Journal of Official Statistics, vol. 29, n. 4 (2013)

Preface ..... p. 471
Li-Chun Zhang

Selective Editing: A Quest for Efficiency and Data Quality..... p. 473–488
Ton de Waal

An Optimization Approach to Selective Editing ..... p. 489–510
Ignacio Arbue´s, Pedro Revilla, David Salgado

Automated and Manual Data Editing: A View on Process Design and Methodology..... p. 511–538
Jeroen Pannekoek, Sander Scholtus,Mark Van der Loo

A Contamination Model for Selective Editing ..... p. 539–556
Marco Di Zio, Ugo Guarnera

Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey ..... p. 557–582
Peter Lundquist, Carl-Erik Särndal

Utilising Expert Opinion to Improve the Measurement of International Migration in Europe .....p. 583–608
Arkadiusz Wiśniowski, Jakub Bijak, Solveig Christiansen, Jonathan J. Forster, Nico Keilman, James Raymer, Peter W.F. Smith

Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time ..... p. 609–622
Anja Mohorko, Edith de Leeuw, Joop Hox

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, p. 471, http://dx.doi.org/10.2478/jos-2013-0043

Preface

In the present issue of JOS, we are proud to feature a special section on selective editing. Statistical data editing is an important and resource-demanding activity at national statistical institutes and methodological improvements are crucial for a sound practice. This issue aims to bring forward notable recent theoretical and methodological works on selective editing, to provide an overview of the historic developments and current status of the field, and to inspire future research.

The work on this special section was initiated at the UNECE Work Session on Statistical Data Editing in Oslo, Norway, in September 2012. The JOS Editorial Board acknowledges all contributors.

Li-Chun Zhang, Guest Editor
Annica Isaksson and Ingegerd Jansson, Co-Editors-in-Chief

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 473–488, http://dx.doi.org/10.2478/jos-2013-0036

Selective Editing: A Quest for Efficiency and Data Quality

Ton de Waal[1]

National statistical institutes are responsible for publishing high quality statistical information on many different aspects of society. This task is complicated considerably by the fact that data collected by statistical offices often contain errors. The process of correcting errors is referred to as statistical data editing. For many years this has been a purely manual process, with people checking the collected data record by record and correcting them if necessary. For this reason the data editing process has been both expensive and time-consuming. This article sketches some of the important methodological developments aiming to improve the efficiency of the data editing process that have occurred during the past few decades. The article focuses on selective editing, which is based on an idea rather shocking for people working in the production of high-quality data: that it is not necessary to find and correct all errors. Instead of trying to correct all errors, it generally suffices to correct only those errors where data editing has substantial influence on publication figures. This overview article sketches the background of selective editing, describes the most usual form of selective editing up to now, and discusses the contributions to this special issue of the Journal of Official Statistics on selective editing. The article concludes with describing some possible directions for future research on selective editing and statistical data editing in general.

Key words: Errors; score function; selective editing; statistical data editing.

1. Introduction

National statistical institutes (NSIs) play a vital role as providers of objective statistical information about society. Statistical figures published by NSIs are used to inform policies and actions in government, trade unions, employer organisations and so on. The statistical figures are also used as a basis for researching the "societal story": what is the current economic and sociological state of society and what main economical and sociological changes have taken place over time? For these purposes it is of the utmost importance that the statistical information provided by NSIs is of high quality.

Let us go several decades back in time and consider such an NSI. As do most NSIs, it has well-trained and excellent statisticians to produce high-quality statistics. The NSI carefully plans a survey, develops the questionnaire and a clever sampling design. Next, it spends a lot of money, time and energy to actually collect the data. After this painstaking process, the NSI is ready to analyse the observed data and publish the statistical outcome.

[1] Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands, Email: t.dewaal@cbs.nl
Acknowledgments: The views expressed in this article are those of the author and do not necessarily reflect the policies of Statistics Netherlands. I am grateful to Li-Chun Zhang, Natalie Shlomo, Bart Bakker and authors of the three articles on selective editing in this issue for their useful comments.

Now, let us suppose that statisticians at this NSI, while analysing the observed data, discover that the collected data contain errors. The data, often literally, do not add up. For example, components of a total do not add up to the overall total, or data of some respondents are an unlikely number of times larger than the data of similar respondents. Such an NSI would obviously try to correct these errors. And what could be more natural than trying to find as many errors as possible and correcting them all? This was the situation for many years at NSIs all over the world. As noted by Granquist (1997, p. 383), implicitly in those days the process was governed by the paradigm: "The more checks and recontacts with the respondents, the better the resulting quality".

In a sense the situation has not changed much over the years: NSIs still have well-trained and excellent statisticians to produce high-quality statistics. In another sense, much has changed over the last few decades. At most NSIs, there are fewer resources to do the work while output expectations have increased. So the work has to be done much more efficiently, for instance by relying more on automated procedures (see, e.g., Pannekoek et al. in this issue), while striving to ensure high quality statistics. Over the years, staff at NSIs have become much more proud of doing their work as efficiently as possible.

In this overview article to the special issue of the Journal of Official Statistics on selective editing, I will sketch some of the important methodological developments with respect to processing data at NSIs that have taken place during the past few decades, and that are still taking place as testified by the articles on selective editing in this issue.

The major change in the statistical process at NSIs that I want to discuss is a thought that is rather shocking for people working in the production of high-quality data: that it is not necessary to find and correct all errors, even if things just do not add up. Instead of trying to correct all errors, one should look at the entire process from a Total Quality Management point of view (see also Granquist 1995, and Granquist and Kovar 1997) and focus on the errors that really matter.

Before we proceed, let us first define statistical data editing in general and selective editing in particular. Statistical data editing is the procedure for detecting and "correcting" errors in observed data. Here I have put correcting in inverted commas, as in practice one generally cannot be sure if one is really correcting the data. In the remainder of this overview article I will not put correcting in inverted commas. The reader should keep in mind that they should be there, as in fact all one can do in general is to *try* to correct errors as well as possible.

De Waal et al. (2011) define selective editing as an editing strategy in which manual editing is limited or prioritised to those errors where this editing has substantial influence on publication figures. According to Granquist and Kovar (1997) selective editing includes any approach which focuses the editor's attention on only a subset of the potentially erroneous microdata items or records that would be identified by traditional manual or interactive editing methods.

Selective editing, or even statistical data editing in general, is a relatively unknown part of (official) statistics in the literature, although NSIs have always put much effort and resources into statistical data editing as they consider it a prerequisite for publishing accurate statistics. For business surveys, the monetary costs of editing at NSIs have even been estimated as high as 40 per cent of the total budget (see Granquist and Kovar 1997). The related problem of estimating missing values (imputation), which can be seen as

detecting and correcting a special, easily detectible kind of error, is much better known in the literature and has been studied in much more detail, not only by NSIs but also, and especially, by academia.

The remainder of this overview article is organised as follows. Section 2 sketches the history of statistical data editing in general, while Section 3 describes the background of selective editing. Section 4 briefly describes the most usual and general form of selective editing up until now, which is based on so-called score functions. Section 5 focuses on the three articles on selective editing by Arbués et al. Di Zio and Guarnera and Pannekoek et al. in this special issue of the Journal of Official Statistics. Finally, Section 6 describes some possible directions for future research on selective editing and statistical data editing in general.

2. A Brief History of Statistical Data Editing

Statistical data editing is likely to be as old as statistics itself. Errors have always been present in statistical data. The data collection stage in particular is a potential source of errors. For instance, a respondent may give a wrong answer (intentionally or not), a respondent may not give an answer (either because he does not know the answer or because he does not want to answer this question), errors can be introduced at the NSI when the data are transferred from the questionnaire to the computer system, and so on. When these errors have been detected, people have tried to correct them. For many years this has been a purely manual process, with people checking the collected data record by record and correcting them if necessary.

We start our brief history of statistical data editing not in these "ancient" times, but somewhere around the 1950s. In the 1950s some NSIs started using electronic computers in the editing process (see Nordbotten 1963). This led to major changes in the editing process. In the early years the role of computers was, however, restricted to checking which edit rules were violated. Edit rules, or edits for short, are user-specified rules that have to be satisfied by the data. Examples of such edits are that the profit of an enterprise should be equal to its total turnover minus its total costs, and that the total turnover of an enterprise should be non-negative. Professional typists entered data into a mainframe computer. Subsequently, the computer checked whether these data satisfied all specified edits. For each record all violated edits were listed. Subject-matter specialists then used these lists to correct the records, that is, they retrieved all paper questionnaires that did not pass the edits and corrected these questionnaires. After they had corrected the data, these data were again entered into the mainframe computer, and the computer again checked whether the data satisfied all edits. This iterative process continued until (nearly) all records passed all edits.

A major problem with this approach was that during the manual correction process, the records were not checked for consistency. As a result, a record that was corrected could still fail one or more specified edits. Such a record hence required more correction. The advent of PCs in the 1980s enabled an improved form of computer-assisted manual editing, called interactive editing. With interactive editing, the consistency of the entered data can be checked during data entry. The computer runs consistency checks and displays a list of edit violations per record on the screen. Subject-matter specialists can manually

edit the data directly. After manual editing, the computer immediately checks the edits again. Each record is edited until it satisfies all edits. Checking and correction can thus be combined into a single processing step. Interactive editing has become so standard over the course of time that manual editing and interactive editing have become synonymous terms.

Nevertheless, even with interactive editing too much effort was spent on correcting errors that did not have a noticeable impact on the figures ultimately published. This has been referred to as "over-editing". Over-editing not only costs money, but also a considerable amount of time, making the period between data collection and publication unnecessarily long. Sometimes over-editing even becomes "creative editing": the editing process is then continued for such a length of time that unlikely, but correct, data are unjustifiably changed into more likely values. Such unjustified alterations can be detrimental for data quality. For more on the dangers of over-editing and creative editing see, for example, Granquist (1995, 1997) and Granquist and Kovar (1997).

There are several editing approaches that aim to reduce the effort spent on correcting data: selective editing, automatic editing, and macro-editing. Macro-editing is sometimes seen as a special form of selective editing. In this article, however, I will consider macro-editing as a separate form of editing and reserve the term selective editing for editing approaches that automatically select or prioritise items or records for manual review without any human interference, apart from specification of metadata or parameters. Here I will briefly discuss automatic editing and macro-editing. Selective editing is discussed in subsequent sections of this overview article.

The aim of automatic editing is to let a computer do all the work. The main role of the human is to provide the computer with metadata, such as edits and imputation models. After the metadata have been specified, the computer edits the data and all the human has to do is examine the output generated by the computer. In case the quality of the edited data is considered too low, the metadata have to be adjusted or some records have to be edited in another way.

In the 1960s and early 1970s, automatic editing was usually based on predetermined rules of the following kind: if a certain combination of edits is violated in a certain way, then a certain action has to be undertaken to correct the data. Freund and Hartley (1967) proposed an alternative approach based on minimising the total deviation between the original values in a record and the corrected values plus the total violation of the edits (the more an edit after correction of the data is violated, the more this edit contributes to the objective function). In this way only the edits had to be specified in order to find the corrected values. The approach by Freund and Hartley never became popular, probably because edits may still be violated after correction of the data – and often are.

In 1976, Fellegi and Holt (Fellegi and Holt 1976) published a landmark paper in the Journal of the American Statistical Association. In their article, Fellegi and Holt described a new paradigm for localising errors in a record automatically. According to this paradigm, the data of a record should be made to satisfy all edits by changing the values of the fewest possible number of variables. This paradigm became the standard on which most systems for automatic editing, such as GEIS (Kovar and Whitridge 1990), SCIA (Barcaroli et al. 1995), CherryPi (De Waal 1996), SPEER (Winkler and Draper 1997), DISCRETE

(Winkler and Petkunas 1997), AGGIES (Todaro 1999), SLICE (De Waal 2001), and Banff (Banff Support Team 2008) are based. The mathematical optimisation problem implied by this paradigm can be solved in several ways. For an overview, I refer to De Waal and Coutinho (2005).

In the 1990s, a new form of editing emerged: macro-editing. Macro-editing offers a solution to some of the problems of micro-editing. In particular, macro-editing can deal with editing tasks related to the distributional aspects of the data. It is common practice to distinguish between two forms of macro-editing. The first form is sometimes called the aggregation method (see e.g., Granquist 1990). It formalises and systematises what every statistical agency does before publication: verifying whether figures to be published seem plausible. This is accomplished by comparing quantities in publication tables with, for instance, the same quantities in previous publications. Only if an unusual value is observed a micro-editing procedure is applied to the individual records and fields contributing to the quantity in error. A second form of macro-editing is the distribution method. The available data are used to characterise the distribution of the variables. Then all individual values are compared with the distribution. Typically, measures of location and spread are computed. Records containing values that could be considered uncommon (given the distribution) are candidates for further inspection and possibly for editing. In macro-editing, graphical techniques are often used to visualise outlying and suspicious records. Generally, there is human interaction to select records for manual review.

For more on these techniques and on how they can be combined into an editing strategy, I refer to De Waal et al. (2011).

3. Background of Selective Editing

The grand idea that it is not necessary to edit all data in every detail was already expressed in the 1950s, although back then it was stated in a reverse way, namely that it was not necessary to do more editing than NSIs already did. Nordbotten (1955) described an early successful attempt to measure the influence on publication figures of errors that remain after manual editing. A random sample of records from the 1953 Industrial Census in Norway was re-edited using every available resource (including re-contacts), and the resulting estimates were compared to the corresponding estimates after ordinary editing (without re-contacts). No significant deviations were found on the aggregate level. With this study, Nordbotten (1955) showed that the less intensive form of manual editing used in practice was sufficient to obtain accurate statistical results. In other words: the experimental "gold standard" editing process would have led to over-editing if used in practice.

The grand idea had to wait until the 1980s and 1990s before it became popular. Up until then, the paradigm "the more edits and corrections, the better the quality" still prevailed. The grand idea forms the basis for selective editing. Studies such as Granquist (1995, 1997) and Granquist and Kovar (1997) have shown that generally not all errors have to be corrected to obtain reliable publication figures. It usually suffices to remove only the most influential errors. They also showed that in practice, editing can indeed be counterproductive and, when taken too far, even detrimental to data quality.

These and other studies show that the cost of editing cannot be justified by quality improvement. A major conclusion from these studies is that too many values are being edited. As noted by Granquist and Kovar (1997, p. 431): "many statistical offices are risking too much in their quest for perfection".

One of the important observations was that small errors in the data often more or less cancel out when aggregated, that is, their sum generally tends to be negligible in comparison to the corresponding publication figure. Another important observation was that, on an aggregated level, the total measurement error due to small measurement errors in individual records is often negligible in comparison to other errors in the survey process, such as the sampling error, under-coverage, over-coverage and nonresponse error.

As figures published by NSIs are aggregated data, such as totals and means, leaving small errors in the data is fully acceptable and does not diminish the quality of the data on an aggregated level, or at least not by much. Moreover, parameters estimated from most statistical models are also derived by some form of aggregated data and therefore it is not necessary for parameter estimation either to correct all data in every detail (see e.g., Pullum et al. 1986, and Van de Pol and Bethlehem 1997).

The studies by Granquist (1995, 1997), Granquist and Kovar (1997) and others have been confirmed by many years of practical experience at NSIs. As a result, research has been focused on effective selective editing methods to single out the records for which it is likely that interactive editing will lead to a significant improvement in the quality of estimates. Besides being referred to as selective editing, these methods are sometimes also known as significance editing.

Methods for selecting records for interactive editing that are specifically designed for use in the early stages of the data collection period are called input editing methods. Sometimes they are also referred to as micro-selection methods or micro-based selective editing methods (see Pursey 1994, and De Waal et al. 2011). These methods can be applied to each incoming record individually. They are based on parameters that are determined before the data collection takes place, often estimated using previous versions of the survey and the values of the target variables in the record under consideration. The purpose of such methods is to start the time-consuming interactive editing as soon as the first survey data are received. Other methods, referred to as output editing, macro-selection methods, or macro-based selective editing methods are designed to be used when the data collection is (almost) completed. These methods use the information from (nearly) all data of the survey to detect suspect and influential values. When (nearly) all survey data are available, estimates of target parameters can be calculated and the influence of editing outlying values on these parameters can be estimated.

The scope of most techniques for selective editing is limited to (numerical) business data. In these data some respondents can be more important than other respondents, simply because the magnitude of their contributions is higher. Social data are usually count data where respondents contribute more or less the same, namely their raising weight, to estimated population totals. In social data it is therefore difficult to differentiate between respondents. For social data micro-integration techniques (see e.g., Bakker 2011) are often used to efficiently integrate data from different data sources, for example a register and a survey. Errors are then corrected by comparing these data sources on a micro-level.

For business data, selective editing has gradually become a popular method and increasingly more NSIs use selective editing techniques.

4. The Basics of Selective Editing

This section briefly describes the basics of the most common form of selective editing up to now, which is based on so-called score functions. For a substantial part, this section is based on De Waal et al. (2011).

4.1. Introduction

The aim of selective editing is to split the data into two streams: the critical stream and the noncritical stream. The critical stream consists of records that are the ones most likely to contain influential errors; the noncritical stream consists of records that are unlikely to contain influential errors. The records in the critical stream are edited in an interactive manner. The records in the noncritical stream are edited not interactively but automatically, or – in some cases – not at all. When selective editing is used, automatic editing, for instance based on the Fellegi-Holt paradigm, is confined to correcting the relatively unimportant errors. One purpose of automatic editing, besides correcting small errors, is then to ensure that the data satisfy the most important edits, so that obvious inconsistencies cannot occur at any level of aggregation.

At present no accepted theory for selective editing exists. In fact, selective editing is an umbrella term for several methods to identify the errors that have a substantial impact on the publication figures (see, for instance, Hidiroglou and Berthelot 1986, Granquist 1990, Latouche and Berthelot 1992, Lawrence and McDavitt 1994, Lawrence and McKenzie 2000, and Hedlin 2003 for examples of such methods). It is hardly possible to describe here all selective editing methods that have been developed over the years. Many selective editing methods are relatively simple *ad hoc* methods based on common sense. A leading principle in most selective editing methods was suggested in Granquist (1997, p. 384): "begin with the most deviating values and stop verifying when (macro-)estimates no longer are changed". This is still the leading principle nowadays. The most frequently applied general approach to implement this principle is to use a score function (see e.g., Hidiroglou and Berthelot 1986).

