selected to establishment of a criterion characterizing this class, $r = 1,2,\dots,k, j = 1,2,\dots,m$.

For a better precision of further analysis, we have now to introduce a definition of a distance of a real number y from the interval $U = [u_1, u_2] \subseteq \mathbb{R}$, $u_1 \leq u_2$. We do this using the formula:

$$\delta(y, U) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } y \in U, \\ \min(|y - u_1|, |y - u_2|) & \text{if } y \notin U. \end{cases} \qquad (1)$$

Note that the definition (1) has a sense also if one of the limits of the interval U is infinite. In such case, if $y \notin U$, we assume as a distance the absolute value of a difference between y and the finite limit of U.

An aggregated distance of *i*–th object (represented by the vector γ_i) from the *r*–th typological class Ω_r described by the criterions Φ_r is defined to be a maximum of partial distances from particular criterion intervals, i.e.

$$d(\gamma_i, \Phi_r) \stackrel{\text{def}}{=} \max_{j=1,2,\dots,m} \delta(x_{ij}, \varphi_{rj}) \qquad (2)$$

$r = 1,2,\dots,k$, $i = 1,2,\dots,n$. That is, the distance from an object to a class is computed by calculation of the distances of data for variables describing a given object from respective intervals describing the class (expressed by (1)) and next a determination of maximum of them. This choice enables one to avoid a compensation of discrepancy in respect to some criterion by a similarity connected with other criterion, what is unfavorable from the point of view of the classification.

The purpose of our analysis is to determine optimum probabilities of assignment of an object represented by the vector γ_i to typological classes determined by the criterions $\Phi_1, \Phi_2, \dots, \Phi_k$. In this context, the key postulate is that this probability should be reversely proportional to a distance of the object from the given class. Therefore, it is proposed to apply the model of so–called probabilistic d–clustering (cf. e.g. A. Ben–Israel and C. Iyigun (2008)). It belongs to the tools of fuzzy classification. In the investigated case we would like to find numbers $p_k(\gamma_i)$, $r = 1,2,\dots,k$, $i = 1,2,\dots,n$, which minimize the value of the target function:

$$f\big(p_1(\gamma_i), p_2(\gamma_i), \dots, p_k(\gamma_i)\big) = \sum_{i=1}^{n} \sum_{r=1}^{k} d(\gamma_i, \Phi_r)\, p_r^2(\gamma_i) \quad (3)$$

with the conditions:

$$\sum_{r=1}^{k} p_r(\gamma_i) = 1$$

$$p_r(\gamma_i) \geq 0$$

for every $r = 1,2,\dots,k$, $i = 1,2,\dots,n$.

The results presented in the cited article can be applied also in this case. Assuming reasonable requirement that $p_r(\gamma_i)\, d(\gamma_i, \Phi_r) = \text{const.}$ (depending on γ_i) for every $r = 1,2,\dots,k$, $i = 1,2,\dots,n$, the optimum probability of assignment of an object represented by the vector γ_i to the class Ω_r described by Φ_r, is given by the formula:

$$p_r^*(\gamma_i) = \frac{\prod_{\substack{q=1,2,\dots,k \\ q \neq r}} d(\gamma_i, \Phi_q)}{\sum_{q=1}^{k} \prod_{\substack{u=1,2,\dots,k \\ u \neq q}} d(\gamma_i, \Phi_u)}, \tag{4}$$

for every $r = 1,2,\dots,k$, $i = 1,2,\dots,n$. The object represented by the vector γ_i is assigned to such class for which the probability of assignment expressed by (4) is the greatest. When the optimal probabilities (4) for two or more classes are identical, then the classification of the object will be determined by maximum partial distance from the particular unit criterion. As we can conclude from the formula (4), a significantly important problem is such choice of limits of criterion intervals that enables to exclude a possibility of an occurrence of zero distances of an object from two or more different classes (what could result in the zero value of the denominator in (4)). Unlike many other classification algorithms, this approach is non-iterative, because the algorithm of assignment and – by the same token – the optimum class for a given farm is exactly computed using the formula (4) derived by mathematical methods. An iteration will be used to obtain the theoretical optimum classes, as we will describe in the next part of this chapter.

Some assessment of quality of obtained classification one could obtain by determination of k interval criterion vectors by an iterative algorithm, originally proposed using some ideas coming from papers by A. Ben–Israel and C. Iyigun (2008) and C. Iyigun (2007) and being a development of some concepts suggested by E. Weiszfeld (1937), adopted to our specific situation. To obtain an optimal criterion division $\Phi_1^*, \Phi_2^*, \dots, \Phi_k^*$ for the exercise (3), we have to consider the problem of differentiation of a function defined on a set of closed intervals contained in a real line. Let $\mathbb{IR} \overset{\text{def}}{=} \{[a,b]\colon a,b \in \mathbb{R}, a \leq b\}$ be this set. Of course, $\mathbb{R} \subseteq \mathbb{IR}$, because a real number is an interval itself (although of a thin form, i.e. with equal limits). Let $Y = [y_1, y_2] \in \mathbb{IR}$ will be non–trivial, i.e. $y_1 < y_2$. Consider the function $g\colon \mathbb{IR} \longrightarrow \mathbb{R}$ defined on its whole domain. Let h be some real number such that $\xi_{Y,h} \overset{\text{def}}{=} [y_1 + h, y_2 - h] \in \mathbb{IR}$. A lower and upper derivation of the function g at its argument Y are defined respectively as:

$$\underline{g'(Y)} = \lim_{h \to 0^-} \frac{g(\xi_{Y,h}) - g(Y)}{h} \text{ and } \overline{g'(Y)} = \lim_{h \to 0^+} \frac{g(\xi_{Y,h}) - g(Y)}{h}.$$

If $\underline{g'(Y_0)} = \overline{g'(Y_0)}$, then we say that the function g is differentiable at the argument Y_0, and its derivation at this argument will be denoted as $g'(Y_0)$ or $\left.\frac{\partial g}{\partial Y}\right|_{Y_0}$.

Let us come back to our distance (1), which – investigated as a function of intervals U – belongs to the analyzed family of interval functions. For its argument $V = [v_1, v_2] \in \mathbb{IR}$, we have then:

$$\left.\frac{\partial\delta(x,U)}{\partial U}\right|_V = \begin{cases} 0 \text{ if } (v_1, v_2) \ni x, \\ 1 \text{ if } v_2 < x \text{ or } v_1 > x. \end{cases}$$

Intervals of the form $[a, x], [x, b] \in \mathbb{IR}$ are places where the function δ is not differentiable. In both cases the lower derivation in such place equals to 0 but the upper amounts to 1.

Taking these observations into account, we can determine theoretical optimum criterion intervals. Our main purpose is to determine such limits of these intervals that the value of the function (3) was minimum. For any class Ω_r, $r = 1,2,\dots,k$, we consider two cases:

Case 1. The class Ω_r ($r \in \{1,2,\dots,k\}$) has such property that no object is strictly identifiable as belonging to it, i.e. $d(\gamma_i, \Phi_r) > 0$ for every $i = 1,2,\dots,n$. Then, the solution of these problems is to find such interval arguments for which the gradient of the function (3) restricted to this class amounts to zero. More formally, we would like for every $j = 1,2,\dots,m$ to find such intervals $\varphi_{rj} = [\alpha_{rj}, \beta_{rj}] \in \mathbb{IR}$, that

$$\frac{1}{m}\sum_{i=1}^{n} \frac{\partial\delta(x_{ij},\varphi_{rj})}{\partial\varphi_{rj}} \frac{\delta(x_{ij},\varphi_{rj})}{d(\gamma_i,\Phi_r)} p_r^2(\gamma_i) = 0. \qquad (5)$$

Taking into account our conclusions about differentiation of the function δ, the equality (5) holds if and only if

$$\sum_{\substack{i=1,2,\dots,n: \\ x_{ij}<\alpha_{rj}}} \frac{(\alpha_{rj} - x_{ij})p_r^2(\gamma_i)}{d(\gamma_i,\Phi_r)} + \sum_{\substack{i=1,2,\dots,n: \\ x_{ij}>\beta_{rj}}} \frac{(x_{ij} - \beta_{rj})p_r^2(\gamma_i)}{d(\gamma_i,\Phi_r)} = 0 \qquad (6)$$

for every $j = 1,2,\dots,m$.

Because both components of the sum on left–hand side of (6) are nonnegative, then the equality (6) holds only if each of them equals to zero. After relevant transformations we obtain the optimum limits of intervals of the form:

$$\alpha_{rj}^* = \sum_{\substack{i=1,2,\dots,n, \\ x_{ij}<\alpha_{rj}^*}} \frac{\frac{p_r^{*2}(\gamma_i)}{d(\gamma_i,\Phi_r^*)} x_{ij}}{\sum_{\substack{h=1,2,\dots,n, \\ x_{hj}<\alpha_{rj}^*}} \frac{p_r^{*2}(\gamma_h)}{d(\gamma_h,\Phi_r^*)}}, \qquad (7a)$$

and

$$\beta^*_{rj} = \sum_{\substack{i=1,2,\dots,n,\\ x_{ij}>\beta^*_{rj}}} \frac{\frac{p_r^{*2}(\gamma_i)}{d(\gamma_i,\Phi^*_r)} x_{ij}}{\sum_{\substack{h=1,2,\dots,n,\\ x_{hj}>\beta^*_{rj}}} \frac{p_r^{*2}(\gamma_h)}{d(\gamma_h,\Phi^*_r)}}, \tag{7b}$$

$r = 1,2,\dots,k,\ j = 1,2,\dots,m.$

Case 2. Let Ω_r ($r \in \{1,2,\dots,k\}$) be a class such that there exist objects strictly identifiable as belonging to it, i.e. for some $i \in \{1,2,\dots,n\}$ we have $d(\gamma_i,\Phi_r) = 0$. Let Ξ_r be a set of objects strictly identifiable as belonging to the class Ω_r. Then the approximation of limits of optimum intervals can be obtained respectively as minimum and maximum values of respective features, i.e.

$$\alpha^*_{rj} = \min_{i:\in\Xi_r} x_{ij} \text{ and } \beta^*_{rj} = \max_{i\in\Xi_r} x_{ij} \tag{8}$$

for every $j = 1,2,\dots,m$.

The algorithm is now iterative. We start from arbitrarily fixed criterion intervals; the optimum probabilities (4) of appurtenance of objects to them are determined. Next, using the formulas (7a), (7b) or (8) and inserting to their right–hand sides all estimated values, we obtain first iteration of the optimum classes. Next, using them, we perform the second iteration and so on. We stop the procedure, when the distance between criterion structures of two successive iterations will be smaller than an arbitrarily established positive threshold ε. The distance of two criterion structures is calculated using the formula (where $\mathbf{\Phi} = (\Phi_1,\Phi_2,\dots,\Phi_k)$, $\mathbf{\Phi}' = (\Phi'_1,\Phi'_2,\dots,\Phi'_k)$ is assumed):

$$d_\#(\mathbf{\Phi},\mathbf{\Phi}') = \frac{1}{k}\sum_{r=1}^{k}\sqrt{\frac{1}{m}\sum_{j=1}^{m} d^2_{\mathcal{H}}(\varphi_{rj},\varphi'_{rj})},$$

where $d_{\mathcal{H}}$ is the Hausdorff distance between respective intervals. The Hausdorff distance between two intervals $U = [u_1,u_2], W = [w_1,w_2] \subseteq \mathbb{R}$, $u_1 < u_2$, $w_1 < w_2$, is defined as:

$$d_{\mathcal{H}}(U,W) = \max(|w_1 - u_1|,|w_2 - u_2|).$$

Therefore, the iteration is continued until $d_\#(\mathbf{\Phi},\mathbf{\Phi}') \le \varepsilon$, where $\mathbf{\Phi}$, and $\mathbf{\Phi}'$ denote structures obtained in two subsequent iterations. Of course, the optimum collection of criterion intervals can contain for some variables also non–disjoint intervals.

It is worth noting that this method differs significantly from other well–known fuzzy classification algorithms. All of them are based on the following objective function.

$$f(p_1(\gamma_i), p_2(\gamma_i), \ldots, p_k(\gamma_i), \bar{\varphi}_r) = \sum_{i=1}^{n}\sum_{r=1}^{k} d^2(\gamma_i, \bar{\varphi}_r)\, p_r^q(\gamma_i), \qquad (9)$$

where $\bar{\varphi}_r$ is the centroid of the group Φ_r, $q \in \mathbb{N}$ is fixed and $d^2(\gamma_i, \bar{\varphi}_r)$ is the distance of the object Γ_i from the centroid of respective group Φ_r ($i = 1,2,\ldots,n$, $r = 1,2,\ldots,k$), is defined in various ways. J. C. Bezdek (1973) and R. J. Hathaway and J. C. Bezdek (1988) define it to be the Euclidean norm on $\mathbb{R}^m$, D. E. Gustafson and W. C. Kessel (1979) as a modified Mahalanobis distance with preserved volume, in Gath–Geva approach (I. Gath and A. B. Geva (1989)) the distance is defined using the posterior probability function assuming that the normal distribution with expected variance and covariance matrix is chosen for generating a datum with prior distribution. Finally, the FCM – NM algorithm (J. – M. Yih and S. – F. Huangh (2010)) is based on the normalized Mahalanobis distance. All these algorithms are iterative and belong to the c–means clustering "family", i.e. they consist of iterations starting either with an initial guess for partitioning on prototype (centroid) vectors $\bar{\varphi}_r$ and is continued until the distances between two successive iterations are sufficiently small. That is, iteration stops with the first $\Phi^{(u)}$ such that $\|\Phi^{(u)} - \Phi^{(u-1)}\| < \varepsilon$, where ε is the arbitrarily established threshold of accuracy, $\Phi^{(u)}$ is the partition obtained at the u–th step, $u = 1,2,\ldots$. Each of these concepts has its disadvantages: the Bezdek's method tends to create spherical clusters, in the Gustafson–Kessel method the added fuzzy covariance matrices in their distance measure are not directly described, in the Gath–Geva algorithm the assumption that the data are multivariate normally distributed can be inappropriate in practice. And, finally, the FCM–NM proposal deforms the original variation of the diagnostic variables.

Our method, although belonging to the fuzzy clustering tools (compare the objective functions (3) and (9)), seems to be much more practically useful. The typological classes are usually defined using the reference (or tolerance) intervals for particular variables and it satisfies this postulate. Moreover, the optimization is very simple and enables to compare practical criterions with their artificial but theoretically optimum equivalences with no significant influence of some inconvenient aspects, e.g. sphericality.

Here one can also compare the form of membership matrix (see for example formula (2) in I. Gath and A. B. Geva (1989)) and the probability of assignment (see formula (4) in our work), what reveals that some ideas are common in both approaches, but particular implementation of the algorithms may be different and sometimes leads to different results. More detailed analysis related to such comparison may be done in the future.

5. Results of classification

According to our assumptions, to perform the classification, one should define effective criterion intervals. They were established using main requirements concerning particular class or, if they are not specified, as typical observation for respective groups of farms occurring in other, but similar, statistical surveys (such as FSS) conducted in the near past. Our final choice is presented in Table 1. The upper element of each cell denotes the lower limit of the respective interval, and the lower one – its upper limit.

Table 1. Arbitrarily assumed criterion intervals

Variable	C1A	C1B	C1C	C1D	C2A	C2B	C2C	C2D	C3A	C3B	C3C	C3D
Land	2 10000	2 10000	2 10000	2 10000	1 1.999	1 1.999	1 1.999	1 1.999	0 0.999	0 0.999	0 0.999	0 0.999
Crop	50 100	0 49.9999	0 0	0 0	50 100	0 49.9999	0 0	0 0	50 100	0 49.9999	0 0	0 10
Animal	0 49.9999	50 100	0 0	0 0	0 49.9998	50 100	0 0	0 0	0 49.9999	50 100	0 0	0 0
Culture	50 100	0 49.9999	0.00001 100	0 0	50 100	0 49.9999	0.00001 100	0 0	50 100	0 49.9999	0.00001 100	0 0
Meadows	10 39.9999	40 100	0 9.9999	0 0	10 39.9999	40 100	0 9.9999	0 0	10 39.9999	40 100	0 9.9999	0 0

Source: Authors' elaboration.