A score for a record is referred to as a record or global score. Such a global score is usually a combination of scores for each of a number of important variables, which are referred to as local scores. A local score is generally defined so that it measures the influence of editing a field on the estimated total of the corresponding variable. In the following subsections I will briefly examine local scores, global scores, and setting threshold values on the global score for splitting the data into the critical and the noncritical streams.

4.2. Local Scores

Local scores are generally based on two components: the influence component and the risk component. Local scores are then defined as the product of these two components, that is,

$$s_{ij} = F_{ij} \times R_{ij}$$

with s_{ij} the local score, F_{ij} the influence component and R_{ij} the risk component for unit i and variable j.

The risk component measures the likelihood of a potential error. This likelihood of a potential error can, for instance, be estimated by the ratio of the absolute difference of the observed raw value and an "anticipated" value which is an estimate of the true value or the value that would have been obtained after interactive editing.

In formula form, the risk component can, for instance, be defined as

$$R_{ij} = \frac{|x_{ij} - x_{ij}^*|}{|x_{ij}^*|},$$

where x_{ij} is the value of variable j in unit i and x_{ij}^* is the corresponding "anticipated" value. Large deviations from the "anticipated" value are taken as an indication that the raw value may be in error. Small deviations indicate that there is no reason to suspect that the value is in error.

The influence component measures the relative influence of a field on the estimated total of the target variable. The influence component can, for instance, be defined as

$$F_{ij} = w_i|x_{ij}^*|, \quad (1)$$

where x_{ij}^* is defined as above and w_i is the design weight of unit i.

Multiplying the risk factor by the influence factor results in a measure for the effect of editing a field on the estimated total. In our example, the local score would be given by

$$s_{ij} = w_i|x_{ij} - x_{ij}^*|,$$

which measures the effect of editing variable j in unit i on the total for variable j.

Large values of the local score indicate that the field may contain an influential error and that it is worth spending time and resources on correcting the field. Smaller values of the local score indicate that the field does not contain an influential error.

In general, an "anticipated" value is modelled as a function of auxiliary variables. For instance, the "anticipated" value of some variables may be modelled as the dependent variable in a regression model with auxiliary variables. Auxiliary variables should be free from gross errors, otherwise the corresponding "anticipated" values can be far from the true values (or the values that would have been obtained after interactive editing) and become useless as reference values. Auxiliary variables and estimates of model parameters can sometimes be obtained from the current survey, but are more often obtained from other sources such as a previous, already edited version of the survey or administrative sources.

4.3. Global Scores

A global score is a function that combines the local scores to form a measure for the whole record. Such a global score is needed to decide whether or not a record should be selected for interactive editing.

The global score should reflect the importance of editing the complete record. In order to combine scores, it is important that the local scores are measured on comparable scales. It is common, therefore, to scale local scores before combining them into a global

score. One method for scaling local scores is by dividing by the (approximated) total of the corresponding variable. Another method is to divide the scores by the standard deviation of the "anticipated" values (see Lawrence and McKenzie 2000). This last approach has the advantage that deviations from "anticipated" values in variables with large natural variability will lead to less high scores and are therefore less likely to be designated as suspect values than deviations in variables with less variability.

Scaled local scores can be combined to form a global score in several different ways. Often, the global score is defined as the sum of the local scores (see e.g., Latouche and Berthelot 1992). As a result, records with many deviating values will get high scores. An alternative is to take the maximum of the local scores (see e.g., Lawrence and McKenzie 2000). The advantage of taking the maximum is that it guarantees that a large value on any one of the contributing local scores will lead to a large global score and hence interactive editing of the record. The drawback of this strategy is that it cannot discriminate between records with a single large local score and records with numerous equally large local scores. Compromises between these two options have been proposed by Farwell (2005) and Hedlin (2008). In fact, the compromise proposed by Hedlin (2008) encompasses taking the sum and taking the maximum as two extreme options. One can also multiply local scores by weights, not to be confused with the design weights in (1), expressing that some variables are considered more important than others (see Latouche and Berthelot 1992).

4.4. Setting Threshold Values

When one wants to apply input editing, a threshold or cut-off value has to be determined in advance so that records with global scores above the threshold are designated as not plausible. These records are assigned to the critical stream and are edited interactively, whereas the other records with less important errors are assigned to the noncritical stream.

The most frequently used method for determining a threshold value is to carry out a simulation study to examine the effect of a range of potential threshold values on the bias in the principal output parameters. In an ideal situation, such a simulation study would be based on a raw unedited data set and a version of the same data set in which all records have been extensively edited interactively so that all true values have been recovered. These data must be comparable with the data to which the threshold values are applied. Often, data from a previous period of the same survey are used for this purpose. The simulation study now proceeds according to the following steps:

- Calculate the global scores according to the chosen selective editing method for the records in the raw version of the data set.
- Simulate that only the first $p\%$ of the records is designated for interactive editing. This is done by replacing the values of the $p\%$ of the records with the highest global scores in the raw data by the values in the edited data.
- Calculate the target parameters using both the $p\%$-edited data set and the true values.

These steps are repeated for a range of values of p. The effect of editing $p\%$ of the records can be measured by the differences between the estimates of the target parameters based on the $p\%$-edited data set and the true values. The costs (resources, timeliness, etc.) are usually estimated by assuming fixed amounts of resources and time per record to be edited.

Such fixed amounts of resources and time can be based on previous experiences with editing these kinds of data. Sometimes different costs are used for different classes of records. The threshold value corresponding to the value of p with the "best" trade-off between costs and data quality is then chosen. What is considered the best value of p is a policy decision to be made by the NSI.

The ideal situation of having a fully interactively edited version of the data set in which all true values have been recovered is, however, very unlikely to arise in practice. Generally, only a subset of the records will have been edited interactively, and not even for those records will all true values be recovered. In such a case, one can only check as well as possible (and hope!) that the edited data set is a good proxy for the true values.

4.5. Other Approaches

Although a score function approach, in one form or another, is thus far the most popular way to implement selective editing, it is by no means the only way to implement a selective editing approach. Other ways of selecting and prioritising records for manual review have been developed and implemented, such as an edit-related approach that measures the extent to which a record fails edit rules (see Hedlin 2003). For other approaches see Arbués et al. in this issue and Chapter 6 of De Waal et al. (2011).

5. The Current Issue of the Journal of Official Statistics

In this issue of the Journal of Official Statistics, we are witnessing the formalisation of selective editing. Whereas until now NSIs relied on rather *ad hoc* methods for selective editing, such as those described in the previous section, in this issue theoretical frameworks for selective editing are being developed.

As seen in the three articles different kinds of frameworks are being developed. The article by Di Zio and Guarnera is most closely related to the traditional score function approach and is important because it offers a statistical framework from which the local score function can be derived.

Di Zio and Guarnera base their approach on a so-called contamination model. In this contamination model, they posit a model for the true data and a separate model for the error mechanism. In their application, Di Zio and Guarnera assume a multivariate normal model for the true data and an error mechanism where only a proportion of the data is contaminated with an additive error, which in their study is also assumed to be normally distributed. Such an error mechanism, where only part of the observed units is affected by errors, is typical for economic surveys at NSIs.

The statistical framework for selective editing developed by Di Zio and Guarnera automatically generates a local score function: the combined use of a model for the true data and a model for the error mechanism allows the derivation of a score function that can be interpreted as an estimate for the error affecting the observed data. This in turn allows the use of the score function to select a set of units for manual review so that the expected remaining error in the data is below a user-specified threshold.

For economic surveys, a lognormal model for both the true data and the errors is often more realistic than a normal model. Di Zio and Guarnera therefore extend their approach to deal with lognormally distributed data and errors. The approach by Di Zio and Guarnera

can, in principle, be developed further by assuming different distributions for the true data or the errors. Their approach allows them to use auxiliary variables unaffected by errors. Finally, they extend their model so it can deal with missing values in the data. The use of auxiliary variables and the ability to deal with missing values make the approach more applicable for practical situations.

The approach can be used when only data from the current data set to be edited are available. In this case the editing approach should be considered as output editing, since a substantial part of the data from the current survey are then needed to estimate the model parameters. One can also estimate the model parameters using data from a previous period of the survey. In that case the approach can be used as an input editing approach.

The article by Arbués et al. has an even more ambitious goal. Their aim is not just to select and prioritise records, but to do this in an optimal way. In a sense, this could even be seen as a change of paradigm. In past implementations of selective editing approaches, NSIs were not overly worried about possibly selecting too many records. The focus lay on prioritising the records to be edited, and then the staff actually doing the editing were relied on not to edit too many records. This approach hence relied on the expert judgement of the staff involved. In the approach by Arbués et al., such expert judgement is no longer required. The approach automatically identifies which records are to be edited and the order in which they are to be edited.

Arbués et al. aim to minimise the number of records for manual review. To this end they develop a generic optimisation problem. Depending on the availability or non-availability of all observed data for the current survey, this generic optimisation problem gives rise to two different versions. If not all observed data of the current survey are available, they derive a stochastic optimisation problem. In this case, the approach may be classified as input editing. If all observed data of the current survey are available, they derive a combinatorial optimisation problem. In this case, the approach may be classified as output editing.

Similarly to Di Zio and Guarnera, Arbués et al. use models for the true data and the errors. Arbués et al. combine these models in a so-called observation-prediction model, that is, a multivariate statistical model for the true data and the measurement errors. By setting user-specified bounds on loss functions, such as the modelled mean squared error or bias of the survey estimators, the developed approach allows Arbués et al. to find the optimal set of units for manual review. As usual, the costs are measured by assuming a fixed amount per record to be edited. The approach can, in principle, be developed further to differential costs for different (classes of) records.

By extending their approach Arbués et al. not only succeed in selecting units for manual review, but also in prioritising these units. This is especially useful when time or resources run out before all units in the optimal set are edited, or conversely, when one decides not to limit oneself to only the optimal set of units after all and interactive editing simply continues until either time or resources run out. In a sense, when the approach of Arbués et al. is used to prioritise units, it leads to a kind of score function again, albeit an implicit score function with complicated coefficients.

As Arbués et al. point out, their approach is reminiscent of the Fellegi-Holt approach used in automatic editing. In both approaches an optimisation model is developed. In the Fellegi-Holt approach, the aim is to minimise the number of fields to change in a certain

record so that it will satisfy all edits. In the approach by Arbués et al., the aim is to minimise the number of records to be edited manually so that certain loss functions, such as the modelled mean squared error, are below upper bounds.

The approach by Arbués et al. leads to an optimal selection of records to edit manually, which is obviously very desirable for an NSI. However, everything comes at a price. In this case, the price to be paid seems to be the higher complexity. The approach by Arbués et al. may be more sensitive to misspecification of the model(s) than a traditional score function approach, even advanced forms as those by Di Zio and Guarnera. A misspecified model will lead to the wrong records being selected for manual editing. This is not a major problem if one edits more records than just the optimal set. It may be a problem when one limits the manual editing strictly to the optimal set.

Another practical problem of the increased complexity of the optimisation approach is that it may be harder to understand for staff applying it in practice than a traditional score function.

A completely different kind of framework is offered by Pannekoek et al. Whereas the frameworks offered by Di Zio and Guarnera and Arbués et al. are both statistical in nature, the framework offered by Pannekoek et al. is focused on processes.

Pannekoek et al. take the point of view that as many records as possible should be edited automatically. Only data that are influential and cannot be treated automatically without jeopardising data quality should be edited manually. They point out that it is useful to distinguish between systematic errors and nonsystematic (random) errors, as some kinds of generic systematic errors, such as unit measure errors, simple typing errors and sign errors, can often be corrected quite easily in an automatic manner.

Pannekoek et al. break down the statistical data editing process into a taxonomy of subprocesses, which they refer to as statistical or data editing functions, and discuss automatic editing in terms of these statistical functions. Examples of such statistical functions are verification functions that verify edit rules or compute quality indicators and selection functions that select a record or field for further treatment. Not all of these statistical functions can be carried out automatically while guaranteeing sufficient data quality. For those that cannot, human interaction remains necessary. Selective editing is a necessary step to identify the records or fields for which manual editing is required. The taxonomy allows NSIs to decide which statistical functions can be handled automatically and for which statistical functions manual review is needed.

Such a breakdown of the statistical data editing process into statistical functions also facilitates the development of reusable software components for the statistical data editing process, which leads to lower development and maintenance costs. It identifies for which statistical functions one should, or at least could, develop reusable software modules. Finally, the breakdown also enables the identification of which of these modules should be able to communicate with one other by passing data and metadata, in the form of input and output parameters. This allows one to easily connect the modules, and thus quickly build an entire editing system in a "plug & play" manner for a certain survey.

The ideas presented in the article by Pannekoek, et al. are closely related to using an architectural framework, which in turn is an instrument for achieving a higher degree of standardisation with respect to methods, processes and software tools (see e.g., Struijs et al. 2013).

6. Future Directions of Selective Editing Research

With the introduction of the frameworks for selective editing in this issue of the Journal of Official Statistics an important step forward has been taken. However, this does not mean that research on selective editing should be considered complete. So what are the main research topics in selective editing for the near future?

In my opinion, the most important research question for the near future is: how do we apply the developed frameworks for statistical editing in practice? Important practical questions here are:

- Are staff able to apply the frameworks correctly in practice?
- Do they trust the results of the frameworks or do they tend to overrule the results of the selective editing frameworks with the results of their own analyses?
- If staff are not able to apply the frameworks correctly, how can we support them? Should we modify the frameworks so they become easier to apply, or should we provide more training?
- If staff overrule the results of the selective editing frameworks with the results of their own analyses, does this mean we should improve the frameworks, or does this mean we should pay more attention to convincing staff to trust the results of these frameworks?

Another practical aspect is how to estimate the model parameters of the approaches by Di Zio and Guarnera and, especially, Arbués et al., described in this issue. The optimal situation would be to use a double data set with raw values and true values, or good approximations of the true values such as values edited according to a "gold standard", to estimate these model parameters. However, (a good approximation of) such an optimal situation usually only exists when one starts using selective editing for the first time for a certain survey. After that one usually only has data edited by means of a selective editing approach from a previous period and raw data from the current period. The approach by Di Zio and Guarnera is able to use only data from the current period. The approach by Arbués et al. may need to be extended.

Di Zio and Guarnera and Arbués et al. propose two different statistical frameworks for selective editing. The framework by Arbués et al. is the more ambitious of the two. It seems more complex to apply, but it potentially offers more benefits to the NSI. It is an open question at the moment which of the two frameworks, if any, will eventually prevail for a given survey.

The current frameworks, including the framework by Pannekoek et al., that is also described in this issue, have all been designed with traditional survey data in mind. An important research topic for the near future is the extension of the frameworks to administrative data and Big Data. Groves (2011) distinguishes between "designed-data", that is, data that have been collected especially for statistical purposes by the NSI itself, and "organic data", that is, data – in most cases electronic data – that somehow grow by themselves. Examples of organic data given by Groves (2011) are Twitter that generates tweets continuously, traffic cameras counting cars and scanners collecting information on purchases. Survey data are designed-data, Big Data are generally organic data, and administrative data are usually somewhere in between.

Developing selective editing and other editing techniques for organic data is much harder than for designed-data. The population (if any), concepts (if any), definitions of variables (if any) underlying organic data are generally unknown to the NSI, whereas they are known for designed-data. For organic data it is much more difficult to know what can be anticipated than for designed-data. A score function with "anticipated" values for organic data is therefore much harder to construct.

With respect to the roles of automatic editing and selective editing, there are two competing points of view.

(1) According to one point of view, automatic editing should be the most important way of editing used for the vast majority of records. Only for those records for which automatic editing cannot provide an acceptable solution, one should resort to interactive editing. Pannekoek et al. in this issue seem to adhere to this point of view. When taking this point of view to the extreme, there is no selection of records for interactive editing at all, except in exceptional cases.
(2) The other point of view is that selective editing is the most important part of the editing process. Once the records selected for manual review have been edited interactively, it does not really matter if (and how) the other records are edited. Granquist (1995, 1997) seems to adhere to this point of view. When taking this point of view to the extreme, automatic editing is only used for "cosmetic" purposes, namely just to ensure that edits are satisfied.

Only time can tell which of these point of views will become the dominant one. In practice the truth is likely to lie in the middle, and the "best" process will probably involve a bit of selective editing and a bit of automatic editing.

The final research topic I want to mention is a research topic for statistical editing in general. This topic has been mentioned since the 1960s. In those days, people already recognised that detecting and correcting errors is not the most important aspect of editing. For instance, Pritzker et al. (1965) observe that a more useful aspect of statistical data editing is, or in any case should be, to identify error sources or problem areas of the survey. As Granquist and Kovar (1997. p. 430) say about statistical data editing: "its more productive role lies in its ability to provide information about the quality of the collected data and thus form the basis for future improvement of the whole survey process".

According to Granquist (1984), statistical data editing has the following three goals:

- Identify and collect data on problem areas, and error causes in data collection and processing, producing the basics for the (future) improvement of the survey vehicle.
- Provide information on the quality of the data.
- Identify and handle concrete important errors and outliers in individual data.

In order to achieve these goals, the focus of statistical data editing should be shifted from detecting and correcting errors to obtaining more knowledge of the sources of errors arising in the data. This information can subsequently be used to further improve future versions of the survey. As noted by Granquist (1997, p. 385): "Editing should be considered a part of the total quality improvement process, not the whole quality process". Editing should be a coherent step in the chain of processes from data collection up to estimation and dissemination of the final results.

Much work has been done over the past decades on statistical data editing in general and selective editing in particular. Despite all the hard and clever work done, the more general, and likely more productive, goals of statistical data editing mentioned by Granquist (1984) have thus far proven very difficult to achieve in practice.

7. References

Bakker, B. (2011). Micro-integration. Statistical Methods 201108, Statistics Netherlands.

Banff Support Team (2008). Functional Description of the Banff System for Edit and Imputation. Technical Report, Statistics Canada.

Barcaroli, G., Ceccarelli, C., Luzi, O., Manzari, A., Riccini, E., and Silvestri, F. (1995). The Methodology of Editing and Imputation of Qualitative Variables Implemented in SCIA. Internal Report, Istituto Nazionale di Statistica, Rome.

De Waal, T. (1996). CherryPi: A Computer Program for Automatic Edit and Imputation. UN/ECE Work Session on Statistical Data Editing, 4–7 November, Voorburg.

De Waal, T. (2001). SLICE: Generalised Software for Statistical Data Editing. Proceedings in Computational Statistics, J.G. Bethlehem and P.G.M. Van der Heijden (eds). New York: Physica-Verlag, 277–282.

De Waal, T. and Coutinho, W. (2005). Automatic Editing for Business Surveys: an Assessment for Selected Algorithms. International Statistical Review, 73, 73–102.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New York: John Wiley and Sons.

Farwell, K. (2005). Significance Editing for a Variety of Survey Situations. Paper presented at the 55th session of the International Statistical Institute, 5–12 April, Sydney.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Freund, R.J. and Hartley, H.O. (1967). A Procedure for Automatic Data Editing. Journal of the American Statistical Association, 62, 341–352.

Granquist, L. (1984). Data Editing and its Impact on the Further Processing of Statistical Data. In Workshop on Statistical Computing, 12–17, Budapest.

Granquist, L. (1990). A Review of Some Macro-Editing Methods for Rationalizing the Editing Process. Proceedings of the Statistics Canada Symposium, 225–234.

Granquist, L. (1995). Improving the Traditional Editing Process. In Business Survey Methods, B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott (eds). New York: John Wiley & Sons, 385–401.

Granquist, L. (1997). The New View on Editing. International Statistical Review, 65, 381–387.

Granquist, L. and Kovar, J.G. (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality, L.E. Lyberg, P. Biemer, M. Collins, E.D. De Leeuw, C. Dippo, N. Schwartz and D. Trewin (eds). Hoboken, NJ: Wiley Series in Probability and Statistics, Wiley, 416–435. DOI: http://www.dx.doi.org/10.1002/9781118490013.ch18

Groves, R.M. (2011). Three Eras of Survey Research. Public Opinion Quarterly, 75, 861–871. DOI: http://www.dx.doi.org/10.1093/poq/nfr057

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistics, 19, 177–199.

Hedlin, D. (2008). Local and Global Score Functions in Selective Editing. UN/ECE Work Session on Statistical Data Editing, 21–23 April, Vienna.

Hidiroglou, M.A. and Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12, 73–83.

Kovar, J. and Whitridge, P. (1990). Generalized Edit and Imputation System; Overview and Applications. Revista Brasileira de Estadistica, 51, 85–100.

Latouche, M. and Berthelot, J.-M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. Journal of Official Statistics, 8, 389–400.

Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. Journal of Official Statistics, 10, 437–447.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243–253.

Nordbotten, S. (1955). Measuring the Error of Editing the Questionnaires in a Census. Journal of the American Statistical Association, 50, 364–369.

Nordbotten, S. (1963). Automatic Editing of Individual Statistical Observations. Statistical Standards and Studies No. 2. UN Statistical Commission and Economic Commission of Europe, New York.

Pritzker, L., Ogus, J., and Hansen, M.H. (1965). Computer Editing Methods–Some Applications and Results. Bulletin of the International Statistical Institute, Proceedings of the 35th Session, Belgrade, 395–417.

Pullum, T.W., Harpham, T., and Ozsever, N. (1986). The Machine Editing of Large-Sample Surveys: The Experience of the World Fertility Survey. International Statistical Review, 54, 311–326.

Pursey, S. (1994). Current and Future Approaches to Editing Canadian Trade Import Data. In proceedings of the Survey Research Methods Section, American Statistical Association, 105–109.

Struijs, P., Camstra, A., Renssen, R., and Braaksma, B. (2013). Redesign of Statistics Production within an Architectural Framework: The Dutch Experience. Journal of Official Statistics, 29, 49–71.

Todaro, T.A. (1999). Overview and Evaluation of the AGGIES Automated Edit and Imputation System. UN/ECE Work Session on Statistical Data Editing, 2–4 June, Rome.

Van de Pol, F. and Bethlehem, J. (1997). Data Editing Perspectives. Statistical Journal of the United Nations ECE, 14, 153–171.

Winkler, W.E. and Draper, L.A. (1997). The SPEER Edit System. In Statistical Data Editing (Volume 2); Methods and Techniques, 51–55, Geneva: United Nations.

Winkler, W.E. and Petkunas, T.F. (1997). The DISCRETE Edit System. In Statistical Data Editing (Volume 2); Methods and Techniques, 56–62, Geneva: United Nations.

Received June 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 489–510, http://dx.doi.org/10.2478/jos-2013-0037

An Optimization Approach to Selective Editing

Ignacio Arbués[1], *Pedro Revilla*[1], *and David Salgado*[1]

We set out two generic principles for selective editing, namely the minimization of interactive editing resources and data quality assurance. These principles are translated into a generic optimization problem with two versions. On the one hand, if no cross-sectional information is used in the selection of units, we derive a stochastic optimization problem. On the other hand, if that information is used, we arrive at a combinatorial optimization problem. These problems are substantiated by constructing a so-called observation-prediction model, that is, a multivariate statistical model for the nonsampling measurement errors assisted by an auxiliary model to make predictions. The restrictions of these problems basically set upper bounds upon the modelled measurement errors entering the survey estimators. The bounds are chosen by subject-matter knowledge. Furthermore, we propose a selection efficiency measure to assess any selective editing technique and make a comparison between this approach and some score functions. Special attention is paid to the relationship of this approach with the editing fieldwork conditions, arising issues such as the selection versus the prioritization of units and the connection between the selective and macro editing techniques. This approach neatly links the selection and prioritization of sampling units for editing (micro approach) with considerations upon the survey estimators themselves (macro approach).

Key words: Selective editing; optimization; observation-prediction model; selection efficiency measure.

1. Introduction

Data editing is a crucial step in the survey statistics production process. It impinges on several dimensions of survey quality such as accuracy, timeliness, response burden or cost effectiveness. This production phase comprises both the detection and treatment of nonsampling errors, mainly of nonresponse and measurement errors. Over time, a typology of errors has been developed, identifying systematic errors, random errors, influential errors, outliers, inliers or missing values, not to mention particular errors within these classes as measurement unit errors or rounding errors. This diversity has given rise to different techniques and algorithms to detect and treat them, such as interactive editing, automatic editing, selective editing, macro editing, and so on (see De Waal et al. 2011 for a comprehensive overview). Nowadays it is widely accepted that no single technique can

[1] D.G. Metodología, Calidad y TIC, Instituto Nacional de Estadística, Paseo de la Castellana, 28071 Madrid, Spain. Emails: ignacio.arbues.lombardia@ine.es, pedro.revilla.novella@ine.es, and david.salgado.fernandez@ine.es

Acknowledgments: We acknowledge graphics computing support from S. Saldaña. We are indebted to C. Pérez-Arriero and M. Herrador for their invaluable suggestions regarding the selection efficiency measure. The authors are grateful to T. de Waal, M. Di Zio, U. Guarnera, J. Pannekoek, M. van der Loo and S. Scholtus for comments and suggestions. We express special thanks to L.-C. Zhang for invaluable suggestions to improve the readability of the article.

deal with all kinds of errors. Thus they must be conveniently combined in so-called editing and imputation (E&I henceforth) strategies, specifically designed and fine-tuned for a given survey.

Selective editing focuses upon influential errors so that a selection of influential units is performed to thoroughly treat their errors (mostly with interactive editing), underlining the importance of recognizing and analyzing their source in order to prevent them when the survey is conducted on future occasions (Granquist 1997). In the last two decades, this editing modality has been recognized as a key element in E&I strategies. However, its principles are heuristics. By and large, selective editing comprises four stages (Lawrence and McKenzie 2000), namely (i) the construction of anticipated values $\hat{y}_k$ for each sample unit k according to an editing model; (ii) the construction of local score functions; (iii) the construction of a global score function; and (iv) the choice of cut-off values below which no further unit is selected. In general terms, the rationale is that those questionnaires k with a large discrepancy between the anticipated values $\hat{y}_k$ and the reported values y_k will be selected.

As a first general remark, our proposal can be succinctly described using the recent taxonomy of data editing functions by Pannekoek et al. (in this issue). They identify six types of editing tasks, called editing functions, according to the accomplishment of either error detection only (as data quality verification or field/record selection) or also including error treatment. These six editing functions are (i) rule checking, (ii) compute scores, (iii) field selection, (iv) record selection, (v) amend observations, and (vi) amend unit properties (see Pannekoek et al. in this issue for details). In this context, our proposal is to be understood as a record selection editing function.

We set out two general principles to approach selective editing (Arbués et al. 2012b). In keeping with Latouche and Berthelot (1992), who stated that "*in the development of an effective recontact and follow-up strategy, we have to minimize the amount of resources used without affecting the overall data quality and timeliness of the survey*", we claim that

i) editing must minimize the amount of resources deployed to recontacts, follow-ups and interactive tasks, in general;
ii) data quality must be ensured.

This framework is ample enough to give room to the preceding score function approach, but its rigorous derivation seems difficult to us. In this article we propose a mathematical translation of these principles into a general optimization problem, whose solution is the selection of units. In our formulation, interactive editing resources are tantamount to the number of selected questionnaires, whereas data quality is reduced to the accuracy of estimators. Thus a general optimization approach is to minimize the number of selected units, subjected to bounds on loss functions defined for a chosen number of variables of interest. These loss functions may be targeted at the bias, mean squared error (MSE), variance or other measures of the estimation uncertainty. They may be heuristic in nature, such as the so-called pseudo-bias related measures traditionally used for score functions, or they may be explicitly derived under some measurement-error models that are suitable for the data. One example is the contamination model (Di Zio and Guarnera in this issue), which is specified in terms of the full distribution of the true data and the conditional distribution of the observations given the true data.

Two versions of the optimization problem are provided, corresponding to the two typical scenarios for the implementation of selective editing. In the first case, selection is carried out unit by unit, in such a way that whether a given unit is selected or not does not depend on the selection of the other units. This mode of execution is suitable for input editing, where in principle the selection can be made in real time on arrival of each questionnaire. We refer to this as the stochastic optimization problem, because the real-time performance of the solution can only be established with respect to hypothetical repetitions of the selection process. In the second case, selection is carried out jointly for all (or a group of) units. This mode of execution is suitable for output (or macro) editing, which takes place at a later stage of the data collection after a sufficient number of observations have become available. We refer to this as the combinatorial optimization problem, where the performance of the solution can be established conditional on the actual sample observations under some specified measurement-error model.

Selection of units does not produce an order of priority by which the units are sorted according to their respective "urgency" to be edited. But prioritization of units is helpful for coping with the contingency of editing fieldwork. It is intrinsically related to selection since it should be possible in some sense to regard the highest prioritized unit as the optimal selection of a single unit, the second highest prioritized unit as the optimal selection of a single unit given that the highest prioritized unit has been selected, and so on. The combinatorial optimization problem can be adapted to yield prioritization. Not only is this a useful variation for practice, but sometimes it is theoretically necessary for obtaining a unique optimization solution, as we shall explain.

To perform a comparison with any other selective editing technique, we propose a selection efficiency measure. The rationale of this measure is to choose as an input the number of units to select and to compare our selection with an averaged random selection of this number of units. The comparison is based on the reduction of the absolute relative pseudo-bias of the survey estimators. In our view, the sooner the influential units are selected (hence the faster the reduction of the absolute relative pseudo-bias), the more efficient the technique will be. We perform a comparison with some score functions in the literature (Latouche and Berthelot 1992) using real data from the Spanish Industrial Turnover Index (ITI) and Industrial New Orders Received Index (INORI) survey.

The article is organized as follows. In section 2 we formulate the generic optimization problem as a mathematical translation of the above two principles. After fixing the notation and setting out the problem in general terms in Subsection 2.1, we show how the choice of the actual information used in this problem drives us either to a stochastic optimization version (Subsection 2.2) or to a combinatorial optimization version (Subsection 2.3). In Section 3 we show the general principles of the construct of any observation-prediction model, as well as a general proposal for continuous variables. In Section 4 we deal with the editing fieldwork and show how to choose the bounds and how to go from the selection to the prioritization of units under the combinatorial optimization approach. In Section 5 a selection efficiency measure is proposed and a comparison with several score functions is carried out using real data from the Spanish ITI and INORI survey. Finally we include an ample discussion in Section 6 in an attempt to assess this proposal in the current framework of selective editing with score functions.

2. The Optimization Problem

Before identifying the variables, the objective function and the restrictions of our optimization problem, we need to introduce the following notation. The sampling design according to which a probability sample s is selected will be denoted by $p(\cdot)$. The sample size will be denoted by n and the corresponding sampling weights by w_{ks}. The sample dependence of the sampling weights implicitly assumes that they do not need to be the design weights. For example, in a ratio estimator of the form $\hat{Y}^{rat} = X \cdot \frac{\hat{Y}^{HT}}{\hat{X}^{HT}}$, where x is a known auxiliary variable from the sampling frame, $X = \sum_{k \in U} x_k$ is a known population total, and $\hat{Y}^{HT} = \sum_{k \in s} \frac{y_k}{\pi_k}$ (analogously for $\hat{X}^{HT}$) stands for the Horvitz-Thompson estimator of the population total $Y = \sum_{k \in U} y_k$, the sampling weights are given by $w_{ks} = \frac{X}{\hat{X}^{HT}} \frac{1}{\pi_k}$, where π_k is the first-order inclusion probability for unit k. More complex situations are embedded under this notation. The true, observed and edited values of a variable $y^{(q)}$, $q = 1, \ldots, Q$ (for ease of notation we drop the superscript (q) hereafter except when strictly necessary), for unit k will be denoted, respectively, by y_k^0, y_k and y_k^*. We assign a binary variable $r_k \in \{0, 1\}$ to each unit k to indicate whether it is selected ($r_k = 0$) or not ($r_k = 1$). The vector $\mathbf{r} = (r_1, \ldots, r_n)^t$ for the whole sample will be referred to as the *selection strategy*. The counterintuitive assignment allows us to relate the preceding three values by the equation $y_k^*(\mathbf{r}) = (1 - r_k) \cdot y_k^0 + r_k \cdot y_k$, where we have made explicit the dependence of the edited values upon the selection strategy. Note that we are implicitly assuming that the editing work drives us from the observed to the true values. If we denote the corresponding measurement error by $\boldsymbol{\epsilon}_k = y_k - y_k^0$, then we can write $y_k^*(\mathbf{r}) = y_k^0 + r_k \boldsymbol{\epsilon}_k$. Note that these edited values are in fact those to be plugged into the survey estimators at this point of the E&I strategy. That is, if we are to estimate the population domain total $Y_{U_d} = \sum_{k \in U_d} y_k^0$ (for ease of notation we will drop the subscript U_d hereafter), then we denote the corresponding chosen estimator by $\hat{Y}^*(\mathbf{r}) = \sum_{k \in s_d} w_{ks} y_k^*(\mathbf{r})$. However, note that this estimator will not be the final estimator after the whole E&I strategy has been executed. Some later procedures such as weight adjustment or outlier treatment may follow. The selection of units proposed herein divides the sample into a critical and a noncritical stream, the treatments of which are decided by the statistician. We will restrict ourselves to population totals and linear estimators. All auxiliary covariates not included in the questionnaire for unit k will be denoted by $\mathbf{x}_k$.

So far the preceding variables are numeric. To use statistical modelling techniques, we promote these numeric variables to random variables according to a model m in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. As usual, this promotion will not be specifically indicated in the notation, except for the selection strategy, so that $\mathbf{R}$ will denote the random selection strategy, and $\mathbf{R}(w) = \mathbf{r}$, with $w \in \Omega$, will be a particular numeric realization called the *selection*. A predicted value of variable y_k according to the chosen model m will be denoted by $\hat{y}_k$. Note that the statistical model m embraces all promoted random variables different from the probability sample s itself. When random variables are used in survey estimators, we write indistinctly $\hat{Y}^0 = \sum_{k \in s_d} w_{ks} y_k^0$, $\hat{Y} = \sum_{k \in s_d} w_{ks} y_k$ and $\hat{Y}^*(\mathbf{R}) = \sum_{k \in s_d} w_{ks} y_k^*(\mathbf{R})$ for the survey estimators targeted at Y. We will denote by $\mathbf{Z}$ the set of random variables actually used by the statistician to select the units in the E&I strategy.

In particular, we will consider two options, namely, either $\mathbf{Z} = \mathbf{Z}^{long} \equiv s$ or $\mathbf{Z}^{long} \equiv \{s, \mathbf{X}\}$ for the stochastic problem (see below for the difference) or $\mathbf{Z} = \mathbf{Z}^{cross} \equiv \{s, \mathbf{X}, \mathbf{Y}\}$

for the combinatorial version. When this cross-sectional information is restricted to unit k, we shall write accordingly $\mathbf{Z}_k^{cross} = \{s, \mathrm{x}, y_k\}$. The use of information is represented as conditioning upon the corresponding random variables. The auxiliary covariates $\mathbf{X}$ are chosen by the statistician according to the chosen statistical model to be used in the problem (see below). They play a similar role to the auxiliary variables in the sampling design or the known auxiliary variables in the weight calibrating process. Indeed, they may coincide partially or totally with these auxiliary variables used in other parts of the estimation process.

2.1. The General Optimization Problem

As stated in the introduction, we want to minimize the number of questionnaires to edit provided that the chosen loss functions of the survey estimators $\hat{Y}^*$ targeted at the population total Y are bounded. To formally set up the optimization problem we need (i) the variables, (ii) the function to optimize, and (iii) the restrictions. Apart from identifying these elements, it is important to show how the available information enters into the formulation of the problem.