The 'master record' is a file being a compilation of data from several various administrative sources, such as Tax Register of Real Estates or database maintained by the Agency for Restructuring and Modernization of Agriculture. Due to significant differences between these sources in terms of timeliness and scope of information, many contradictions within the data could be observed. The most important of them are:

- for 493 records the area of agricultural land maintained in good agricultural culture is larger than the total area of agricultural land,
- for 2 other records the area of meadows and pastures is larger than the total area of agricultural land,
- for other 112 records the area of arable land is larger than the total area of agricultural land,
- area of meadows and pastures is positive, but the area of agricultural land maintained in good agricultural culture amounts to 0 (next 790 records).

Summarizing, 1397 records (i.e. 37.7%) were defective and had to be removed from further analysis. Moreover, the 'master record' contains no data on neither sown area nor structure of the **basic** crops. One can find there only information on some "peripheral" crops, i.e. flax, hemp and hop. Therefore, we cannot also indicate farms where the agricultural production exceeds the thresholds adopted within EU. Moreover, the Agency for Restructuring and Modernization of Agriculture does not register the farms for which agricultural land area is smaller than 1 ha. These problems (taking into account also the fact that the Agency gathers no data on the total agricultural land, but only on the land maintained in the good agricultural culture) are the main difficulties in harmonization of the analyzed registers.

Due to lack of data on crops, a direct computation of the quantity *grc* necessary to determine the classification variables *Crop* and *Animal* is impossible. Therefore, estimation is needed. We have done it using the relevant data gathered during the Farm Structure Survey in 2007. That is, we have constructed a linear regression model with *grc* as explained variable and arable land area (*aland*) as explanatory variable. The regression function is of the form:

$$grc = 0.98455 \cdot aland - 0.03105. \qquad (10)$$

The value of the Student's t–test for intercept amounts to -3.02 (p=0.0025) and for the slope 5446.25 (p<0.0001). The analysis of variance and adjustment is presented in Table 2.

Table 2. Regression of *grc* according to arable land area – analysis of variance and assessment of adjustment

Analysis of variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	502043681	502043681	2.966E7	<.0001
Error	185159	3133940	16.9257		
Corrected Total	185160	505177621			

Adjustment of the model

Root MSE	4.1141	R-Square	0.9938
Dependent Mean	20.6096	Adj R-Sq	0.9938
Coeff Var	19.9620		

Source: Authors' elaboration using the SAS Enterprise Guide 4.2 environment.

On the basis of these results we can conclude that this model is well established and may be an effective tool for estimation. We have used it. If an estimate of *grc* obtained on the basis of regression function was negative (it is sometimes possible for very small farms), we have assumed it to be zero. The function (10) was finally used to estimate the variables *Crop* and *Animals*.

Now, we present the classification obtained using the "master record" set. The Table 3 contains specification of number of farms belonging to each class and average value of particular classification variables within it. To avoid some misclassification which is undesirable from the practical point of view, during the maximization of the probabilities (4) (with the distance of object from classes computed using (2)) , we have preferred those elements which are more strictly desired from the practical point of view. That is, we have minimized (4) only within the farms of the same size type (in terms of *Land* variable) as the classified object. That is, we have looked for such optimum class, for which the land area of the given farm belong to the respective land area interval describing this class. If necessary, this additional criterion was extended also to the variables *Crop* or *Animal*.

The quality of received clustering was assessed using three indices. The coefficient of homogeneity of clusters is given as

$$hm = \frac{1}{k}\sum_{r=1}^{k}\frac{1}{n_r}\sum_{\substack{i=\{1,2,\dots,n\}\\ \gamma_i\in\Omega_r}} d_e(\gamma_i,\bar{\gamma}_r),$$

and the coefficient of their heterogeneity, i.e. mutual separation level (assuming that k>1):

$$ht = \frac{1}{k(k-1)} \sum_{r=1}^{k} \sum_{\substack{s=1 \\ s \neq r}}^{k} d_e(\bar{\gamma}_r, \bar{\gamma}_s),$$

where $\bar{\gamma}_r$ is the centroid of the class Ω_r, i.e. the vector, which coordinates are arithmetic means of observations of respective variables for objects belonging to this class, r = 1, 2, …, k, and $d_e(\cdot,\cdot)$ denotes the Euclidean distance. The coefficient of correctness of clusters is a ratio of these two quantities (i.e. it equals to hm/ht). The closer to zero it is, the better the quality of clustering is.

Table 3. Classification of farms using the 'master record' data

Class	Number of farms in the class	Land	Crop	Animal	Culture	Meadows
C1A	911	16.2812	84.9414	15.0586	71.9838	20.1087
C1B	327	17.5099	33.2624	66.7376	87.6120	28.7586
C1C	17	6.1618	0	0	84.5360	0
C1D	31	9.4671	0	0	0	0
C2A	156	1.4583	98.9344	1.0656	45.2971	12.2058
C2B	13	1.5741	32.2028	67.7972	80.6275	32.5879
C2C	4	1.4850	0	0	90.0211	0
C2D	8	1.5625	0	0	0	0
C3A	190	0.4109	99.6811	0.3190	0	0
C3B	35	0.0479	1.5527	98.4473	0	0
C3C	0	0	0	0	0	0
C3D	438	0.0099	0	0	0	0

Source: Authors' elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

The coefficient of homogeneity amounts to 16.014, the coefficient of heterogeneity of clusters equals to 91.650 and hence the coefficient of correctness amounts to 0.1747. It is a very satisfactory result and therefore the division can be perceived as effective.

Using the procedure described in paragraph 4 we have determined the limits of classes in an econometrically optimum division. We have then obtained 11 non–trivial classes, which are uniquely described by the following intervals (for details, see Table 4).

Table 4. Optimum classes for the “master record”

Variable	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Land	2	2	2.020	1.030	0.490	1.100
	135.760	27.270	46.810	1.995	2.390	1.740
Crop	50.191	0	0	51.916	0	0
	100	0	0	100	0	0
Animal	0	0	0	0	0	0
	49.810	0	0	48.084	0	0
Culture	50.969	23.579	0	56.853	0	84.118
	100	100	0	100	85.057	100
Meadows	10	0	0	12.817	0	0
	39.877	0	0	39.288	0	0

Variable	Class 7	Class 8	Class 9	Class 10	Class 11
Land	1.010	1.320	0	0	0
	1.860	1.530	1.290	1.750	0.990
Crop	0	33.887	0	0	0
	0	35.401	0	72.930	0
Animal	0	64.599	0	0	0
	0	66.113	0	58.569	0
Culture	0	84.967	0	0	0
	0	86.364	0	0	0
Meadows	0	53.788	0	0	0
	0	57.516	0	0	0

Source: Authors’ elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

These classes reflect better the actual structure within the analyzed database and seem to be rather easy to interpret. They are also similar to those presented in the Table 1. The one slightly more significant difference between them is that no optimum class is described by values of the *Animal* being above 67% and *Crop* below 33% (except for zero). Also values of the *Meadows* belonging to the interval (0, 10) were also omitted. The main probable reason of this phenomenon could be a fact that for most of farms with prevalence of animal production its domination over the crop production is not especially significant. On the other hand, the distance of the variable *Animal* from the interval (67,100] can be smaller than, e.g. a distance of *Meadows* from the thin interval [0,0]. For example, if for some farm the values of the classification variables are as follows: *Land*=1.50, *Crop*=25%, *Animal*=75%, *Culture*=85% and *Meadows* = 20%, then it will be classified to the class 8. Such situations affect the final results. The nature of such behavior can be explained also taking into consideration that there are relatively small numbers of farms with prevalence of animal production (what is evident for example for C2B and C2C classes), what – with some similarities for variables other than crop and animal in C1 and C2 classes – may cause that such farms were omitted in final classification. An advantage of these classes is the clear presentation of farms of various size and type which have conducted no agricultural production (classes 2, 3, 5, 6, 9, 11). It is confirmed by the comparative classification done using the new classes –1, 4, 8 or 10 with the average values of *Crop* and *Animal* amounting to, respectively, 86.88% and 15.12%, 99.14% and 0.86%, 34.64% and 65.36% and, finally, 22.61% and 17.39%.