The ultimate variables are the selection strategy $\mathbf{r}^T = (r_i, \ldots, r_n)$ for the sample units $s = \{1, \ldots, n\}$, where $r_k = 0$ if the unit k is selected and $r_k = 1$ otherwise. However, since the measurement error $\boldsymbol{\epsilon}_k = y_k - y_k^0$ is conceived to be random in nature conditional on the realized sample s, and given the available information $\mathbf{Z}$ chosen to make the selection of units, this selection can vary depending on the realized $\mathbf{y}$, $\mathbf{y}^0$ and $\mathbf{Z}$. Thus let $\mathbf{R}$ denote the stochastic selection strategy so that (i) $\mathbf{R}(w) = \mathbf{r}$ is a realized selection and (ii) $\mathbb{E}_m[\mathbf{R}|\mathbf{Z}]$ is the vector of probabilities of nonselection under the specific model m given the chosen information $\mathbf{Z}$. The objective function to optimize, given the information $\mathbf{Z}$, is then written as $\mathbb{E}_m[\mathbb{1}^T\mathbf{R}|\mathbf{Z}]$, whose maximization amounts to minimizing the number of selected units.

The constraints derive from the application of a loss function to the survey estimators. Let us concentrate on the two loss functions most used in practice, namely the absolute loss $L = L^{(1)}(a, b) = |a - b|$ or the squared loss $L = L^{(2)}(a, b) = (a - b)^2$. Then it is straightforward to prove (see appendix A) that $\mathbb{E}_m[L^{(r)}(\hat{Y}^*(\mathbf{R}), Y)|\mathbf{Z}] \leq \eta$ warrants $\mathbb{E}_{pm}[L^{(r)}(\hat{Y}^*(\mathbf{R}), Y)] \leq \left(\eta^{1/r} + \mathbb{E}_{pm}^{1/r}\left[L(\hat{Y}^0, Y)\right]\right)^r$, where $O(\cdot)$ stands for the well-known big O. In other words, each constraint controls the loss of accuracy in terms of the chosen loss function L due to nonselected units, up to sampling design variability.

For these loss functions, each constraint can always be written as a bound on a quadratic form, denoted by $\mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$ (see Appendix A). Particular forms suitable for the stochastic and combinatorial problems will be explained in Subsection 2.2 and 2.3. The $n \times n$ matrix Δ specifies the potential losses at the unit level. Measures of bias and/or MSE seem natural in practice and they stem from the choice of the absolute or the squared loss function respectively. These measures can be heuristic in nature, such as the pseudo-bias for traditional score functions, or explicitly derived under some appropriate measurement-error model. In particular, non-zero off-diagonal terms of Δ allow for cross-unit terms to be included in the "overall" loss.

The choice of the matrix Δ is naturally linked to the choice of the loss function L, hence the term loss matrix (see Appendix A for details). Thus, if Δ is diagonal with entries

$|w_{ks}\epsilon_k|$, then we are choosing the absolute loss so that $\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0)|\mathbf{Z}]$ is also bounded by η (up to sampling design factors). This is targeted at the bias. Similarly, if $\Delta_{kl} = w_{ks}w_{ls}\epsilon_k\epsilon_l$, then we are choosing the squared loss so that $\mathbb{E}_m[(L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0)|\mathbf{Z}]$ is also equally bounded. In turn, this is targeted at the mean squared error. In both cases, model-based techniques using data from the current time period can be applied in the combinatorial version, whereas in the stochastic version we are obliged to resort to auxiliary information from other periods.

For instance, the (local) score for a given y-variable is usually conceived as the product of a "risk" component and an "influence" component. A generic measure can be given using a model-based approach. Let $p_k = P\left(y_k^0 \neq y_k|y_k\right)$, that is, the posterior probability that the true value is different from the observed one. Let $\tilde{\mu}_k = \mathbb{E}_m\left(y_k^0|y_k, y_k^0 \neq y_k\right)$, that is, the conditional expectation of the true value given that it is different from the observed one. Then, we have

$$\mathbb{E}_m\left(y_k^0|y_k\right) = (1 - p_k)y_k + p_k\tilde{\mu}_k \quad \text{and} \quad \delta_k = y_k - \mathbb{E}_m\left(y_k^0|y_k\right) = p_k(y_k - \tilde{\mu}_k)$$

It follows that $w_k\delta_k$ can be used to construct the local score of unit k with respect to y, which is the product of "risk" measured by p_k and "influence" measured by $w_k(y_k - \tilde{\mu}_k)$, where w_k can be the sample weight, for example. Di Zio and Guarnera (in this issue) derive such a measure under the contamination model, which is suitable for the combinatorial problem. For the stochastic problem, where scoring does not use observations other than the unit at hand, $\tilde{\mu}_k$ cannot be evaluated for the current sample data and instead information from preceding realizations of this survey or similar surveys must be used. It is customary to replace it with some reference value, such as y_k from a previous time point, giving rise to a pseudo-bias. Nor can the "risk" component be assessed properly, and some heuristics measure might be used, such as in the SELEKT approach of Statistics Sweden (see for example Lindgren 2011). The auxiliary information, which we exploit in the observation-prediction model (see Section 3), is fundamental.

The main difference between both versions arises when considering their actual application. The stochastic problem, supplemented by the assumption that ignores the cross-unit terms, allows the construction of score functions to be applied independently to each unit. The supplementary assumption amounts to considering these cross-terms more or less constant over time, hence playing no significative role in the selection. Conversely, the combinatorial problem needs a sufficient number of observations available to carry out the selection jointly for all units.

Taking into account the possibility of multiple constraints, we now arrive at the following general optimization problem:

$$\begin{aligned} [P_0] \quad & \max \ \mathbb{E}_m[\mathbf{1}^T\mathbf{R}|\mathbf{Z}] \\ & \text{s.t.} \quad \mathbb{E}_m[\mathbf{R}^T\Delta^{(q)}\mathbf{R}|\mathbf{Z}] \leq \eta_q, \quad q = 1, 2, \ldots, Q, \\ & \mathbf{R} \in \Omega_0 \end{aligned}$$

where Ω_0 denotes the admissible outcome space of $\mathbf{R}$, and q refers to the different constraints. Manipulation of Ω_0 creates extra flexibility for adoption. For instance, the problem can be recast for selection conditional on the units that have already been selected, by restricting Ω_0 such that certain R_k s are fixed at 0. The different constraints

may arise from the fact that there are multiple y-variables of interest, or the constraints may be directed at the different population domains even when there is only a single y-variable. In particular, the loss matrices $\Delta^{(1)}, \ldots, \Delta^{(Q)}$ may all be derived under a single multivariate model for the joint data, even when the bounds are marginally specified for each target quantity on its own.

Variations of the optimization problem stated above are possible, by either adopting a different function for optimization and/or different forms of constraints. For instance, maximization may be changed to minimization as long as suitable alterations of the selection variables and the loss functions are provided. Alternatively, one may for example use $w_k\delta_k$ in Δ but state the constraint as $\mathbb{E}_m[|\mathbf{R}^T\Delta\mathbf{R}|\,|\mathbf{Z}] \leq \eta$. We do not explicitly consider such variations of the problem in this article, but note that (i) it is possible to adapt the solutions presented below, should such variations be desirable in practice, and (ii) the expounded optimization approach can be carried out in the same spirit.

2.2. *The Stochastic Optimization Problem*

As stated above, the main assumption in this version of problem P_0 is neglecting the cross-unit terms in each constraint. Then these constraints can be rewritten as $\mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\mathrm{diag}(\Delta)|\mathbf{Z}]$. Furthermore, the distinction between $\mathbf{Z}^{long} = s$ and $\mathbf{Z}^{long} \equiv \{s, \mathbf{X}\}$ is a matter of choice. In the former case, the restrictions are required to be fulfilled only on average for all realizations of the survey, whereas in the latter case they are imposed on the current realization, given the realizations of preceding time periods. The deduced stochastic optimization problem is solved in Arbués et al. (2012a) by using the duality principle, the sample average approximation and the interchangeability principle. The solution resulting from this linear problem is given in terms of matrices $\boldsymbol{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$. This dependence on $\mathbf{Z}^{cross}$ may seem misleading, but only momentarily. Since this selection scheme is to be applied unit by unit upon receipt of each questionnaire, and no cross-sectional information except that regarding each unit k separately will be actually used, the formal conditioning upon $\mathbf{Z}^{cross}$ reduces effectively to conditioning upon the information $\mathbf{Z}_k^{cross} = \{s, \mathbf{x}, \mathbf{y}_k\}$ of each unit. Thus we write $\boldsymbol{M}^{(q)} = \mathbb{E}_m[\Delta|\mathbf{Z}^{cross}] = \mathrm{diag}\left(\mathbb{E}_m\left[\Delta_{kk}^{(q)}|\mathbf{Z}_k^{cross}\right]\right) = \mathrm{diag}\left(M_{kk}^{(q)}\right)$. On the other hand, in order to obtain the optimal Lagrange multipliers λ_q^* involved in the dual problem, a historic double-data set with raw and edited values is necessary. Putting it all together we arrive at the final solution, which only requires the diagonal entries of the matrices $\boldsymbol{M}^{(q)}$:

$$R_k = \begin{cases} 1 & \text{if} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} \leq 1, \\ 0 & \text{if} \quad \sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)} > 1. \end{cases} \tag{1}$$

Note that since the scheme is "trained" on the historic data, the evaluation of $M_{kk}^{(q)}$ given the observations in the current sample necessarily yields a pseudo-measure, regardless of the definition of the loss matrices.

This provides a score function for unit-by-unit selection. In the special case of $Q = 1$, unit k is selected provided $M_{kk} > 1/\lambda^*$, so that M_{kk} can be regarded as a single score and $1/\lambda^*$ as the threshold value. Equivalently, one may consider $\lambda^* M_{kk}$ as a "standardized"

score, in the sense that the threshold value is generically set to 1. The latter extends in a straightforward manner to the setting with multiple constraints, where each $\lambda_q^* M_{kk}^{(q)}$ is a standardized local score, and $\sum_{q=1}^{Q} \lambda_q^* M_{kk}^{(q)}$ is the standardized global score, with the generic global threshold value 1.

The global scoring derives from the linear structure of the dual problem and few variations are allowed without a substantial modification of problem P_0. As an exception, if a global score is initially envisaged as the weighted sum of local scores, then one may incorporate each weight into the constraint that generates the corresponding standardized local score to begin with.

The stochastic problem thus clarifies the fact that the performance of unit-by-unit selection can only be established over hypothetical repetitions of the selection process. At the end of each selection process, we have the realized selection strategy $\mathbf{r}$, and the realized loss $\sum_{k=1}^{n} r_k M_{kk}^{(q)}$, which can either be higher or lower than the specified bound η_q, for $q = 1, \ldots, Q$. Upon any hypothetical repetition of the selection process, however, y_k and y_k^0 will vary, and so will the corresponding $M_{kk}^{(q)}$ and r_k. It is over such hypothetical repetitions that the constraint $\mathbb{E}_m[\mathbf{R}\Delta^{(q)}\mathbf{R}|\mathbf{Z}] \leq \eta_q$ can possibly be satisfied, but not for each particular realization of the selection process.

2.3. The Combinatorial Optimization Problem

The combinatorial problem deals with the selection among all (or a group of) units. Cross-unit terms are now allowed and the information actually used is that given by the sample s, the auxiliary covariates $\mathbf{X}$ and the variables of interest $\mathbf{Y}$, that is by $\mathbf{Z} = \mathbf{Z}^{cross}$. Notice that all this information is available only after all questionnaires have been collected, thus it is only applicable as a form of output editing. It is easily proved that each constraint reduces to $\mathbb{E}_m[\mathbf{R}^T\Delta^{(q)}\mathbf{R}|\mathbf{Z}^{cross}] = \mathbf{r}^T\mathbf{M}^{(q)}\mathbf{r}$, where $\mathbf{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$, which can now be possibly evaluated under some measurement-error model. Consequently, it becomes possible to establish the performance of the realized selection strategy directly. The optimization problem can be rephrased as

$$
\begin{aligned}
[P_{co}(\mathbf{M}, \eta, \Omega_0)] \quad & \max \mathbf{1}^T\mathbf{r} \\
& \text{s.t.} \quad \mathbf{r}^T\mathbf{M}^{(q)}\mathbf{r} \leq \eta_q, \quad q = 1, 2, \ldots, Q, \\
& \mathbf{r} \in \Omega_0
\end{aligned}
$$

Note that a more direct derivation can be obtained by not promoting the selection strategy vector $\mathbf{r}$ to a random vector $\mathbf{R}$ when modelling the measurement errors.

This combinatorial problem is solved in two different forms using two greedy algorithms, which run in $n^4 \cdot Q$ and $n^3 \cdot Q$ times, respectively. The solution of both algorithms is not exact a priori but suboptimal with a good degree of approximation. The faster algorithm is noticeably less precise than the slower one. This lack of precision entails a small amount of overediting in practice, that is, more units than those optimally obtained will be selected. The fourth and third power dependence on n may appear discouraging for practical applications. However, firstly, the input size P in problem P_{CO} is actually $P = O(n^2)$, thus the algorithms run in $O(P^2)$ and $O(P^{3/2})$, which are acceptable speeds for combinatorial problems. Secondly, in practice the problem is intended to be

applied not to entire samples but to their breakdowns into publication cells, which are the figures upon which precision is called for (see Section 6). These heuristic algorithms locally search the optimum in each iteration until the current solution satisfies all the restrictions. To do this we introduce infeasibility functions $h_i(\mathbf{r})$ for each algorithm $i = 1$, 2 (see Salgado et al. 2012 for details) indicating whether a solution satisfies all the restrictions ($h(\mathbf{r}) = 0$) or not ($h(\mathbf{r}) > 0$). Both algorithms start from the initial solution $\mathbf{r} = 1$ and in each iteration select the next unit in a locally optimal way until all restrictions are satisfied. The infeasibility functions will also be used later when constructing the prioritization of units.

Finally, we can regard both versions as related to two different approaches to the problem of optimization under uncertainty (see e.g., Wets 2002). The combinatorial version is consistent with the wait-and-see approach, since it puts off all decisions until all the information is available. The stochastic version is, at least partially, a here-and-now approach, since the decision about the procedure or rule of selection (although not the selection itself) is made before the data collection.

3. The Observation-Prediction Model

To substantiate the constraints in both versions of the optimization problem, we need to compute the loss matrices $\mathbf{M}^{(q)} = \mathbb{E}_m[\Delta^{(q)}|\mathbf{Z}^{cross}]$ and to choose the bounds η_q. We now show how to undertake the former whereas the latter is dealt with in the next section.

To compute the loss matrices we make use of the standard model-based techniques, but not in a conventional way. Let us digress very briefly. When facing the editing tasks and, in particular, the selection of units, one resorts to the very best auxiliary information available at that precise moment. With full generality, this will comprise (i) the reported values of the variables of analysis $\mathbf{y}_k^{(t)}$ for the present ($t = T$) and preceding ($t < T$) time periods, (ii) the true values of these variables $\mathbf{y}^{(0,\,t)}$ for those edited units in the past $t < T$, (iii) and the values of auxiliary covariates $\mathbf{x}_k^{(t)}$ for all time periods. In the notation of preceding sections, we have $\mathbf{y}_k = \mathbf{y}_k^{(T)}$, $\mathbf{y}_k^0 = \mathbf{y}_k^{(0,T)}$ and $\mathbf{x}_k = \mathbf{y}_k^{(t_1)}, \mathbf{y}_k^{(0,t_1)}, \mathbf{x}_k^{(t_2)}$, with $t_1 < T$ and $t_2 \leq T$. Note that some of these values can be coincidentally equal (e.g., when the measurement error is null) and that $\mathbf{y}_k^0$ is only known after accomplishing the editing work. But this is not everything. We also know (at least we can know) a point prediction $\hat{y}_k$ for each y-variable based on these auxiliary variables. For instance, we can make use of a time series model $\left\{y_k^{(0,t)}\right\}_{t<T}$ to make a point prediction $\hat{y}_k^{(T)}$. Different choices arise depending on the amount and type of auxiliary information. These predictions will enter into the selection problem as auxiliary covariates, so that $\mathbf{x}_k = \mathbf{y}_k^{(t_1)}, \mathbf{y}_k^{(0,t_1)}, \hat{\mathbf{y}}_k^{(T)}, \mathbf{x}_k^{(t_2)}$, with $t_1 < T$ and $t_2 \leq T$.

Let us denote by m^* the auxiliary model used to make the predictions $\hat{\mathbf{y}}_k$, not to be confused with the measurement error model m considered throughout this paper. This measurement error model m is given as usual in terms of (i) the conditional distribution of the predicted values $\mathbf{y}$ upon the true values $\mathbf{y}^0$, and (ii) the distribution of $\mathbf{y}^0$ conditional on the available auxiliary information $\mathbf{X}$. To be specific, for a y-variable we will assume $y_k = y_k^0 + \epsilon_k^{obs}$ and $y_k^0 = \hat{y}_k + \epsilon_k^{pred}$. In other words, we are using the predicted value computed according to the auxiliary model m^* as an exogenous variable for the model regarding y^0. In this sense we refer to this proposal as an observation-prediction model.

Generalizing these ideas, let us consider

i) an observation model $\mathbb{P}_{obs|0}(\mathbf{y}|\mathbf{y}^0)$, that is, a conditional probability distribution for the observed values $\mathbf{y}$ given the true values $\mathbf{y}^0$;
ii) a prediction model $\mathbb{P}_{0|pred}(\mathbf{y}^0|\hat{\mathbf{y}})$, that is, a conditional probability distribution for the true values $\mathbf{y}$ given the predicted values $\mathbf{y}^0$ according to an auxiliary model m^*.

Now let us denote by $\mathbb{P}_{obs|pred}$ the probability distribution of $\mathbf{y}$ conditional on the predicted values $\hat{\mathbf{y}}$ and by $\mathbb{P}_{0|obs,pred}$ the probability distribution of the true values $\mathbf{y}^0$ conditional on the observed values $\mathbf{y}^{obs}$ and the predicted values $\hat{\mathbf{y}}$. Then by Bayes' theorem or a generalization thereof, we can write

$$\mathbb{P}_{0|obs,pred} = \frac{\mathbb{P}_{obs|0} \times \mathbb{P}_{0|pred}}{\mathbb{P}_{obs|pred}} \tag{2}$$

The product must be understood in a suitable generalized form when the distributions are completely general. As usual, if the probability distributions are absolutely continuous with density functions $f.(\cdot)$, Equation (2) can be easily recognized as

$$f_{0|obs,pred}(\mathbf{y}^0) = \frac{f_{obs|0}(\mathbf{y}|\mathbf{y}^0, \hat{\mathbf{y}}) f_0(\mathbf{y}^0|\hat{\mathbf{y}})}{\int_{\mathbb{R}^Q} f_{obs|0}(\mathbf{y}|\mathbf{y}^0, \hat{\mathbf{y}}) f_0(\mathbf{y}^0|\hat{\mathbf{y}}) \mathrm{d}\mathbf{y}^0}.$$

The discrete case also boils down to applying Bayes' theorem. Once we have the distribution $\mathbb{P}_{0|obs,pred}$, the loss matrices can be computed as

$$\mathbf{M}^{(q)} = \mathbb{E}_{0|obs,pred}[\Delta^{(q)}|S, \mathbf{Y}, \hat{\mathbf{Y}}]. \tag{3}$$

To illustrate this proposal, let us consider the following generic example with a continuous variable y. Let us define the observation model $y_k^{obs} = y_k^0 + \epsilon_k^{obs}$ and the prediction model $y_k^{obs} = \hat{y}_k + \epsilon_k^{pred}$, with the following specifications:

1. $\epsilon_k^{obs} = \delta_k^{obs} e_k$.
2. $e_k \simeq Be(p_k)$, where $p_k \in (0, 1)$.
3. $\left(\epsilon_k^{pred}, \delta_k^{obs}\right) \simeq N\left(\mathbf{0}, \begin{pmatrix} \nu_k^2 & \rho_k \sigma_k \nu_k \\ \rho_k \sigma_k \nu_k & \sigma_k^2 \end{pmatrix}\right)$.
4. ϵ_k^{pred}, δ_k^{obs} and e_k are jointly independent of $\mathbf{Z}_k^{cross}$.
5. e_k is independent of ϵ_k^{pred} and δ_k^{obs}.

These are equivalent to stating that unit k has a probability $1 - p_k$ of reporting a value without measurement error $\left(y_k = y_k^0\right)$ and, when reporting an erroneous value, the measurement error distributes as a normal random variable with zero mean and variance σ_k^2. On the other hand, the prediction error distributes as a normal random variable with zero mean and variance ν_k^2. Both errors distribute jointly as a bivariate normal random variable with correlation ρ_k. Reporting an erroneous value is independent of both types of errors.

For the time being let us assume that the parameters $\boldsymbol{\theta} = \left(p_k, \sigma_k^2, \nu_k^2, \rho_k\right)^T$ are known. Let us focus on the squared loss function. Then it is easy to prove (Arbués et al. 2012a) that

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)|s_k, y_k, \hat{y}_k\right] = \nu_k \cdot \frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k} \cdot \left(\frac{y_k - \hat{y}_k}{\nu_k}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right), \tag{4}$$

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)^2|s_k, y_k, \hat{y}_k\right] = \nu_k^2 \cdot \left(\frac{\sigma_k^2 + \rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^2 \cdot$$

$$\left[\frac{\sigma_k^2\left(1 - \rho_k^2\right)\left(\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k\right)}{\left(\sigma_k^2 + \rho_k\sigma_k\nu_k\right)^2} + \left(\frac{y_k - \hat{y}_k}{\nu_k}\right)^2\right] \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right),$$

$$\mathbb{E}_m\left[\left(y_k - y_k^0\right)\left(y_l - y_l^0\right)|s_k, y_k, \hat{y}_k\right] = \mathbb{E}_m\left[\left(y_k - y_k^0\right)|s_k, y_k, \hat{y}_k\right]\mathbb{E}_m\left[\left(y_l - y_l^0\right)|s_k, y_k, \hat{y}_k\right],$$

$$k \neq l,$$

where

$$\zeta_k(x) = \frac{1}{1 + \dfrac{1 - p_k}{p_k}\left(\dfrac{\nu_k^2}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}\right)^{-1/2} \exp\left(-\dfrac{1}{2}\dfrac{\sigma_k^2 + 2\rho_k\sigma_k\nu_k}{\sigma_k^2 + \nu_k^2 + 2\rho_k\sigma_k\nu_k}x^2\right)}.$$

Should we choose the absolute loss function, then, under the same hypotheses, we would have (see Appendix A):

$$\mathbb{E}_m\left[|y_k - y_k^0||s_k, y_k, \hat{y}_k\right] = \sqrt{\frac{2}{\pi}} \cdot \nu_k \cdot {}_1F_1\left(-\frac{1}{2}; \frac{1}{2}; -\frac{(y_k - \hat{y}_k)^2}{2\nu_k^2}\right) \cdot \zeta_k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right), \tag{5}$$

where ${}_1F_1(a; b; x)$ denotes the confluent hypergeometric function of the first kind.

The estimation of the parameters $\boldsymbol{\theta}$ depends on the scenario. For the stochastic problem, as before, we are obliged to use some reference values or heuristic measures. Once more we resort to the auxiliary information. Our choice depends very much on the amount and type of auxiliary information. From the historic double-data sets comprising τ past time periods (e.g., a fixed panel) we can compute

$$\hat{p}_k = \frac{1}{\tau}\sum_{t=1}^{\tau} I_{y_k^{(t)} \neq y_k^{(0,t)}},$$

$$\hat{\sigma}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)^2,$$

$$\hat{\nu}_k^2 = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)^2, \text{ where } \epsilon_k^{(t)} = \hat{y}_k^{(t)} - y_k^{(0,t)},$$

$$\hat{\rho}_k = \frac{1}{\tau - 1}\sum_{t=1}^{\tau}\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right)\left(\epsilon_k^{(t)} - \bar{\epsilon}_k\right).$$

In case of rotating panels or sampling designs with too short a continuity in the sample for a number of units, we are forced to make simplifying assumptions such as partitioning

the sample $s = \cup_{i=1}^{I} s_i$ and positing $\boldsymbol{\theta}_k = \boldsymbol{\theta}_i$ if $k \in s_i$. We can also adopt these assumptions for some of the parameters. The extreme case would $\boldsymbol{\theta}_k = \boldsymbol{\theta} = (p, \sigma^2, \nu^2, \rho)^T$ for all $k \in s$, which can be further supplemented with extra hypotheses such as $\rho = 0$.

On the other hand, for the combinatorial problem we do have (almost) the complete current sample so that we can make use of these data, although with important limitations. It is clear that it is impossible to estimate each $\boldsymbol{\theta}_k$ using only the current sample. We are obliged to make some simplifying assumptions, as above. In practice, however, it is advisable to use not only data from the current time period ($t = T$), but also from preceding periods ($t < T$). The stationarity across time periods of the response mechanism supports this course of action.

Alternatively, the contamination model by Di Zio and Guarnera (in this issue) is a relevant example of a model-based technique which uses exclusively data from the current period (except for the covariates for the model) to estimate the model parameters. The usage of statistical models to make the selection of units allows us to cherish the hope of extending this approach to qualitative and semicontinuous variables, thus paving the way for the use of selective editing in household surveys.

4. Fieldwork: Selection and Prioritization of Units

The problem is not completely specified until we choose the bounds η_q to formulate the optimization problem completely. The bound η on a given constraint $\mathbb{E}_m[\mathbf{R}^t \Delta \mathbf{R} | \mathbf{Z}] \leq \eta$ can be set either absolutely or relatively in terms of a chosen figure of merit or reference value. This can be, for example, the a priori variance used in the sampling design phase so that the constraint establishes a bound for the loss of accuracy as a fraction of the desired precision. The decision will necessarily involve some subject-matter knowledge.

So far, the formulation of the selective editing problem as an optimization problem is complete, providing a *selection* of units expressed by the solution $\mathbf{r}$. However, in practice having a selection of units must be confronted with the actual conditions of fieldwork. In particular, both controllability and availability of resources, such as person hours for example, are important issues in this respect. Given a particular selection, either we may run out of resources and cannot edit all selected units or we may finish the editing field work ahead of time and thus miss the opportunity to achieve better accuracy. In this sense it seems natural to have at our disposal a set of selections to optimize the actual use of resources. We achieve this by having a *prioritization* of units. Next we show how to prioritize units in the optimization approach. In Section 6 we discuss in more detail this issue of the selection/prioritization of units in relation with the fieldwork.

From the preceding sections it is clear that it does not make sense to prioritize units in the stochastic formulation. On the other hand, to prioritize units in the combinatorial version we propose combining different selections by choosing a sequence of appropriate values as bounds. The basic idea is to choose large initial bounds which drive us to select no unit, then to decrease the bounds until one unit is selected and to flag this unit for future selections. Then we again decrease the bounds until a new unit is selected and flagged for future selections. The procedure is repeated until all units have been flagged.

Let $f^{[k]} \subset s = \{1, \ldots, n\}$ denote the set of flagged units at iteration k and $\Omega_0^{[k]}$ the outcome space of the combinatorial problem at iteration k. For any given strategy vector $\mathbf{r}$

we denote by $\mathcal{I}^{-1}(\mathbf{r})$ the set of strategy vectors $\bar{\mathbf{r}}$ obtained from $\mathbf{r}$ transforming exactly a component 1 into 0. For example, $\mathcal{I}^{-1}((1,1,0)^T) = \{(0,1,0)^T, (1,0,0)^T\}$. Let h denote the infeasibility function used in the greedy algorithms (see Subsection 2.3).

The algorithm of prioritization reads as follows:

1. Set $f^{[0]} = \emptyset$, $\Omega_0^{[0]} = \{0,1\}^{\times n}$, $\mathbf{s}^{[0]} =$ and $\boldsymbol{\eta}^{[0]} = (\mathbf{s}^{[0]T}M^{(1)}\mathbf{s}^{[0]}, \ldots, \mathbf{s}^{[0]T}M^{(Q)}\mathbf{s}^{[0]})^T$.
2. FOR $k = 0$ TO $k = n$
   i. Set $\mathbf{s}^{[k+1]} = \arg\min_{\mathbf{s}\in\mathcal{I}^{-1}(\mathbf{s}^{[k]})}(h(\mathbf{s}))$. In case of multiple $\mathbf{s}^{[k+1]}$ choose one at random.
   ii. Set $l^* \in s$ such that $s_{l^*}^{[k+1]} \neq s_{l^*}^{[k]}$.
   iii. Set $f^{[k+1]} = f^{[k]} \cup \{l^*\}, \Omega_0^{[k+1]} = \Omega_0^{[k]} - \{\mathbf{s}^{[k]}\}$ and $\boldsymbol{\eta}^{[k+1]} = (\mathbf{s}^{[k+1]T}M^{(1)}\mathbf{s}^{[k+1]}, \ldots, \mathbf{s}^{[k+1]T}M^{(Q)}\mathbf{s}^{[k+1]})^T$.
3. FOR $k = 0$ TO $k = n$
   i. Set $\mathbf{r}^{[k]} = \arg\max[P_{co}(\mathbf{M}, \boldsymbol{\eta}, \Omega_0^{[k]})]$.
4. Set $\boldsymbol{s} = \sum_{k=0}^{n}\mathbf{r}_k$.

The vector $\mathbf{s}$ provides the prioritization: unit k must be edited in the s_kth place. Notice that steps 1 and 2 provide a sequence of bounds $\boldsymbol{\eta}^{[k]}$ and a sequence of outcome sets $\Omega_0^{[k]}$ which are used in step 3 to solve $n+1$ concatenated combinatorial problems. Two comments are in place here. On the one hand, in practice, Step 3 indeed reduces to the first point in Step 2 since $\mathbf{r}^{[k]} = \mathbf{s}^{[k]}$ because h is the infeasibility function of the optimization algorithm.

On the other hand, this invites us to reconsider the role of the infeasibility function in the prioritization of units: this depends on the choice of h. Should we choose, instead of the original infeasibility function $h_1(\mathbf{r}) = \sum_{q=1}^{Q}\left(\mathbf{r}^t M_{kl}^{(q)}\mathbf{r} - m_q^2\right)^+$ of algorithm 1, the function $h(\mathbf{r}) = \sum_{q=1}^{Q} w_q\left(\mathbf{r}^t M_{kl}^{(q)}\mathbf{r} - m_q^2\right)^+$, where $w_q \geq 0$ are positive weights expressing the different priority given to the accuracy of each variable $y^{[q)}$, we would arrive at a different prioritization. This can also be viewed more geometrically. To produce a sequence of bounds we begin by having no selected units, that is, by $\boldsymbol{\eta}_0 = (\mathbb{1}^t M^{(1)}\mathbb{1}, \ldots, \mathbb{1}^t M^{(Q)}\mathbb{1})^t$, and we need to produce a sequence of points in $\mathbb{R}^Q$ such that its final point is $\mathbf{0}$. There exist infinitely many possibilities (see Figure 1). In this context, the prioritization amounts to choosing a path from $\boldsymbol{\eta}_0$ to $\mathbf{0}$. This path expresses the priority which the statistician gives to the accuracy of the different estimators along the process of prioritization of units. The original infeasibility functions of the algorithms confer the same relevance on every estimator $\hat{Y}^{(q)}$.

5. A Selection Efficiency Measure: Comparison with the Score Function Approach

To make a comparison of the selection undertaken under any approach, we propose the following selection efficiency measure for an estimator $\hat{Y}$. Beforehand, we need a double data set with raw and edited values according to a gold standard so that when a unit is selected, its raw values are substituted by their corresponding edited counterparts, considered true. We will denote by $\hat{Y}^{sel}(n_{ed})$ the estimator obtained when n_{ed} questionnaires have been selected according to a selective editing technique *sel* and edited correspondingly. Note that $\hat{Y}^{sel}(n_{ed} = n) = \hat{Y}^0$. As a figure of merit for the

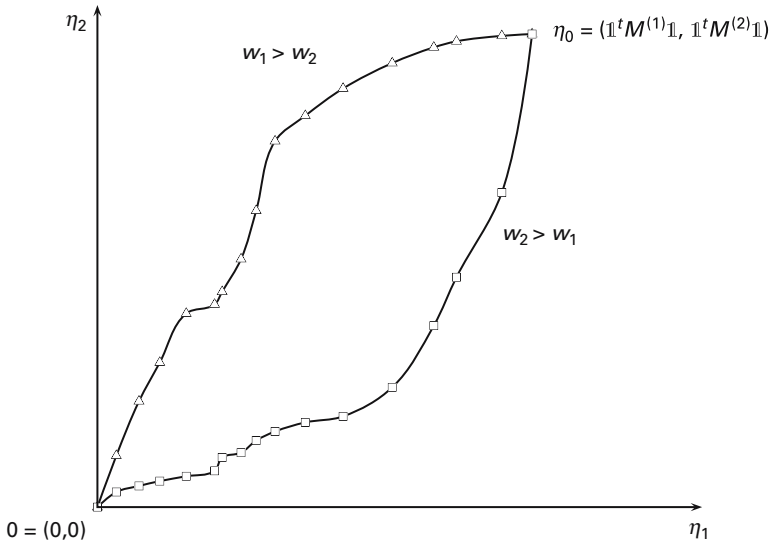

Fig. 1. Example of two different sequences of bounds with $Q = 2$ arising from different choices of the weights w_q.

selection of units we will focus upon the absolute relative pseudo-bias of an estimator $\hat{Y}$, given by $\widetilde{\mathrm{ARB}}\ (\hat{Y}^{sel}(n_{ed})) = \left|\frac{\hat{Y}^{sel}(n_{ed}) - \hat{Y}^0}{\hat{Y}^0}\right|$.

The rationale of the proposed measure is the comparison with a random selection of units. The idea is to compare $\widetilde{\mathrm{ARB}}(\hat{Y}^{sel}(n_{ed}))$ for a selective editing technique *sel* with $\widetilde{\mathrm{ARB}}(n_{ed}) \equiv \widetilde{\mathrm{ARB}}(\mathbb{E}[\hat{Y}^{ran}(n_{ed})])$, where *ran* stands for an equal-probability selection and $\mathbb{E}$ is the expectation with respect to this random selection. It is immediate to show that $\widetilde{\mathrm{ARB}}_0\,(n_{ed}) = \left(1 - \frac{n_{ed}}{n}\right)\widetilde{\mathrm{ARB}}(\hat{Y}^{ran}(0))$. Let us denote by $\gamma_0(n_{ed})$ and $\gamma^{sel}(n_{ed})$ the straight and polygonal lines with vertices $\gamma_0(n_{ed}) \simeq \{(0, \widetilde{\mathrm{ARB}}_0\,(0)), (n_{ed}, \widetilde{\mathrm{ARB}}_0)(n_{ed}))\}$ and $\gamma^{sel}(n_{ed}) \simeq \{(0, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(0))), (1, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(1))), \ldots, (n_{ed}, \widetilde{\mathrm{ARB}}_0\,(\hat{Y}^{sel}(n_{ed})))\}$, respectively. Let us also denote by $A_\gamma(n_{ed})$ the signed area of the surface between the curve γ and the horizontal axis to the left of the vertical line at n_{ed} (see Figure 2). The area is agreed to be positive if the polygonal line lies below the straight line and is otherwise negative. We propose the following definition for the efficiency of the technique *sel*:

$$\epsilon^{sel}(n_{ed}) \equiv (A_{\gamma_0}(n_{ed}) - A_{\gamma^{sel}}(n_{ed}))/A_{\gamma_0}(n_{ed}) = 1 - \frac{A_{\gamma^{sel}}(n_{ed})}{A_{\gamma^0}(n_{ed})}.$$

Note that this measure depends on the number of units to select. This allows us to recognize those techniques which prioritize the most influential units first. A typical situation is depicted in Figure 2.