For a better comparison, we will present now results of classification using the data collected "from nature", i.e. by direct interviewing the farmers (the "gold record" file). These data are, of course, much more detailed and therefore some classification variables computed using them are of higher quality than those determined using the "master record" database. The earlier established collection of criterion intervals (Table 1) remains without any change. The classification is presented in Table 5.

Table 5. Classification of farms on the basis of the 'gold record' (restricted to farms considered in Table 3.)

Class	Number of farms in the class	Land	Crop	Animal	Culture	Meadows
C1A	1021	12.6959	78.1635	21.8365	97.1076	27.1654
C1B	281	14.9352	40.4804	59.5196	97.9527	29.9669
C1C	19	11.8884	0	0	98.6842	0
C1D	25	4.1644	0	0	0	0
C2A	208	1.4361	87.1073	12.8927	96.7985	27.4768
C2B	30	1.4993	30.3071	69.6929	91.9563	21.9160
C2C	16	1.5250	0	0	94.9925	0
C2D	24	1.4688	0	0	0	0
C3A	138	0.4288	92.7006	7.2994	96.6787	13.5861
C3B	55	0.3685	24.8366	75.1634	83.2707	10.7372
C3C	12	0.5250	0	0	100	0
C3D	225	0.0547	0	0	0	0

Source: Authors' elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

The coefficient of homogeneity amounts to 20.4302, the coefficient of heterogeneity of clusters equals to 81.5134 and hence the coefficient of correctness amounts to 0.2417. It is a very good result. Comparing this structure of classification with the result obtained on the basis of the 'master record' and presented in Table 3, using the three most popular tests for location (i.e. for the hypothesis that the expected value of the distance between them equals zero), we can observe that they are consistent – Student's t statistics amounts to 0. 283698 (p=0.7819), sign test statistics equals to -2 (p=0.3877) and Wilcoxon signed rank statistics is also negative: -7.5 (p=0.5801). The calculations were conducted by UNIVARIATE SAS procedure using the difference between number of farms from Table 3 and Table 5. More details about this procedure can be found in Base SAS 9.2 Procedures Guide (2010), pp. 332–334. This consistency may be sometimes, however, not especially strong due to a fact that some data used to compute *grc* and *pzw* were much more detailed in 'gold record' (e.g. the additional categories of cattle for which special coefficients to calculate them per livestock units are used, were here presented).

Now, we present the limits of classes in an econometrically optimum division established using the procedure described in paragraph 4 and all records included in the 'gold record'. This way, we obtain 11 non–trivial classes, which are uniquely described by the following intervals (see Table 6).

Table 6. Optimum classes for the "gold record"

Variable	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
Land	2	2.070	2	1	0.590	1
	922.660	141.230	11.080	1.990	2.490	1.930
Crop	50.064	0	0	50.171	0	0
	100	0	0	100	0	0
Animal	0	0	0	0	0	0
	49.936	0	0	49.829	0	0
Culture	50.523	35	0	60	0	50
	100	100	0	100	75	100
Meadows	10.062	0	0	11.333	0	0
	39.933	0	0	38.418	0	0

Variable	Class 7	Class 8	Class 9	Class 10	Class 11
Land	1	0.420	0	0.120	0
	1.980	0.990	1.380	0.990	0.960
Crop	0	51.123	0	0	0
	0	100	0	0	0
Animal	0	0	0	0	0
	0	48.877	0	0	0
Culture	0	50	0	50	0
	0	100	75	100	0
Meadows	0	10.101	0	0	0
	0	34.884	0	0	0

Source: Authors' elaboration using original procedure written in the SAS Enterprise Guide 4.2 environment.

These classes are much more consistent with the structure of the analyzed data set than the arbitrarily fixed norms expressed in Table 1. The optimum classes are rather easy to interpret and correspond to the classes presented in Table 1. The only significant difference between both structures is that no class is described by the interval (50, 100] for the variable *Animal* and the interval (0,50] for *Crop*. The reason of this phenomenon is similar as in the case of the optimum classes for 'master record' but the situation observed here is slightly more difficult. The farms which have conducted no agricultural production are also well presented (classes 1, 3, 5, 6, 9, 10, 11). It is confirmed by the comparative classification done using the new classes – the farms conducting the agricultural production were classified only to the classes 1, 4 or 8, with the average values of *Crop* and *Animal* amounting to, respectively, 76.91% and 23.08%, 86.06% and 13.91% and, finally, 89.72% and 10.28%.

6. Conclusions

We have proposed an original method of classification of objects taking various characters of the used criterions into account. They may have less or more strict form, it means that they may be less or more fuzzy. Moreover, the projected groups of objects may be desired to have some arbitrarily fixed properties, resulting from commonly (of which legally) adopted norms. On the other hand, they should be, of course, optimum. Our proposal is a trial to satisfy all these expectations and to show how large is the distance between assumptions established "in advance" and obtained endogenically, only on the basis of the internal properties of used data basis.

Of course, the empirically examined collection of agricultural variables is here relatively small. It is a consequence of small scope of information contained in the "master record" file, which may serve as a classification features. In practice, it is possible to obtain much broader set containing many data that describe the character of a farm using "typical" intervals specific for it (e.g. number of poultry or ostrich units, area of fish ponds, area of mushrooms under cover, etc.). Also, sometimes economists expected to include in the analysis also variables characterizing the economic aspects of the farm activity, such as the standard gross margin (reflecting the value of marketable output), economical size (expressed in European Size Units), commodity output or employment in the farm, fulfillment of some production thresholds established by the EU regulations, etc. The proposed methods theoretically enable one to effectively involve all these postulates to the classification and also asses the internal structure of the data basis being at researcher's disposal. However, these economical variables were not available in our database and therefore they could not be used here. Our method gives the opportunity to introduce it to the model if it will be necessary in the future. It solves also the most difficult problem of usage of interval or ratio variables to the classification. In comparison with other fuzzy clustering methods this one is much more useful from the practical point of view, where classes are often defined using the reference intervals for particular characteristics. It is also more effective in context of the computational capacity.

The only inconvenience connected with this approach seems to be the necessity to establish some additional preferences during maximization of probability of appurtenance of object to particular classes. Despite using the "maximum" formula of distance of an object from a given class, the formula for probability measure (see equation (4)) does not exclude a possibility of compensation of discrepancy in respect to some criterion by a similarity connected with other criterion. The strong practical requirements enforce application of such correction.

However, in general, the proposed method can be assessed as useful in realization of important methodological tasks, such as preparatory works for the national censuses. This task may be realized in practice in any exercise of such type in the following way. Using the typology constructed by means of the proposed method, the basic area and profile groups are determined. On the basis of this division a survey methodology can be established. That is, it could be decided which groups of farms should be investigated by exhaustive survey and which by a sample survey (and in this case it can contribute to find an effective sample size). This enables one to rationalize the costs of statistical undertakings and optimize the quality of their results. The more diversified the used data set is, the more effective the final effects of its application should be. The obtained classification can be further developed by selecting in each class some subclasses by adding more nominal criteria, what is much easier than in data collection analyzed above and each user should be rather able to do it.

It could be also a good basis for a wider discussion on principles and efficiency of such classification method as well as on methods of its possible improvement. Of course, a critical view of our results is fully justifiable. The critics of it may recall in this context the argument that the economical quantities such as structure of agricultural output could be here better variables and the final results obtained using them may be different than the current product (e.g. taking into account that according to the Statistical Yearbook of Agriculture 2009 (GUS (2009)), the nationwide ratio of crop output to animal output is about 56:44 for gross output and 45:55 for market output, whereas in our case the respective relation between these two main types of farms was much more clear. One should remember, however, about two main features of our approach. Firstly, we were not able to use the strictly economical variables because they were not available in administrative sources used for the census. Secondly, due to the above reason we had to analyze the **physical** structure of production which could be significantly different from their **economical**, monetary value. Of course, if we had more information on the economical aspects at our disposal, the quality of the classification would be better.

REFERENCES

BEN – ISRAEL A., IYIGUN C. (2008) *Probabilistic d–Clustering*, Journal of Classification, vol. 25, pp. 5–26.

BEZDEK J, C. (1973) *Fuzzy Mathematics in Pattern Classification*, PhD Dissertation, Cornell University, Ithaca, New York.

Commission Regulation (EC) *No 1242/2008 of 8 December 2008 establishing a Community typology for agricultural holdings,* OJ L 335, 13.12.2008, pp. 3–24 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:335:0003:0024:EN:PDF

EVERITT, B. S., LANDAU, S., & LEESE, M. (2001) *Cluster analysis* (4th ed.). London: Arnold.

GATH I., GEVA, A. B. (1989) *Unsupervised optimal fuzzy clustering*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol 11(7), pp 773–781.

GUS(2009) *Statistical Yearbook of Agriculture*, Central Statistical Office of Poland (GUS), Warszawa.

GUSTAFFSON D. E,, KESSEL W. C. (1979) *Fuzzy Clustering with a Fuzzy Covariance Matrix, Clustering with a Fuzzy Covariance matrix*, Proceedings of the IEEE Conference Decision Contribution, San Diego, CA, USA, p. 761–766.

HATHAWAY R. J., BEZDEK J. C. (1988) *Recent Convergence Results for the Fuzzy c-Means Clustering Algorithm*, Journal of Classification, vol. 5, p. 237–247.

IYIGUN C. (2007) *Probabilistic Distance Clustering*, A dissertation submitted to the Graduate School – New Brunswick Rutgers, The State University of New Jersey in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Graduate Program in Operations Research. Written under the direction of Professor Adi Ben–Israel, New Brunswick, New Jersey, November, 2007, http://www.benisrael.net/Iyigun-Thesis-Nov-07.pdf.

LIPIŃSKA H., GAJDA J. (2006) *Area of farms versus fodder base and cattle population in specialized dairy farms*, Annales Universitatis Mariae Curie-Skłodowska Lublin – Polonia, vol. LXI, Sectio E, pp. 225–236 (in Polish).

Regulation (EC) No 1166/2008 of the European Parliament and of the Council of 19 November 2008 on farm structure surveys and the survey on agricultural production methods and repealing Council Regulation (EEC) No 571/88, OJ L 321, 1.12.2008, p. 14–34 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:321:0014:0034:EN:PDF.

SAS Institute Inc. (2010) *Base SAS® 9.2, Procedures Guide: Statistical Procedures*, Third Edition. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/cdl/en/procstat/63104/PDF/default/procstat.pdf.

TONINI A. (2007) *Agriculture and Dairy in Eastern Europe after Transition focused on Poland and Hungary*, PhD Thesis, Wageningen University, The Netherlands, http://library.wur.nl/wda/dissertations/dis4133.pdf.

Tonini A, Jongeneel R. (2007) *The Distribution of Dairy Farm Size in Poland: a Markov Approach Based on Information Theory*. Applied Economics, vol. 1, pp.1–15.

WEISZFELD E. (1937) Sur le point pour lequel les sommes des distances de n points donné et minimum, Tahoku Mathematical Journal, vol. 34, pp. 355–386.

YIH J. – M., HUANGH S. – F. (2010) *Unsupervised Clustering Algorithm Based on Normalized Mahalanobis Distance*, [in:] S. Chen and H. Wu (eds.) Proceedings of the 9th WSEAS Int. Conference on Applied Computer and Applied Computational Science, Electrical and Computed Engineering Series. A Series of Reference Books and Textbooks, WSEAS Press.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 639—647

BOOK REVIEW

National Research Council (2009). *Coverage Measurement in the 2010 Census.* Panel on Correlation Bias and Coverage Measurement in the 2010 Decennial Census, Robert M. Bell and Michael L. Cohen (Eds.). Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press, 180 pages

Methods of data quality assessment of population censuses were discussed at the Committee of Demographic Research of the Polish Academy of Sciences during last three years (Dygaszewicz, 2008; Kordos, 2008, 2010; Nowak, 2008), and presented in several Polish statistical publications (Gołata, 2008, 2009; Kordos, 2007; Paradysz, 2008, 2010). It is mainly connected with preparation for the 2011 Census of Population and Housing in Poland, and increasing concern in quality of statistical data.

Preparation for the US 2010 Census of Population started directly after the 2000 Census (National Research Council, 2000). Among other problems to be solved, the coverage measurement in the 2010 Census has been considered as the most important. The National Research Council appointed a Panel on Coverage Evaluation and Correlation Bias. Findings of the National Research Council's Panel are published in the book *"Coverage Measurement in the 2010 Census"* with Robert. Bell and Michael L. Cohen, as Editors.

This book consists of five chapters, three appendixes, and nearly 100 references and bibliography, including 13 recommendations for the Census Bureau, which is in charge for US Censuses of Population. This book, which details the findings of the National Research Council's Panel on Coverage Evaluation and Correlation Bias, strongly supports the Census Bureau's change of goal. However, the panel finds that the current plans for data collection, data analysis, and data products are still too oriented towards measurement of net coverage error to fully exploit this new focus. Although the Census Bureau has taken several important steps to revise data collection and analysis procedures and data products, this book recommends further steps to enhance the value of coverage measurement for the improvement of future census processes.

Chapter 1

Introduction. This introductory chapter stressed that since the 1950 census there has been efforts by the Census Bureau to estimate the size of error in census counts for areas and demographic groups and to use the information to improve census processes. The programs to measure census coverage error are referred to as coverage measurement programs. In recent years, coverage measurement programs included a third objective—correcting the census for enumeration error, referred to as census adjustment. The techniques used in coverage measurement programs to understand the extent of enumeration errors are sample surveys, dual-systems estimation (DSE), and demographic analysis.

This introductory chapter is followed by four chapters and three appendices:

Chapter 2

Fundamentals of Coverage Measurement. Coverage measurement is a collection of techniques that measure the differences between census enumerations and the corresponding true counts for groups or areas. Coverage measurement is the quantitative aspect of coverage evaluation, which also encompasses more qualitative techniques, such as ethnographic observation. The differences between census counts and the corresponding true counts at the level of the individual (or the household) are referred to collectively as census coverage errors, and in this chapter are categorized as types of census coverage error and indicate methods that can be used for their summarization. They then detail the three primary (potential) uses of census coverage measurement that rely on summarizations. Finally, they provide a brief overview of the methods that are currently used in the U.S. census for coverage measurement. This chapter also presents short histories of the U.S. census coverage measurement programs from 1950 to 1990, including a description of accuracy and coverage evaluation program, the coverage measurement program for the 2000 census.

Chapter 3

Plans for the 2010 Census. This chapter examines how the 2010 census differs from the 2000 census with respect to the impact on the coverage measurement program for 2010. It looks in some depth at the treatment of duplicates in the 2010 census and the 2010 coverage measurement program, including the possibility of contamination of the 2010 coverage measurement data collection through the application of the coverage follow-up interview. The chapter also discusses how the use of administrative records could potentially assist in both coverage improvement and coverage measurement for the 2010 census.

Chapter 4

Technical Issues. Here, a number of technical topics introduced by various changes made in coverage measurement for 2010 is discussed, including: (i) the sample design for the census coverage measurement post enumeration survey in 2010; (ii) the use of logistic regression modeling as a substitute for post stratification in modeling net coverage error; (iii) how one compares competing models in this situation; and (iv) the treatment of missing data in net coverage error modeling, including the Census Bureau's current plans for addressing missing data prior to fitting the logistic regression models in 2010. In relation to the issue of missing data, the chapter includes a description of an attempt by the Census Bureau to greatly reduce the number of cases that are considered to have insufficient information to support matching. The chapter concludes with a discussion of how to improve demographic analysis for use in census coverage measurement in 2010.

Chapter 5

Analytic Use of Coverage Measurement Data. First, the Census Bureau's framework for defining and estimating components of census coverage error is briefly outlined. Then, potential variables for use in statistical models to assess correlates of components of census coverage error are considered. The chapter ends with a consideration of the purpose of the key output from the census coverage measurement program in 2010—the analytic capability to develop statistical models linking census coverage errors of various types to individual and household characteristics and census process variables.