We have carried out a comparison of the preceding proposal of prioritization of units with that obtained from some score functions in the literature. In order to avoid possible interferences with missing data and units recently added to the sample, we have used a rectangular subset of the sample data of the Spanish ITI and INORI surveys (INE Spain 2010). For clarity's sake we shall concentrate on one particular score function, illustrate the corresponding results and make some comments regarding the similar behavior of all of them. We have used a slightly enhanced version of the RATIO function of Latouche and

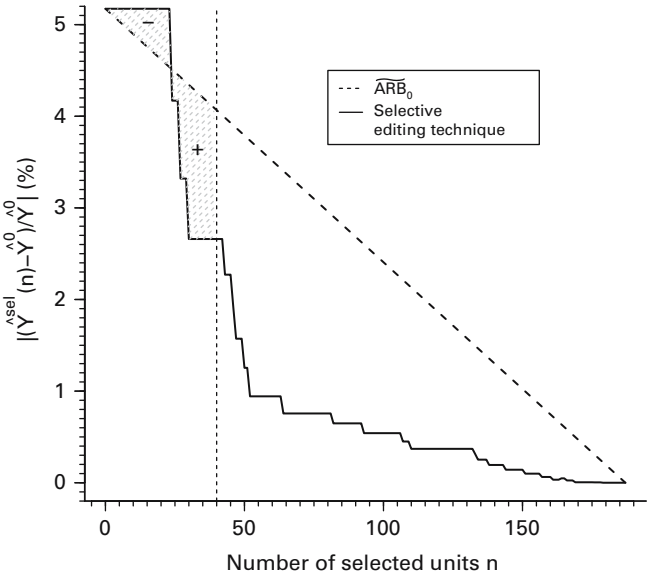

Fig. 2. Absolute relative pseudo-bias vs. number of selected units

Berthelot (1992). Let $r_k^{(t)} = \frac{y_k^{(t)}}{y_k^{(t-1)}}$ and define

$$\bar{r}_k^{(t)} = \begin{cases} \left| \dfrac{r_k^{(t)}}{\text{median}_k\left(r_k^{(t)}\right)} - 1 \right| & \text{if } r_k^{(t)} > \text{median}_k\left(r_k^{(t)}\right), \\ \left| 1 - \dfrac{r_k^{(t)}}{\text{median}_k\left(r_k^{(t)}\right)} \right| & \text{otherwise.} \end{cases}$$

Also define $g_k^{(t)} = w_{ks} \times \bar{r}_k^{(t)} \times \sqrt{\max\left(y_k^{(t)}, y_k^{(t-1)}\right)}$ and then the local score $s_k^{(t)} = \frac{|g_k^{(t)} - \text{median}_k\left(g_k^{(t)}\right)|}{\text{IQR}_k\left(g_k^{(t)}\right)}$, where IQR stands for the interquartile range. For $q = 1, \ldots, Q$ variables, these combine in the global score function defined as RATIO2$(k,t) = S_k^{(t)} = \sum_{q=1}^{Q} s_k^{(q,t)}$. The enhancement arises due to the fact that only data from the time period $t-1$ is used and not from $t-2$, as in the original proposal. Thus this function RATIO2 can only be used as a form of output editing after all data have been collected (as the combinatorial approach, which we are making the comparison with).

Regarding the prioritization of units computed under the combinatorial approach, firstly we must specify the auxiliary model m^* to find the predicted values $\hat{y}_k$. For each unit we have fitted three alternative time series models $\xi_1 : (1-B)_{z_t} = a_t$, $\xi_2 : (1-B^{12})_{z_t} = a_t$ and $\xi_3 : (1-B)(1-B^{12})_{z_t} = a_t$, where B stands for the backshift operator, $zt = \log\left(m + y_t^0\right)$ (m being a nuisance parameter estimated by maximum likelihood) and a_t denotes white noise. Each predicted value $\hat{y}_k$ is computed according to the corresponding best model ξ^* (in terms of the minimal estimated mean squared error). Since the sample is a fixed panel selected by cut-off, the sampling weights w_{ks} are all equal to 1.

Next, we have applied the generic univariate observation-prediction model illustrated in Section 3 to the logarithmic transforms of the turnover and the new orders received

independently. The common error probability $p_k = p$ and observation variance $\sigma_k^2 = \sigma^2$ have been estimated from the past three months using a double-data set. The prediction variance ν_k^2 has been computed according to the corresponding chosen best model ξ^* for each unit. As loss matrices, we have chosen both the squared and the absolute loss function with entries given by Equations (4) and (5), respectively.

Finally, to make the comparison with a random selection of units, we have computed the absolute relative pseudo-bias for 50 equal-probability random selections. We have calculated the mean and first and third quartiles of the corresponding distribution. This provides a confidence-like interval for each number of selected units (see Figure 3). The motivation is to provide an insight not only into the average random selection but also of its distribution.

We have carried out this comparison for 23 NACE Rev. 2 divisions and subdivisions (aggregations of groups according to subject-matter knowledge). Firstly, RATIO2 showed a better performance than the rest of score functions (RATIO, DIFF, FLAG ITI, FLAG INORI;

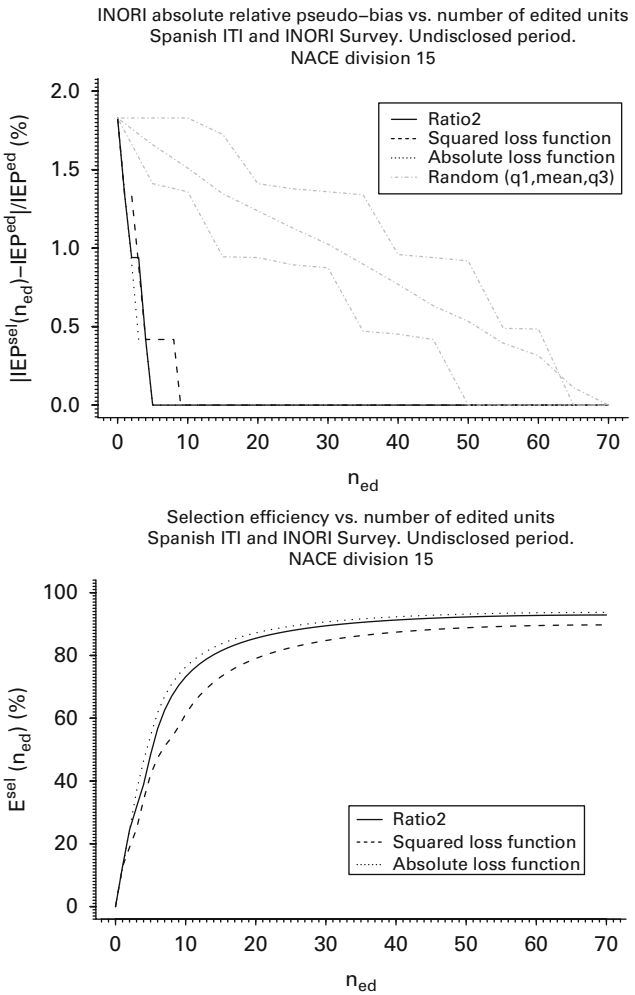

Fig. 3. Absolute relative pseudo-bias and editing efficiency vs. number of selected units

see Latouche and Berthelot 1992). In 15 cases the absolute loss yielded the most efficient prioritization, with nine of these cases having the RATIO2 score function as more efficient than the squared loss choice (Figure 3 illustrates this behavior). However in five cases it is the squared loss function that outperforms the other two choices, in which the absolute loss also did better than the score function. In the remaining three cases, RATIO2 slightly overcame the absolute loss, which in turn performed better than the squared loss.

Thus, in general, the absolute loss is more efficient than the squared loss in terms of the pseudo-bias, as expected. This also happens with the score function RATIO2, since it is also targeted at the bias. In general, the absolute loss is also more efficient than the score functions. However, in actual production conditions, both missing data and respondents newly added to the sample must be taken into account. In these cases, in the optimization approach the prediction values $\hat{y}_k$ must be imputed or fixed under some supplementary scheme, since the considered time series models fail to produce these values. As an elementary test, we assigned $\hat{y}_k = y_k$ in these cases in order for them not to be selected at first positions. The general result was a slight deterioration of the performance of the score functions for all values of n_{ed}, while in the optimization approach, the behavior was as good as before for the most influential units ($n_{ed} = 1, 2, \ldots$), but noticeably poorer for the last units ($n_{ed} \geq n/2$). We have not considered these issues in the preceding comparison, since they belong to sophistications of the observation-prediction model.

In our opinion it is important to note that the above results have been obtained with crude time series models and extremely simplifying assumptions, and they do not incorporate any subject-matter knowledge. Thus there is more room to elaborate further on them (using better parameters, building multivariate models, etc.). In this line of thought the most attractive point will arise if working models can be built for discrete or semicontinuous variables, paving the way for the use of selective editing techniques also in household surveys. The possibility of using well-established tools such as time series models or statistical models in general, reinforces the statistical defensibility of the data editing work.

6. Discussion and Concluding Remarks

Once we have detailed the methodological proposal, we now proceed to discuss several issues regarding this optimization approach from different perspectives. As two immediate objections, a cautious reader can point out the limitation to linear estimators and the polynomial running time of the algorithms. Firstly, the limitation to linear estimators, which contrasts with the common use in practice of some nonlinear estimators such as ratio estimators or regression estimators, can be easily overcome as follows. In practice most nonlinear estimators $\hat{Y}^{nl}_{U_d}$ are functions of linear estimators $\hat{Y}^{nl}_{U_d} = f\left(\hat{Y}^{(1)}_{U_d}, \ldots, \hat{Y}^{(M)}_{U_d}\right)$. Then instead of considering the corresponding restriction for the MSE of $\hat{Y}^{nl}_{U_d}$, we consider a restriction for each linear estimator $\hat{Y}^{(m)}_{U_d}$, $m = 1, \ldots, M$. The rationale amounts to expecting an accurate nonlinear estimator if each linear estimator is accurate. Moreover, a bounded growth in the number of restrictions is expected, since nonlinear estimators are usually built from different combinations of survey variables, whose number is fixed by the questionnaire. Secondly, the polynomial running time of the selection algorithms is not a practical concern, at least in Spanish sampling sizes standards, as we will now explain. On the one hand, the estimation problem in a finite population U is

essentially a multivariate problem seeking accurate and numerically consistent estimations in given partitions of the population U. These partitions are fixed according to the breakdown established by the statistical dissemination plan of each survey. Thus the selection or prioritization should be applied to each of these publication cells, since no lack of accuracy is rightfully allowed in any published figure. On the other hand, we have applied this approach to the Spanish ITI and INORI survey as a pilot experience at INE Spain (details will be published elsewhere). In these monthly short-term business statistics, the sampling size amounts to around 12,000 industrial establishments broken down into 37 publication cells with sizes ranging up to 1,500 units at most. The prioritization of units in all cells took a total of three hours on a desktop PC, which is a reasonable working time.

As a deeper concern, one can inquire why the roles of the two basic principles of our formulation are not interchanged, that is, why data quality is not optimized (minimizing the loss function) restricting the amount of resources used (number of questionnaires to recontact). We give two reasons to support our proposal. From a broad perspective, in a statistical office it appears desirable to minimize the cost of each survey in order to optimize resources to face and embrace as many other surveys in the statistical production as possible. In our view, this is a natural decision given the increasing demand for information from stakeholders. From a more methodological standpoint, the multivariate feature of the problem again arises. If we interchanged the roles of both principles, we would need to minimize the loss function of the different variable estimators corresponding to each publication cell restricted to the number of questionnaires to be recontacted. As a matter of fact this is a multiobjective optimization problem, which ineludibly needs some decisions to compute a solution (see e.g., Marler and Arora 2004). In this respect, our position in official statistics production is to minimize the number of decisions taken by the survey conductor, which is clearly expressed in the following citation by Hansen et al. (1983): "[. . .] it seems desirable, to the extent feasible, to avoid estimates or inferences that need to be defended as judgments of the analysts conducting the survey".

As a matter of fact, the question of the number of decisions is a first relevant point to establish a comparison with the score function approach. Nowadays the score function approach is undisputedly the favored technique for selecting influential units in the editing production phase. Thus it provides the framework to assess advantages and disadvantages of any other technique. Furthermore, in our opinion, a comparison will help us reveal fundamental aspects of the editing production phase irrespective of the particular techniques. Regarding the number of decisions, let us recall that the score function approach comprises four main decisions to determine a selection of units (Lawrence and McKenzie 2000), namely (i) an editing model to construct the anticipated values, (ii) each local score function, (iii) a global score function, and (iv) a cut-off value. On the one hand, in the optimization approach the first three decisions are jointly substituted and integrated into a single step: the construction of the observation-prediction model or an alternative statistical error-modelling technique, and the subsequent formulation of the optimization constraints. Furthermore, in our view, this integration renders this selection procedure more natural within the statistical language, in contrast to a score function, which can seem extraneous. In this sense, let us point out that the construction of an observation-prediction model is a multivariate exercise, so the integration of the choices of both local and global score functions comes naturally together

with the construction of the statistical model. On the other hand, the choice of the cut-off value is now substituted for the choice of the bounds in the optimization problem. In the score function approach, this value must be chosen normally using data from previous realizations of the survey and using a heuristic or empirical connection between this value and the chosen loss function of the survey estimators. In the optimization approach, the choice of the bounds makes use of a priori values of variances (or some other similar measure) as in the survey design stage and shows a neater connection with the loss function, thus fitting again more naturally into the whole survey statistics production process. Indeed, we have shown how the prioritization of units under the score function approach can be reproduced and slightly overcome with a very simple model. Furthermore, although admittedly still too far, this proposal points toward enlarging the traditional sampling strategy $(\mathcal{D}, T)$ comprising the sampling design $\mathcal{D}$ and the construction of the estimator T (see e.g., Hedayat and Sinha 1991) with a selection strategy R, so that we would have a triplet $(\mathcal{D}, R, T)$. This follows the spirit of the total survey design.

The selection/prioritization issue goes hand in hand with the double version of the optimization approach. This issue arises mainly from resource availability and controllability, mainly of timeliness and person-hours in the editing fieldwork. When having a selection of units in practice we face two situations: Either we run out of resources to accomplish the interactive editing of all selected units, or we end up ahead of time and then we miss the opportunity to gain more accuracy. Now, since editing near the source is a must for this production phase, it is advisable to have a real-time selection mechanism on each questionnaire, as pointed out in the introduction, independently of the rest of the sample. Conversely, on later stages it is preferable to prioritize units to edit (interactively) the most influential first. In this line of thought, the stochastic approach suits the selection whereas the combinatorial approach suits the prioritization. Furthermore, since both approaches derive from a common general framework focused on the exploitation of auxiliary information, we envisage a more complex, although unified, editing process. Let us parameterize the auxiliary information used in the editing work in terms of its longitudinal, cross-sectional and multivariate dimensions. By longitudinal we mean the value of variables for each unit in previous time periods. By cross-sectional we refer to the information stemming from the sample at the current period. Finally, by multivariate we mean the information arising from the multidimensional character of the survey (always several variables are investigated). If we focus on the longitudinal and cross-sectional dimensions of the auxiliary information, Figure 4 represents the transition from micro-selective to macro editing as the data collection is being completed. In our view, these two editing techniques appear as the head and tail of a time-continuous process driven by the evolution of the data collection. We envisage that intermediate techniques combining both the longitudinal and available cross-sectional information as a time-continuous process during the data collection will be of practical usefulness.

Regarding the optimization approach, we want to point out that both versions fit naturally as the head and tail of this time-continuous editing process, so that the stochastic version corresponds to exploiting longitudinal information as in traditional selective editing techniques, whereas the combinatorial version arises as a macro editing technique focusing upon the cross-sectional information. In contrast, the score function approach and traditional macro editing techniques can hardly be seen under the same methodological

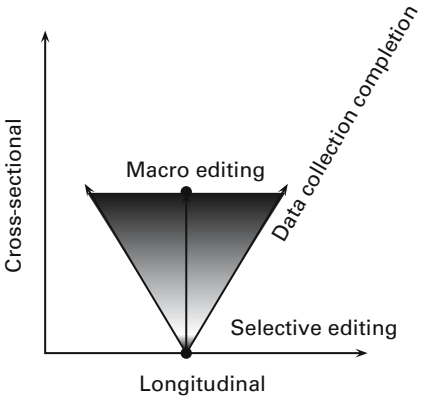

Fig. 4. Schematic representation of the transition from micro-selective to macro editing as data collection is completed. As data collection is completed, more cross-sectional information is available

principles. It remains open for future work to find a more general formulation for this proposed time-continuous process embedding both optimization versions.

A complementary comparison can be made with the automatic data editing techniques based on the Fellegi-Holt methodology, in particular with the different approaches to the error localization problem, which also make an extensive use of optimization techniques (see De Waal et al. 2011). The common points reduce to the fact that mathematical optimization appears as a natural translation of the proposed data editing principles. Conversely, the Fellegi-Holt methodology focuses upon each questionnaire, seeking to minimize the number of items to change satisfying all edits. In this approach we focus upon the whole sample, seeking to minimize the number of units to be recontacted satisfying restrictions upon the loss functions using a statistical model instead of edits.

To conclude, as immediate future prospects, we have recently begun to analyze the inclusion of these techniques in the current E&I strategies in most business surveys in INE Spain. A pilot experience with the ITI and INORI survey fosters our hope to reduce current recontact rates and consequently both editing costs and the response burden at our office. R packages and SAS macros implementing this optimization approach are under intense development and being tested in these pilot experiences. Apart from this, more methodological research is needed to find generic multivariate models fitting the observation-prediction model and to generalize them to both qualitative and semicontinuous variables. In this context, multivariate models already present in the literature for data editing (Di Zio and Guarnera, in this issue) appear as a fruitful alternative. In addition, we already have a first adaptation of the preceding greedy algorithms to be applied to surveys with self-weighting samples and qualitative variables. We are collaborating with experts from the Spanish National Health Survey to produce an observation-prediction model adapted to these variables.