There are three appendixes:

Appendix A

A Framework for Components of Census Coverage Error.

A major goal and challenge for coverage measurement in 2010 is to design a survey that measures the components of coverage error, namely erroneous enumerations and omissions. The Census Bureau's previous coverage measurement surveys were designed primarily to estimate net census error using Dual System Estimation (DSE). To improve the accuracy of estimates of net error, the Census Bureau's DSE has relied on balancing some of the components of error, meaning some census omissions offset some erroneous inclusions in a manner that preserved the net error. As a result, the process produced inflated estimates of omissions and erroneous inclusions. This appendix summarizes Mulry and Kostanich (2006) paper, which provides a framework for overcoming these inflated estimates of component errors. It also explicitly defines the individual components of error and how these components relate to traditional net error concepts.

Appendix B

Logistic Regression for Modeling Match and Correct Enumeration Rates.

This is very technical appendix which provides details on the use of logistic regression models as a substitute for post stratification. More information is available in: (Malec, D., and Maples, 2005; Mulry et al., 2005; Schindler, 2006). This research suggests that inclusion of small-area effects could substantially improve coverage estimates. Several questions remain: how best to treat the complex sample design, how many random effects can be included and at what level of aggregation, the best way to estimate the model parameters, and how the model fit should be assessed. The panel is impressed with this high-caliber research that addresses an important issue in coverage modeling; further work in this area would be very valuable.

Appendix C

Biographical Sketches of Panel Members and Staff, provides biographical sketches of panel members and staff.

The panel offers 13 recommendations concerning coverage measurement plans for 2010.

To achieve this new goal, instead of only measuring net census error, the Census Bureau also plans to measure the four components of census coverage error: (1) census omissions, (2) census duplications, (3) erroneous census enumerations, and (4) census enumerations in the wrong location. The panel supports these plans, since different types of coverage errors are caused by different interactions between census processes and housing units and their occupants. The estimation of these four components of coverage error can be supported by the general structure of the data collection and matching that is carried out in support of dual-systems estimation, though modified and expanded to support this different purpose. The panel finds, however, that the Bureau's plans could be more fully developed for this purpose. Additionally, the panel recommends to allocate sufficient sources for research program on decennial census improvement in future.

Recommendation 1

The Census Bureau should more completely shift its focus in coverage measurement from that of collecting data and developing statistical models with the goal of estimating net coverage error to that of collecting data and developing statistical models that support the improvement of census processes.

Recommendation 2

The Census Bureau should allocate sufficient resources, including funding and staff, to assemble and support an ongoing intercensal research program on decennial census improvement. Such a group should focus on using the data from the census and the census coverage measurement programs to identify deficient census processes and to propose better alternatives. The work of this group should be used to help design the census tests early in the next decade.

Recommendation 3

The Census Bureau should retain comprehensive data on the functioning of the coverage follow-up interviews for a substantial sample of cases, especially for those cases in the CCM block clusters, to support detailed follow-up analysis of the functioning of the follow-up interviews and to help suggest modifications and alternatives for use in 2020.

Recommendation 4

The Census Bureau should organize census and coverage follow-up data collection so that data collection for the census coverage measurement (CCM) program is initiated as soon as possible after the completion of the census. In particular, the post enumeration survey in a particular area should start as soon as possible after the completion of the great majority of the census data collection—hopefully before late July. The Census Bureau should also consider census designs for 2010 in which there is some modest overlap between coverage follow-up and CCM data collections.

Recommendation 5

The Census Bureau should use the various testing opportunities in both the 2010 census and in the early part of the 2010–2020 intercensal period to assess how administrative records can be used in the 2020 census.

Recommendation 6

The Census Bureau should compare its sample design for the 2010 census coverage measurement post enumeration survey with alternative designs that give greater sampling probability to housing units that are anticipated to be hard to enumerate. If an alternative design proves preferable for the joint goals of estimating component coverage error and net coverage error estimation, such a design should be used in place of the current sample design.

Recommendation 7

The Census Bureau should develop missing data techniques, in collaboration with external experts if needed, which preserve associations between imputed and observed variables, condition on variables that are predictive of the missing values, and incorporate imputation uncertainty into estimates of standard errors. These ideas should be utilized in modeling the census coverage measurement data collected in the 2010 census.

Recommendation 8

The Census Bureau should give priority to research on improving demographic analysis in the four areas: (1) improving the measurement of undocumented and documented immigrants, (2) development of sub-national geographic estimates, (3) assessment of the uncertainty of estimates from demographic analysis, and (4) refining methods for combining estimates from demographic analysis and post enumeration survey data.

Recommendation 9

The Census Bureau should further develop and refine its framework for defining the four basic types of census coverage error and measuring their frequency of occurrence. The Census Bureau should also develop plans for operationalizing the measurement of these components using data from the census and the census coverage measurement program.

Recommendation 10

In developing the logistic regression models or other types of discriminant-analysis models of match status, correct enumeration status, and components of census coverage error, the Census Bureau should consider:

- Use of several approaches before focusing on a specific model; besides logistic regression, alternatives should include use of other link functions, discriminant analysis, and various data mining approaches, such as classification trees, support vector machines, and neural nets.
- Thorough examination of the subset of predictors that is best suited to each individual statistical model; the predictors for these various statistical models need not be identical; however, there may be a benefit to constraining the (logistic regression) models of match rate and correct enumeration rate to have identical variables in the estimation of net coverage error, and research should be carried out to assess whether this benefit outweighs the benefit of selecting variables that are optimal for each of these two logistic regression models.
- To effectively blend information from auxiliary sources at various levels of geographic and demographic aggregation, random effects modeling and Bayes' methods also should be examined.

Recommendation 11

The primary output of the Census Bureau's coverage measurement program in 2010 should be an analytic database that is used to support the development of statistical models to inform census process improvement. The production of summary tabulations should be of lesser priority.

Recommendation 12

The Census Bureau should develop regression models that elucidate the various types of census coverage error, using specified dependent and predictor variables. To the extent that the database supporting these models can be made available to external researchers, it is extremely important that the Census Bureau pursue all viable avenues to involve outside researchers in the development of such models.

Recommendation 13

For a sample of households, the Census Bureau should retain data that provide a comprehensive picture of the census processes used to enumerate it, and the individuals residing in it, to facilitate subsequent evaluation. To allow linking assessment of census coverage error with a history of the census processes, this sample should substantially overlap with the CCM sample.

Improvement the quality of subsequent censuses

Although it is important to assess census coverage, it would also be extremely helpful to use that assessment to improve the quality of subsequent censuses. Consequently, an important use of coverage measurement is to help to identify important sources of census coverage errors and possibly to suggest alternative processes to reduce the frequency of those errors in the future. Although drawing a link between census coverage errors and deficient census processes is a challenging task, the Census Bureau thinks that substantial progress can be made in this direction. Therefore, the 2010 coverage measurement program has the goal of identifying the sources of frequent coverage error in the census counts. This information can then be used to ***allocate resources toward developing alternative census designs and processes*** that ***will provide counts with higher quality in 2020***. It is conceivable that use of such a feedback loop could also provide substantial savings in census costs, in addition to improvement in census quality because the trade-off between the effect on accuracy and on census process costs might now be better understood. The panel fully supports this modification of the objectives of coverage measurement in 2010.

For our readers we would like also to recommend a new published monograph by UN Statistics Division (2010) on ***Post Enumeration Survey***, which is strictly connected with the reviewed book.

Prepared by Jan Kordos, Warsaw School of Economics,
E-mail: jan1kor2@aster.pl

REFERENCES

DYGASZEWICZ, J. (2008), *Project of the 2011 Census of Population and Housing in Poland – the Most Important Strategic Aims.* The Committee of the Demographic Research of the Polish Academy of Sciences (in Polish). http://www.knd.pan.pl/images/stories/pliki/pdf/Dygaszewicz_luty_2008.pdf

GOŁATA, E. (2009), Economic activity in population census 2011 and administration resources. In: Gołata, E. (Ed), *Methods and Sources of Obtaining Information in Public Statistics,* Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. Poznań. (in Polish).

GOŁATA, E. (2010), Indirect Estimation of Economic Activity for the Register-based Census. In: Gołata, E.(Ed), *Measurement and Information in Economy,* Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu, Poznań, pp. 85–104 (in Polish).

HOGAN, H. (2003). The Accuracy and Coverage Evaluation: Theory and Design. Survey Methodology, 29(2):129-138.

KORDOS, J. (2007), Some Aspects of Post-Enumeration Surveys in Poland, *Statistics in Transition-new series,* December 2007, Vol. 8, No. 3, pp. 563–576.

KORDOS, J. (2008), Methods of Census Data Quality Assessment – Current Practice in Poland, and Suggestions Connected with the 2011 Population Census Preparation (in Polish).
http://www.knd.pan.pl/images/stories/pliki/pdf/

KORDOS, J. (2010), Methods of Quality Assessment of Population Census Data.
http://www.knd.pan.pl/images/stories/Census-Q-_sl_-2011.pdf

MALEC, D., and MAPLES, J. (2005). An evaluation of synthetic, small-area census coverage error using a random effects model. Proceedings of the Section on Survey Research Methods of the American Statistical Association. Available:
http://www.amstat.org/Sections/Srms/Proceedings/y2005/Files/JSM2005-000180.pdf.

MULE, T. (2007). Coverage measurement for the 2010 census. Presented at the Washington Statistical Society Meeting, Washington, DC, January 23, 2008. Available: http://scs.gmu.edu/~wss/wss080123slides.pdf [accessed 11/25/08].

MULRY, M.H., SCHINDLER, E., MULE, T., NGUYEN, N., SPENCER, B.D. (2005). Investigation of extreme estimates of census coverage error for small areas. Proceedings of the Section on Survey Research Methods of the American Statistical Association: http://www.amstat.org/Sections/Srms/Proceedings/y2005/Files/JSM2005-000551.pdf.

MULRY, M.H., and KOSTANICH, D.K. (2006). Framework for census coverage error components. Proceedings of the Section on Survey Research Methods of the American Statistical Association. Available: http://www.amstat.org/sections/srms/Proceedings/y2006f.html

NATIONAL RESEARCH COUNCIL. (2000). Redesigning the 2010 Census: First Interim Report. Panel on Research on Future Census Methods, Committee on National Statistics. Washington, DC: National Academy Press.

NOWAK, L. (1998), Quality of Census Data, In: Tendencies of Changes in Structure of Population, Households and Families in 1998-1995 GUS, Warsaw, pp. 22–31 (in Polish).

NOWAK, L. (2008) , Th e 2011 *Census of Population and Housing – Methodology and Topics,* the Committee of Demographic Research of the Polish Academy of Sciences (in Polish). http://www.knd.pan.pl/images/stories/pliki/pdf/Nowak_28_luty_2008.pdf

PARADYSZ, J. (1989), On non-sampling errors in women's fertility survey in the 1970 National Population Census. In: *Problems of Statistical Surveys by Sampling*. GUS, Warszawa, BWS, vol. 36, pp. 154–159. (in Polish).

PARADYSZ, J. (2010), Necessity of Indirect Estimation in National Census. In: Gołata, E. (Ed); *Measurement and Information in Economy*, Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. Poznań, pp. 45–66 (in Polish).

SCHINDLER, E. (2006). 2010 census coverage measurement: The hunt for the magic variables. Proceedings of the Section on Survey Research Methods of the American Statistical Association. Available: http://www.amstat.org/sections/srms/Proceedings/y2006f.html.

UN STATISTICS DIVISION (2010), *Post Enumeration Surveys*, Operational guidelines, Technical Report, New York, April 2010.

STATISTICS IN TRANSITION-new series, December 2010
Vol. 11, No. 3, pp. 649—654

REPORT

The XXIX Conference on Multivariate Statistical Analysis, MSA 2010, 8–10 November 2010 Łódź, Poland

The 29th Conference on **Multivariate Statistical Analysis** took place on November 8th-10th 2010 in Łódź, Poland. The organization of the conference was entrusted to **Professor Czesław Domański**, the Chair of Department of Statistical Methods in University of Łódź and the President of Polish Statistical Association. This year MSA Conference was dedicated to Professor Zdzisław Hellwig.

The conference presented the latest theoretical and empirical achievements in the field of the multivariate statistical analysis and its applications. This was a continuation of the issues undertaken on the past years conferences. The scientific programme of MSA 2010 covered a wide range of various statistical problems, such as multivariate distributions, statistical tests, non-parametric inference, discrimination analysis, Monte Carlo analysis, Bayesian inference, application of statistical methods for finance, insurance, capital markets and risk management.

This year MSA Conference began with the meeting of the Main Council of the Polish Statistical Association.

Altogether there were 70 participants from various academic and research centers in Poland and abroad. Concerning the papers, 41 papers were presented in 11 sessions.

The conference was opened by the Chairman of the Organizing Committee, **Professor Czesław Domański**. On the opening speech also spoke the Vice-rector in Charge of Research of the University of Łódź, **Prof. Antoni Różalski** and the Pro–Dean for Research of the Faculty of Economics and Sociology of the University of Łódź, **Prof.** Jolanta Grotowska-Leder**.**

For the first **plenary session** (The Chair of this session was **Prof. Eugeniusz Gatnar**) 4 papers were delivered:

Prof. Czesław Domański (University of Łódź) reported the paper titled "*Statistical tests based on empty cells*".

Prof. Grażyna Trzpiot (University of Katowice) delivered a lecture titled "*Some properties of the robust trend tests*".

Prof. Joanna Kisielińska (Warsaw University of Life Sciences) presented the paper about "*Bypassing the condition of normality in parametric tests*".

Prof. Grzegorz Kończak (University of Katowice) reported the paper titled *“On testing the significance of the coefficients in the multiple linear regression model”*.

The second **plenary session** (chair: Prof. Mirosław Szreder) was devoted to famous Polish statisticians, including ***Kazimierz Władysław Kumaniecki**, **Marek Fisz**, **Zbigniew Pawłowski**, **Jerzy Greń** and* **Wiktor Oktaba.**

Prof. Czesław Domański (University of Łódź) delivered a lecture about **“***Kazimierz Władysław Kumaniecki – statistician, a founder of Polish Statistical Association***”**. In March 1912 K.W. Kumaniecki submitted the documents to the Imperial and Royal Governance in Lviv, together with a statue, probably written by himself, asking for a permission to launch Polish Statistical Association with headquarter in Cracow. When had obtained the permission, after the election of the first Association's Authorities, he was nominated to its Secretary. The Association's main premises were located in Cracow's Municipal Statistical Office headed by K.W. Kumaniecki.

Prof. Mirosław Krzyśko (Adam Mickiewicz University in Poznań) reported a paper titled *“**Professor Marek Fisz (1910–1963). On the hundredth anniversary of his birth**”*.

Marek Fisz was born on January 15, 1910 in Szydłowiec. He studied mathematics at the University of Warsaw from 1934 to 1939, and in June 1939 he received the Master Degree. On June 23, 1951 the degree of Doctor of Science was conferred on him by the University of Wrocław. From 1954 he was the Head of the Chair of Mathematical Statistics at the University of Warsaw and from 1958 he was the Chief of the Division of Mathematical Statistics in the Mathematical Institute of the Polish Academy of Sciences. In 1960 he moved to the United States. In the United States he held different scientific positions, respectively at the University of Washington (Seattle), Stanford University (Stanford), Columbia University and New York University (New York).

Prof. Janusz Wywiał (University of Katowice) presented a lecture about the following topic: *“**On selected scientific works of Zbigniew Pawłowski on the occasion of the 80th anniversary of the birthday**”*. Professor Zbigniew Maria Pawłowski lived in the years 1930-1981. His scientific career began in the Central School of Planning and Statistics in Warsaw (now known as The Warsaw School of Economics), where he had studied and then began working as a deputy assistant in the Department of Statistics. There he obtained the PhD degree in 1957, and in 1962 - a postdoctoral degree. Since 1962 he continued his career in the Higher School of Economics (later renamed The University of Economics) in Katowice, where he took the position of the Head of the Department of Statistics. In 1957 he received the title of an associate professor and respectively in 1972 – of professor ordinarius. The opinion that prevails in the environment of Polish econometricians, mathematicians and statisticians is that Professor Zbigniew Pawłowski was one of the pioneers of econometrics in Poland.