A. Mathematical Appendix

We include some mathematical proofs. Firstly we prove how the constraints imply a control on the loss of accuracy. In particular, if $L = L^{(r)}$ denotes the absolute ($r = 1$) or

squared loss ($r = 2$) function, we prove that $\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R})\hat{Y}^0)|\mathbf{Z}] \leq \eta$ (where $\mathbf{Z} = \mathbf{Z}^{st}$ or $\mathbf{Z}^{cross}$) implies $\mathbb{E}_{pm}[L(\hat{Y}^*(\mathbf{R}), Y)] \leq \left(\eta^{1/r} + \mathbb{E}_{pm}^{1/r}[L(\hat{Y}^0, Y)]\right)^r$. It is straightforward to prove that $d(A,B) = \mathbb{E}_{pm}^{1/r}[L^{(r)}(A,B))]$ is a metric. Then, by the triangle inequality, we have

$$d(\hat{Y}^*(\mathrm{R}), Y) \leq d(\hat{Y}^*(\mathrm{R}), \hat{Y}^0) + d(\hat{Y}^0, Y).$$

Now, using properties of the conditional expectation, we can write

$$d^r(\hat{Y}^*(\mathrm{R}), \hat{Y}^0) = \mathbb{E}_{pm}[\mathbb{E}_m[L(\hat{Y}^*(\mathbf{R}), \hat{Y}^0|\mathbf{Z}]] \leq \eta,$$

where $\mathbf{Z} = \mathbf{Z}^{st}$ or $\mathbf{Z}^{cross}$. The result follows immediately.

Secondly we show the connection between the loss matrices and the loss function. In the absolute loss case, we have $\mathbb{E}_m[|\hat{Y}^*(\mathbf{R}) - \hat{Y}^0|\,|\mathbf{Z}] = \mathbb{E}_m\left[\left|\sum_{k\in s} R_k w_{ks}\epsilon_k\right||z\right] \leq \left[\sum_{k\in s} R_k^2 |w_{ks}\epsilon_k|\,|\mathbf{Z}\right]$, since $R_k^2 = R_k$. Thus we can write $\mathbb{E}_m[|\hat{Y}^*(\mathbf{R}) - \hat{Y}^0|\,|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$, where Δ is diagonal with entries $\Delta_{kk} = |w_{ks}\epsilon_k|$. In the squared loss case, in turn we have $\mathbb{E}_m[(\hat{Y}^*(\mathbf{R}) - \hat{Y}^0)^2|\mathbf{Z}] = \mathbb{E}_m\left[\sum_{k\in s}\sum_{k\in s} R_k R_l w_{ks}\epsilon_k w_{ls}\epsilon_l|\mathbf{Z}\right]$. Thus, we can also write $\mathbb{E}_m[(\hat{Y}^*(\mathbf{R}) - \hat{Y}^0)^2|\mathbf{Z}] = \mathbb{E}_m[\mathbf{R}^T\Delta\mathbf{R}|\mathbf{Z}]$, where $\Delta_{kl} = w_{ks}\epsilon_k \cdot w_{ls}\epsilon_l$.

The conditional moments (4) and (5) are found along similar lines. Under the hypotheses assumed in Section 3 regarding the observation-prediction model, it follows that $y_k - \hat{y}_k = \epsilon_k^{obs} + \epsilon_k^{pred}$ and $\mathbb{E}_m\left[\left(y_k - y_k^0\right)^r|s_k, y_k, \hat{y}_k\right] = \mathbb{E}_m\left[\delta_k^{(obs)r}|s_k, y_k, \hat{y}_k\right]$. $\mathbb{E}_m[e_k|s_k, y_k, \hat{y}_k]$, with $r = 1, 2$. Conditioning on $s_k, y_k, \hat{y}_k$ amounts to conditioning on $s_k, \epsilon_k^{obs}, \hat{y}_k$, thus we can rewrite these conditional expectations as $\mathbb{E}_m[\cdot|s_k, y_k - \hat{y}_k, \hat{y}_k]$. Now the second term is computed using Bayes' theorem, so that $\mathbb{E}_m[e_k|s_k, y_k - \hat{y}_k, \hat{y}_k] = \zeta k\left(\frac{y_k - \hat{y}_k}{\nu_k}\right)$. For the first term, we notice that the random vector $\left(\delta_k^{obs}, \delta_k^{obs} + \epsilon_k^{pred}\right)^T$ is normally distributed with expectation $\mu = 0$ and variance $\Sigma = \begin{pmatrix} \sigma_k^2 & \sigma_k^2 + \rho_k\sigma_k\nu_k \\ \sigma_k^2 + \rho_k\sigma_k\nu_k & \sigma_k^2 \end{pmatrix}$. The conditional moments follow then from standard properties of the multivariate normal distribution.

7. References

Arbués, I., González, M., and Revilla, P. (2012a). A Class of Stochastic Optimization Problems with Application to Selective Data Editing. Optimization, 61, 265–286. DOI: http://www.dx.doi.org/10.1080/02331934.2010.511670

Arbués, I., Revilla, P., and Salgado, D. (2012b). Optimization as a Theoretical Framework to Selective Editing. UNECE Work Session on Statistical Data Editing, 24–26 September. WP2, 1–10.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New York: Wiley.

Granquist, L. (1997). The New View on Editing. International Statistical Review, 65, 381–387.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-dependent and Probability Sampling Inferences in Sample Surveys. Journal of the American

Statistical Association, 78, 776–793. DOI: http://www.dx.doi.org/10.1080/01621459.1983.10477018

Hedayat, A.S. and Sinha, B.K. (1991). Design and Inference in Finite Population Sampling. New York: Wiley.

INE Spain (2010). Industrial Turnover Indices. Industrial New Orders Received Indices. Base 2005. CNAE-09. Methodological Manual. Available at http://www.ine.es/en/metodologia/t05/t0530053_en.pdf. (accessed January 10, 2013).

Latouche, M. and Berthelot, J.M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. Journal of Official Statistics, 8, 389–400.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243–253.

Lindgren, K. (2011). Selective Editing in the International Trade in Services. Working Paper 19 of 2011 UNECE Metting on Statistical Data Editing. May 9–11, Ljubljana. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2011/wp.19.e.pdf (accessed October 2013).

Marler, R.T. and Arora, J.S. (2004). Survey of Multi-Objective Optimization Methods for Engineering. Structural and Multidisciplinary Optimization, 26, 369–395. DOI: http://www.dx.doi.org/10.1007/s00158-003-0368-6

Salgado, D., Arbués, I., and Esteban, M.E. (2012). Two Greedy Algorithms for a Binary Quadratically Constrained Linear Program in Survey Data Editing. INE Spain Working Paper 02/12. Available at http://www.ine.es. (accessed January 10, 2013).

Wets, R.-B. (2002). Stochastic Programming Models: Wait-and-see Versus Here-and-now. Institute for Mathematics and Its Applications, 128.

Received February 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 511–537, http://dx.doi.org/10.2478/jos-2013-0038

Automated and Manual Data Editing: A View on Process Design and Methodology

Jeroen Pannekoek[1], *Sander Scholtus*[1], *and Mark Van der Loo*[1]

Data editing is arguably one of the most resource-intensive processes at NSIs. Forced by ever-increasing budget pressure, NSIs keep searching for more efficient forms of data editing. Efficiency gains can be obtained by selective editing, that is, limiting the manual editing to influential errors, and by automating the editing process as much as possible. In our view, an optimal mix of these two strategies should be aimed for. In this article we present a decomposition of the overall editing process into a number of different tasks and give an up-to-date overview of all the possibilities of automatic editing in terms of these tasks. During the design of an editing process, this decomposition may be helpful in deciding which tasks can be done automatically and for which tasks (additional) manual editing is required. Such decisions can be made a priori, based on the specific nature of the task, or by empirical evaluation, which is illustrated by examples. The decomposition in tasks, or statistical functions, also naturally leads to reuseable components, resulting in efficiency gains in process design.

Key words: Automatic editing; selective editing; edit rules; process design; process evaluation.

1. Introduction

The quality of raw data available to National Statistical Institutes (NSIs) is rarely sufficient to allow the immediate production of reliable statistics. As a consequence, NSIs often spend considerable effort to improve the quality of microdata before further processing can take place.

Statistical data editing encompasses all activities related to the detection and correction of inconsistencies in microdata, including the imputation of missing values. Data editing, or at least the correction part of data editing, has traditionally been performed manually by data-editing staff with subject-specific expert knowledge. The manual follow-up of a large number of detected inconsistencies is, however, very time consuming and therefore expensive and decreases the timeliness of publications. Therefore, several approaches have been developed to limit this resource-consuming manual editing.

[1] Statistics Netherlands, P.O. Box 24500, 2490 HA The Hague, The Netherlands. Emails: jpnk@cbs.nl, sshs@cbs.nl, and mplo@cbs.nl

Acknowledgments: The views expressed in this article are those of the authors and do not necessarily reflect the policies of Statistics Netherlands. The authors are grateful to T. de Waal, M. Di Zio, U. Guarnera, I. Arbués, P. Revilla and D. Salgado for comments and suggestions and to L.-C. Zhang for fruitful discussions on the subject of this article.

One approach is selective editing (Latouche and Berthelot 1992). This is an editing strategy in which manual editing is limited or prioritised to those errors where this editing has a substantial effect on estimates of the principal parameters of interest. Provided that there is an effective way of determining the influential errors, this strategy can be successful. It has been well established (see the review by Granquist and Kovar 1997) that for many economic surveys only a minority of the records contains influential errors that need to be edited; the remaining errors can be left in without substantial effect on the principal outputs.

An alternative route to reducing manual editing is to perform the editing automatically. Automatic editing is not a single method but consists of a collection of formalised actions that each perform a specific task in the overall editing process. Some well-known tasks performed in automatic editing are the evaluation of edit rules to detect inconsistencies in the data, the localisation of fields that cause these inconsistencies, the detection and correction of systematic errors such as the well-known thousand error, and the imputation of missing or incorrect values. Once implemented, automatic editing is fast, uses hardly any manual intervention and is reproducible. For reasons of efficiency, it should therefore be preferred to manual editing even if the latter is confined to selected records. However, not all data-editing functions can be performed automatically while achieving a result of sufficient quality. Selective manual editing is then a necessary addition.

The relationship between manual and automatic editing as it emerges from the classical literature on selective editing is that all important amendments should be done manually and that the role of automatic editing is confined to the less influential errors: its purpose is mainly to ensure the internal consistency of the records so as to avoid inconsistencies at all levels of aggregation. In this view, the quality of automatic editing has no bearing on the decision to edit a record manually or automatically. Efficiency gains are realised by the selection process only. The point of view taken in this article is that for reasons of efficiency, manual editing should be confined to the data that are influential *and* cannot be treated automatically with sufficient quality. In this view, the quality of automatic editing is important in making the decision to edit manually or not and improvements in automatic editing will lead to efficiency gains.

This article gives an overview of the current state of the art in efficient editing of establishment data. Using numerical results from two example statistics, it is shown how with the current methods, selective editing can be minimised while data quality is retained. We identify methodological research directions that in our view have potential for yielding further efficiency gains.

Besides making the data-editing process more efficient, there is a need to increase the cost effectiveness of designing and implementing data-editing systems. In this article we propose a hierarchical decomposition of the data-editing process into six different task types, called *statistical functions*. This view of the overall process builds on the previous work of Camstra and Renssen (2011) and Pannekoek and Zhang (2012) by adding a taxonomy of editing functions and defining the minimal input and output requirements of each of these functions. Identifying the in- and output parameters of these abstract functions allows us to move towards a modern approach to process design, based on reusable components that connect in a plug-and-play manner.

The remainder of this article is structured as follows. Section 2 discusses some basic aspects of error detection in manual and automatic editing. First we consider the different kinds of errors that can arise and differentiate between errors for which automatic treatment is a possibility and those for which manual treatment is required. Then we discuss the edit rules that are extensively used in data editing, in particular with respect to business surveys. In Section 3 an overview is given of both well-known and more recently developed automatic error detection and correction methods. Section 4 is concerned with a decomposition of the overall data-editing process into data-editing functions based on the action and purpose of these functions. In Section 5 the application of a sequence of different editing functions is illustrated using two real data examples. This section also gives references to the freely available R packages that are used for these illustrations. Finally, in Section 6 we summarise some conclusions.

2. Error Detection in Manual and Automated Editing

2.1. Sources of Errors in Survey Data

In analyses of survey errors it is customary to decompose the total error into more or less independent components that may be treated separately. Well-known decompositions include those by Groves (1989) and Bethlehem (2009). Here, we use Bethlehem's taxonomy of survey errors since it allows us to identify sources of error with common data-editing strategies.

Bethlehem (2009) uses the scheme shown in Figure 1 to distinguish between sources of error in a statistical statement based on surveys. The total error is decomposed into sampling and nonsampling error. The sampling error is further decomposed into selection and estimation error. Selection error consists of differences between the theoretical and realised inclusion probabilities of sampling units, while estimation error consists of the usual variance and bias introduced by the estimation method. Nonsampling errors can be split into observational and nonobservational errors. Observational errors are composed of overrepresentation of elements in the population register (overcoverage), measurement errors (item nonresponse, completion errors, etc.) and processing errors at the NSI (e.g., data entry errors). Nonobservational errors are caused by omission of elements from the population register (undercoverage) and unit nonresponse.

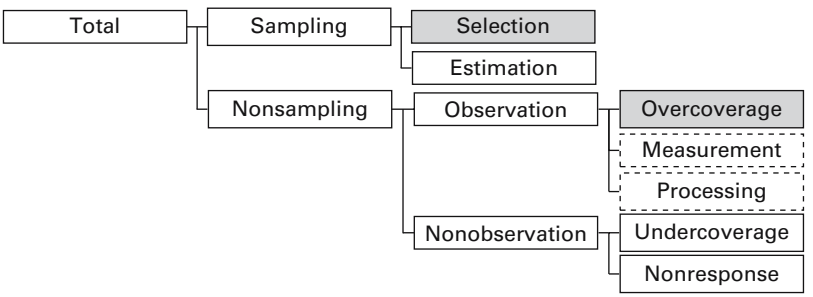

Fig. 1. Bethlehem's (2009) taxonomy of survey errors. Errors in grey boxes are commonly solved by manual data editing while automated techniques are usually more suited for error causes indicated in dotted boxes

Traditionally, automated data-editing methods have more or less focused on errors occuring at the measurement or processing stage. That is, many automated data-editing methods focus on the observed variables rather than the identifying or classifying variables already available in the population register. For example, in a stratified hot-deck imputation scheme, the values of stratifying variables are assumed correct to begin with. In contrast, data-editing staff often do not make such assumptions and may frequently reclassify units.

Since many automated data-editing methods are based on mathematical modelling, they usually assume that some kind of structured auxiliary information is available. In many cases historic records, auxiliary register variables or totals from related statistics can be used to estimate values for erroneous or missing fields in a survey data set. By contrast, data-editing staff may use unstructured auxiliary information to edit records. Such information may, for example, include written financial reports or information from websites, as well as recontacts. These two differences between manual and automated data editing enable data-editing staff to correct for errors not caused at the moment of measurement.

In 2010, thirteen of Statistics Netherlands' data-editing employees working on the short-term business survey were informally interviewed on commonly found errors and data-editing practices. Besides a number of commonly found measurement errors (reporting of net instead of gross turnover, reporting of value of goods instead of invoices, etc.) many causes of error that were mentioned are nonobservational or sampling errors in Bethlehem's taxonomy. Examples include misclassifications such as retailers being registered as wholesalers, population effects such as bankruptcies, splits and mergers, and differences between legal units (chamber of commerce), tax units (of the tax office) and economic units (of Statistics Netherlands). Such errors are detected and/or solved by looking at auxiliary information such as figures and articles from sector organisations and (financial) newspapers, a website dedicated to registering bankruptcies, publicly available information on wages and retirement funds in a sector and so on. Subject-matter experts also use (often unstructured) domain knowledge on branch-specific transient or seasonal effects to detect errors. Examples of such effects include weather conditions (energy and construction), holidays (food industry, printers, etc.) and special events (tourist sector).

For the various measurement errors mentioned by the interviewees, conventional automatic data-editing methods can in principle be applied. For nonobservational errors like population errors and misclassifications, the error detection and correction process is based on fuzzier types of information and therefore harder to automate. At the moment, we are not aware of methods that can exploit such information for data-editing purposes automatically.

2.2. *Edit Rules for Automatic Verification*

Prior knowledge on the values of single variables and combinations of variables can be formulated as a set of edit rules (or edits for short), which specify or constrain the admissible values. For single variables such edits are range checks; for most variables in business surveys these amount to a simple non-negativity or positivity requirement such as:

e_1: *Number of employees* ≥ 0
e_2: *Turnover* > 0

Edits involving multiple variables describe the admissible combinations of values of these variables in addition to their separate range restrictions. For numerical business data, many of these edits take the form of linear equalities (balance edits) and inequalities. Some simplified examples of such edit rules are:

e_3: *Result* = *Total revenues* − *Total costs*
e_4: *Total costs* = *Purchasing costs* + *Personnel costs* + *Other costs*
e_5: *Turnover* = *Turnover main activity* + *Turnover other activities*
e_6: *Employee costs* $<$ 100 $\times$ *Number of employees*

The inequality and equality edits e_1−e_5 are examples of *fatal* or *hard* edits: they must hold true for a correct record. This class of edits is opposed to the so called *soft* or *query* edits, the violation of which points to highly unlikely or anomalous (combinations of) values that are suspected to be in error although this is not a logical necessity. The edit e_6 could be interpreted as a soft edit.