Prof. Jan Kordos (Warsaw School of Economics) reported the paper *“**In memory of Professor Jerzy Greń (1936–1985)**”*. This year it is 25th anniversary of Professor Jerzy Greń's death, who was one of the most famous Polish statisticians and a researcher of the Institute of Econometrics, Warsaw School of Economics. Professor Jerzy Greń was born on 14 March 1936 in Cięcin, in working-class family. University Diploma of the Warsaw School of Economics-received with honors in 1958. The academic title of Professor was given to Jerzy Greń by the President of Poland in 1979. In the years 1962-1981 he worked at the Commission Mathematical of the Central Statistical Office of Poland.

In years 1981–1984 he worked in Ethiopia as an international expert in econometrics, preparing the Integrated System of Food and Agricultural Statistics in Ethiopia. He trained in econometrics Ethiopian statisticians, prepared project documentation, and published:

- Handbook on Applied Econometrics, Central Statistical Office in Ethiopia, Addis Ababa, 1983, 210 pages;
- Forecasting in Agricultural Statistics, Statistical Techniques in Developing Countries No 2., ESS/STDC/2, FAO, Rome 1983, p. 21.

He died early, at the age of 49, but left a significant legacy of scientific books, scientific articles, different analysis and opinions.

Prof. Bronisław Ceranka (University of Life Sciences in Poznań) delivered a lecture about "***Wiktor Oktaba***" . Wiktor Oktaba (born on April 16, 1920 in Kiev, died on September 6, 2009 in Lublin) was a Polish mathematician and statistician. One of the champions of biometry science in Poland. Oktaba studied mathematics at the University of Warsaw and Maria Curie-Skłodowska University of Lublin. Oktaba was professor and the Chair of Department of Mathematical Statistics and Institute of Application of Mathematics Academy of Agricultural of Lublin.

Dr. Milda Maria Burzała (Adam Mickiewicz University in Poznań) reported the paper titled "***In memory of Professor Bogusław Guzik***". Professor Bogusław Guzik unexpectedly passed away on 6th of July 2009, at the age of 63, when he was in his prime. Since the beginning of his studies, he was connected with University of Economics in Poznań. In 1969 he achieved Master of Science degree there. In 1975, he obtained the degree of Doctor of Economic Science, and in 1980 - the postdoctoral degree. The title of professor was given to him in 1991. He was the author of about 300 scientific publications in the area of econometrics, forecasting, multidimensional comparative analysis and also analysis of economic efficiency.

Titles of the papers of the next sessions of the MSA Conference, with the authors' names are presented respectively below:

9 November 2010

Session III A:
The Chair: Prof. **Janusz Wywiał**

- *Methodology and example of Discriminant Analysis to separate a study population by treatment subgroups in a Phase 2 clinical trial* (**Lev Sverdlov, Merck Research Labs, USA)**
- *Theory of multiple sequences and series. A review of the elements, which might be useful in works on discrete distribution* (**Janusz Kupczun, Łódź)**
- *Some remarks on bootstrap, confidence interval and calibration in social surveys* (**Jan Kordos, Warszawa)**
- *Assessment of entropy based methods for choosing uniformly distributed variables noisy in the context of cluster analysis* (**Jerzy Korzeniewski, Łódź**)

Session III B:
The Chair: Prof. Walenty Ostasiewicz

- *The integration of data in statistics of Entrepreneurship* (**Grażyna Dehnel, Elżbieta Gołata, Poznań)**
- *Test of independence among random vectors based on copula functions* (**Joanna Tomanek, Katowice)**
- *Territorial differentiation in dynamics of enterprises' population in Poland – cluster analysis* (**Aneta Ptak-Chmielewska, Warszawa**)
- *Attempt to use administrative registers in the study of commuting* **(Hanna Gruchociak, Poznań)**

Session IV A:
The Chair: Prof. **Grażyna Trzpiot**

- *On generation of correlated pseudo-random binary numbers* (**Wojciech Gamrot, Katowice)**
- *On pseudo-EBLUP under some model for longitudinal data with auxiliary variables* **(Tomasz Żądło, Katowice)**
- *Sampling with SAS for practical use* (**Dorota Bartosińska, Warszawa**)
- *On the Data Imputation using R* (**Małgorzata Misztal, Łódź**)

Session IV B:
The Chair: Prof. **Wojciech Zieliński**

- *Subjective methods of extraction and data visualization in factor analysis* **(Piotr Tarka, Poznań)**
- *Applicability of the multi group confirmatory factor analysis to constructing sentiment measures* (**Piotr Białowolski, Warszawa**)

- *Principal component analysis method for symbolic interwal data* (**Andrzej Dudek, Wrocław)**
- *About adaptation of unfolding analysis for symbolic data* (**Artur Zaborski, Marcin Pełka, Wrocław)**

Session V A:
The Chair: Prof. **Mirosław Krzyśko**

- *Comparison of stability of classical taxonomy bagging metod with bagging based on co-occurence data* **(Dorota Rozmus, Katowice**)
- *Dynamic Grouping on the basis of self-learning network GNG* (**Kamila Migdał Najman, Krzysztof Najman, Gdańsk)**
- *Evaluation of post-secondary school students' preferences, using the analysis of hidden classes* (**Marcin Pełka, Aneta Rybicka, Wrocław)**

Session V B:
The Chair: Prof. **Marek Walesiak**

- *Hierarchical log-linear models for contingency tables* (**Justyna Brzezińska, Katowice**)
- *Feature selection in high dimensional regression problem* **(Mariusz Kubus, Opole**)
- *Using permutation test in multiple correlation investigation* (**Jacek Stelmach, Katowice)**

10 November 2010:

Session VI A:
The Chair: Prof. **Jan Kordos**

- *Special cases of some general formula for price indices* (**Jacek Białek, Łódź**)
- *Review of selected applications of domes in finance and insurance* (**Krzysztof Janas, Agnieszka Świerczyńska, Łódź)**
- *Statistical properties of a control design Supreme Chamber of Control* **(Wojciech Zieliński, Warszawa)**
- *Multiple decision procedures* **(Dariusz Parys, Łódź**)

Session VI B:
The Chair: Prof. **Bronisław Ceranka**

- *Bayesian exponential survival model in the analysis of unemployment duration determinants* (**Wioletta Grzenda, Warszawa**)
- *Convergence of food share expenditures as a problem in welfare analysis* **(Hanna Dudek, Warszawa)**
- *Modified soft model of sustainable development* (**Dorota Perło, Białystok**)

Plenary Session VII
The Chair: Prof. **Zofia Rusnak**

- *On some construction of regular A-optimal spring balance weighing design* (**Małgorzata Graczyk, Poznań**)
- *Notes on the optimum chemical balance weighing design* (**Bronisław Ceranka, Małgorzata Graczyk, Poznań)**
- *Selected Variance Estimation Methods and their Application in the Case of Income Concentration Measures* (**Alina Jędrzejczak, Łódź**)

The next Conference ***on Multivariate Statistical Analysis*** has been planned on November 7th-9th, 2011 and will take place in Lodz. The Chairman of the Organizing Committee, Professor **Czesław Domański** informs it will be the 30th anniversary edition of the Conference and kindly invites all interested Scientists, Researchers and Students to take part in it. If interested, please send an application to the Scientific Secretary of the MSA 2011 Conference at the following address:

Aleksandra Fijałkowska
30th Conference MSA, Łódź 2011
Department of Statistical Methods, University of Łódź
90-214 Łódź, Rewolucji 1905 r. nr 41, Poland
Contact number : (+48) 42 635 51 78, Fax number: (+48) 42 635 53 07
E-mail address : msa@uni.lodz.pl, Web: www.msa.uni.lodz.pl

Prepared by:
Monika Zielińska-Sitkiewicz,
Anna Witaszczyk