More generally, an inequality edit k can be expressed as $\sum_{j=1}^{J} a_{kj}x_j \leq b_k$, with the x_j denoting the variables, the a_{kj} coefficients, and b_k a constant. In e_1 and e_2, $b_k = 0$ and the a_{kj} are zero for all variables except one, for which a_{kj} is -1. Linear equalities such as e_3, e_4 and e_5 can similarly be expressed as $\sum_{j=1}^{J} a_{kj}x_j = b_k$.

Note that these edits are connected by certain common variables, which is true for many of the edits used in business statistics and has consequences for error localisation and adjustment for consistency. In such situations it is convenient to re-express the edits as a system of K linear equations and inequalities, in matrix notation:

$$\boldsymbol{Ex} \odot \boldsymbol{b}, \tag{1}$$

with $\boldsymbol{E}$ the $K \times J$ *edit matrix* with elements a_{kj}, $\boldsymbol{x}$ a J-vector containing the variables and $\boldsymbol{b}$ a K-vector with elements b_k. The symbol $\odot$ should here be interpreted as a vector of operators (with values $<$, $=$ or $\leq$) appropriate for the corresponding (in)equalities.

Each of the edit rules can be verified for each record. If we have N records and K edits, all the failure statuses can be summarised in a binary $N \times K$ *failed-edits matrix* $\boldsymbol{F}$, corresponding to all the record-by-edit combinations. The failure statuses can be the input to an error localisation function that selects variables from those involved in failed edits with values that are to be considered erroneous and need to be changed in order to resolve the edit failures (see Subsection 3.3).

The number of edit rules varies greatly between statistical domains. The structural business statistics (SBS) are an example with a large number of edit rules. An SBS questionnaire can be divided into sections. It contains, for instance, sections on employees, revenues, costs and results. In each of these sections a total is broken down into a number of components. Components of the total number of employees can be part-time and full-time employees and components of total revenues may be subdivided into turnover and other operating revenues. The total costs can have as components: purchasing costs, depreciations, personnel costs and other costs. The personnel costs can be seen as a subtotal, since they can again be broken down in subcomponents: wages, training and

other personnel costs. Each of these breakdowns of a (sub)total corresponds to a (nested) balance edit. SBS questionnaires also contain a profit and loss section where the revenues are balanced against the costs to obtain the results (profit or loss), which leads to the edit e_3. This last edit connects the edits from the costs section with the edits from the revenues section. Soft edits for the SBS form are often specified as bounds on ratios. For instance, ratios between a component and the associated total, between the number of employees and the personnel costs, between purchasing costs and turnover, etc.

3. Methods for Automatic Detection and Amendment of Missing or Erroneous Values

The overall editing process can be seen as a sequence of statistical functions applied to a data set. Such functions, for example selecting records for manual editing, may be implemented as an automated or a manual subprocess. In this section we summarise a number of data-editing methods that can be performed automatically.

Since these methods often detect or correct different types of errors, they will usually be applied one after another so as to catch as many errors as possible. A detailed exposition of the statistical methodology for each of these functions is beyond our scope, but below we summarise the types of methods that could be used and/or give some simple examples. More detailed descriptions can be found in De Waal et al. (2011) and the references cited there.

3.1. Correction of Generic Systematic Errors

From a pragmatic point of view, a systematic error is an error for which a plausible cause can be detected and knowledge of the underlying error mechanism enables a satisfactory treatment in an unambiguous deterministic way. De Waal et al. (2012) distinguish between generic systematic errors and subject-related systematic errors. A generic systematic error is an error that occurs with essentially the same cause for a variety of variables in a variety of surveys or registers. Subject-related systematic errors on the other hand occur for specific variables, often in specific surveys or registers.

3.1.1. Unit of Measurement Error

A well-known generic systematic error is the so-called unit of measurement error which is the error of, for example, reporting financial amounts in Euros instead of the requested thousands of Euros. Unit of measurement errors are often detected by a simple ratio criterion that compares the raw value x_{raw} with a reference value x_{ref}. Such a rule can be expressed as

$$\frac{x_{raw}}{x_{ref}} > t, \tag{2}$$

with t some threshold value. The reference value can be an approximation to the variable x that is unaffected by a unit of measurement error, such as an edited value for the same unit from a previous round of the same survey or a current or previous stratum median of x. The detection of unit of measurement errors may be improved by dividing the financial variables by the number of employees (e.g., costs or revenues per employee) to eliminate

the variation in these variables due to the size of the unit. If a thousand error is detected, the affected values are divided by one thousand. See, e.g., Di Zio et al. (2005) and Al Hamad et al. (2008) for further discussion and more advanced methods for detecting unit of measurement errors.

Thousand errors are often made in a number of financial variables simultaneously, yielding what is known as a uniform thousand error in these variables. Thousand errors will not violate balance edits if they are uniform in all variables involved; therefore, they cannot be detected by such edits. Incidental thousand errors may be detected by balance edits when the error is made in one or more of the components or their total but not in all these variables.

3.1.2. Simple Typing Errors, Sign Errors and Rounding Errors

Some inconsistencies are caused by simple typing errors. Recently, methods have been developed to reliably detect and correct these types of errors (Scholtus 2009; Van der Loo et al. 2011). The algorithm correcting for typing errors uses the edit rules to generate candidate solutions and accepts them if the difference with the original value is not larger than a prespecified value. The difference is measured with the restricted Damerau-Levenshtein distance (Damerau 1964; Levenshtein 1966). This distance measure counts the (possibly weighted) number of deletions, insertions, alterations and transpositions necessary to turn one character string into another (the restriction entails that substrings, once edited, cannot be edited again).

The typo-correction can also correct simple sign errors. More complex sign errors, such as those caused by swapping *Cost* and *Turnover* in a questionnaire where the rule $Profit = Turnover - Cost$ must hold, can be solved by a binary tree algorithm that tests whether (combinations of) swapping options decrease the number of violated edits (Scholtus 2011).

Rounding errors cause edit violations by amounts of a few units of measurement at most. It is therefore of less importance which variables are adapted to resolve these inconsistencies. The scapegoat algorithm of Scholtus (2011) uses a randomisation procedure to adapt one or more variables by a small amount so that the number of equality violations is decreased.

3.2. Domain-Specific Correction Rules

In contrast to generic systematic errors, subject-related or domain-specific systematic errors occur for specific variables, often in specific surveys or registers. Such errors are often caused by a common misunderstanding of certain definitions among respondents. Restaurants, for instance, often incorrectly classify their main revenues as revenues from trade (because they sell food) rather than revenues from services as it should be. As another example, reporting net rather than gross turnover may occur frequently in some domains.

Direct if-then rules can be used to correct such errors. These rules are of the form

if *condition* then *action*,

where *condition* is a logical expression that is true if an error is detected and *action* is an amendment function that assigns new values to one or more variables.

Apart from being used for correction of subject-specific systematic errors, such rules are also used for selection and imputation. For selection of records for manual editing, the action consists of assigning TRUE to an indicator variable for manual treatment. For instance, if for large units crucial variables such as *Employment* or *Turnover* are missing or inconsistent, the unit may be selected for manual treatment. For the selection of fields to be changed, the action consists of changing some fields to NA (which stands for Not Available or missing). For instance, if the costs per employee are outside the admissible range, the number of employees (in Full Time Equivalents, FTE) may be selected as erroneous rather than the employee costs because it is known that the financial variables are reported more accurately. For imputation the condition specifies which missing value can be imputed by the rule and under what conditions. For instance

$$\text{if } \textit{Wages for temp. employees} = \text{NA and } \textit{No. of temp. employees} = 0$$

$$\text{then } \textit{Wages for temporary employees} \equiv 0,$$

We use the symbol $\equiv$ when we need to distinguish assignment from mathematical or logical equivalence ($=$). Even the evaluation of an edit rule can be seen as a rule in this if-then form. The condition is in that case the edit rule itself and the action is the assignment of a TRUE-FALSE status to a column of the matrix $\boldsymbol{F}$.

These rules are called *direct* correction/selection/imputation rules because the implementation of the condition and the action follows trivially from the rule itself. In contrast, the generic systematic errors discussed above such as typos and rounding errors are also based on rules, because they use the edit rules, but in those cases the implementation cannot be formulated in a single simple if-then rule but requires a more sophisticated algorithm. The same is true for Fellegi-Holt-based error localisation and model-based imputation with estimated parameters, to be discussed below.

3.3. Error Localisation

Error localisation is the process of pointing out the field(s) containing erroneous values in a record. Here, we assume that all fields should be filled, so an empty field (NA) is also assumed erroneous. If there are N records with J variables, the result of an error localisation process can be represented as a boolean $N \times J$ matrix $\boldsymbol{L}$, of which the elements L_{ij} are TRUE where a field is deemed erroneous (or when it is empty) and FALSE otherwise.

Automated error localisation can be implemented using direct rules, as mentioned in Subsection 3.2. In such a case a rule of the form

$$\text{if } \textit{condition} \text{ then } L_{ij} \equiv \text{TRUE} \tag{3}$$

can be applied. It should be noted that this method takes no account of edit restrictions, and does not guarantee that a record can be made to satify all the edits by altering only the content of fields with $L_{ij} = \text{TRUE}$; Boskovitz (2008) calls this the *error correction guarantee*.

Error localisation becomes more involved when one demands that 1) it must be possible to impute fields consistently with the edit rules and 2) the (weighted) number of fields to alter or impute must be minimised. These demands are referred to as the principle of Fellegi and Holt (1976). Identifying 1 and 0 with the boolean values TRUE and FALSE respectively, the localisation problem for each row $\boldsymbol{l}$ of $\boldsymbol{L}$ can be denoted mathematically as

$$\boldsymbol{l} \equiv \underset{\boldsymbol{u}\in\{0,1\}^J}{\arg\min}\, \boldsymbol{w}^T\boldsymbol{u} \tag{4}$$

under the condition that the set of (in)equality restrictions in Equation (1) has a solution for the x_j with $l_j = 1$, given the original values of the x_j with $l_j = 0$. The vector $\boldsymbol{l}$ points out which variables are deemed wrong (1) and which are considered correct (0). In addition, $\boldsymbol{w}$ is a non-negative weight vector assigning weights to each of the J variables. These weights are referred to as reliability weights, because they can be used to express the degree of trust one has in each original value x_j. Note that increasing w_j makes it less likely that x_j will be chosen as a candidate for amendment, as feasible solutions with lower weights are more likely to be available.

A special case occurs when only univariate (range) edits are considered – that is, when every edit contains just one variable. Denote by $\boldsymbol{C}$ the $K \times J$ boolean matrix that indicates which variables (columns) occur in which edits (rows), and denote by $\boldsymbol{X}$ the $N \times J$ numerical data matrix. In this special case, the matrix $\boldsymbol{C}$ contains at most $2J$ nonzero elements, since each variable can be bounded from above or below or both. The matrix $\boldsymbol{L}$ can then be computed as

$$\boldsymbol{L} \equiv (\boldsymbol{FC} > 0) \vee (\boldsymbol{X} = \mathsf{NA}). \tag{5}$$

Here, $\boldsymbol{F}$ is the $N \times K$ failed-edits matrix defined in Subsection 2.2, and the logical and comparison operators ($<$ and $=$) on the right-hand side should be evaluated elementwise. The symbol $\vee$ indicates the elementwise or operation.

Several algorithms have been developed for error localisation under interconnected multivariate linear constraints. See De Waal et al. (2011) and the references therein for a concise overview of available algorithms. Regardless of the algorithm used, the special case of Equation (5) can be applied to the subset of univariate edits prior to one of the more complex algorithms to reduce computational complexity. The branch-and-bound algorithm of De Waal and Quere (2003) and approaches based on a reformulation of the error localisation problem as a mixed-integer problem (MIP) have recently been implemented as a package for the R statistical environment by De Jonge and Van der Loo (2011).

3.4. Imputation of Missing or Discarded Values

Imputation is the estimation or derivation of values that are missing due to nonresponse or discarded for being erroneous (as indicated by $\boldsymbol{L}$ in the previous section). Below we discuss deductive and model-based imputation methods.

3.4.1. Deductive Imputation of Missing or Discarded Values

In some cases the values for the empty fields can be derived uniquely from edit rules by mathematical or logical derivation. For example, when one value in a balance edit is missing, the only possible imputed value that will satisfy the balance edit can be computed from the observed values. For the interrelated systems of linear edits that are typical for the SBS, it is generally not obvious if some of the missing values are determined uniquely by the edit rules. By filling in the observed values from a record in the edit rules, a system of (in)equalities is obtained with the missing values as unknowns. Specifically, if $\boldsymbol{x}$ is partitioned as $\boldsymbol{x} = (\boldsymbol{x}_{obs}, \boldsymbol{x}_{mis})$ where $\boldsymbol{x}_{obs}$ denotes the subvector of $\boldsymbol{x}$ containing the observed values and $\boldsymbol{x}_{mis}$ the subvector with missing values and $\boldsymbol{E}$ is partioned conformably as $\boldsymbol{E} = (\boldsymbol{E}_{obs}, \boldsymbol{E}_{mis})$, then we have from $\boldsymbol{Ex} \odot \boldsymbol{b}$,

$$\boldsymbol{E}_{mis}\boldsymbol{x}_{mis} \odot \boldsymbol{b} - \boldsymbol{E}_{obs}\boldsymbol{x}_{obs}, \tag{6}$$

where the right-hand side is calculated from the observed values and $\boldsymbol{x}_{mis}$ contains the unknown missing values. The problem now is to determine which, if any, of these unknowns can be solved from this system and consequently deductively imputed. There exist simple algorithms that can find the values of all uniquely determined values for the unknowns in this system (De Waal et al. 2011).

3.4.2. Model-Based Imputation

Deductive imputation will in general only succeed for part of the missing values. For the remaining missing items, models are used to predict the values and these predictions are used as imputations. Here the term "model" is used in a broad sense, covering not only parametric statistical models but also nonparametric approaches such as nearest-neighbour imputation.

For business surveys with almost exclusively numerical variables, the predominant methods are based on linear regression models including, as special cases, (stratified) ratio and mean imputation (cf. De Waal et al. 2011, Ch. 7). Important for the efficiency of the application of regression imputation is that models for each of the variables that need imputation are specified in advance or selected automatically without the need for time-consuming model selection procedures by analysts at the time of data editing. When available, a historical value is often a good predictor for the current value.

An alternative, if all variables are continuous, is to use a multivariate regression approach where all variables that are observed in a record are used as predictors for each of the missing values. Thus, for each record, the variables are partitioned into two sets; the variables observed in record i and the variables missing in that record. The subvectors of $\boldsymbol{x}$ corresponding to these two sets will be denoted by $\boldsymbol{x}_{obs(i)}$ and $\boldsymbol{x}_{mis(i)}$ and the value of $\boldsymbol{x}_{obs(i)}$ in record i by $\boldsymbol{x}_{i.obs}$. If it is assumed that $\boldsymbol{x}$ is multivariate normally distributed, the conditional mean of the missing variables, given the values of the observed variables in record i, $\boldsymbol{\mu}_{i.mis}$ say, can be expressed as

$$\boldsymbol{\mu}_{i.mis} = \boldsymbol{\mu}_{mis(i)} + \boldsymbol{B}_{mis(i),obs(i)}(\boldsymbol{x}_{i.obs} - \boldsymbol{\mu}_{obs(i)}), \tag{7}$$

with $\boldsymbol{\mu}_{mis(i)}$ and $\boldsymbol{\mu}_{obs(i)}$ the unconditional means of $\boldsymbol{x}_{mis(i)}$ and $\boldsymbol{x}_{obs(i)}$ and $\boldsymbol{B}_{mis(i),obs(i)}$ an $n_{mis(i)} \times n_{obs(i)}$ matrix with rows containing the coefficients for the $n_{mis(i)}$ regressions

of each of the missing variables on the observed ones. Estimates of the conditional means $\boldsymbol{\mu}_{i.mis}$ are the regression imputations and can be applied for continuous variables for which the linear model is a good approximation, without necessarily assuming normality.

An estimator of the coefficient matrix $\boldsymbol{B}_{mis(i),obs(i)}$ can be obtained from an estimator of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{x}$ by using

$$\boldsymbol{B}_{mis(i),obs(i)} = \boldsymbol{\Sigma}^{-}_{obs(i).obs(i)}\boldsymbol{\Sigma}_{obs(i).mis(i)} \tag{8}$$

with $\boldsymbol{\Sigma}_{obs(i).obs(i)}$ the submatrix of $\boldsymbol{\Sigma}$ containing the (co)variances of the variables observed in record i and $\boldsymbol{\Sigma}_{obs(i).mis(i)}$ the submatrix containing the covariances among the variables observed in record i and the variables missing in this record. Note that once we have estimated the covariance matrix $\boldsymbol{\Sigma}$ and mean vector $\boldsymbol{\mu}$ for all variables, we can perform all regressions needed to impute each of the records, with their different missing data patterns, by extracting the appropriate submatrices and subvectors. In (8) we used a generalised inverse, denoted by "$-$", instead of a regular inverse because the covariance matrix involved can be singular due to linear dependencies of the variables implied by equality constraints.

A nice property of this multivariate regression approach with all observed variables as predictors is that linear dependencies in the data used to estimate $\boldsymbol{\Sigma}$ will be transferred to each imputed record. Therefore, all equality edits will be satisfied by the imputed data provided that $\boldsymbol{\Sigma}$ is estimated on data consistent with these edits (cf. De Waal et al. 2011, ch. 9). A possible data set to be used for estimation is the set of complete and consistent records from the current data. If there are not (yet) enough such records, cleaned data from a previous round of the survey provide an alternative. If the current data are used it is possible to also include the records with missing values in the estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by applying an EM algorithm (see Little and Rubin 2002).

3.5. Adjustment of Imputed Values for Consistency

Imputed values will often violate the edit rules, since most imputation methods do not take the edit rules into account. The multivariate regression approach (7) takes equalities into account but not inequalities. More involved imputation methods have been developed that can take all edit rules into account (De Waal et al. 2011, Ch. 9), but for many unsupervised routine applications such models become too complex. The inconsistency problem can then more easily be solved by the introduction of an adjustment step in which adjustments are made to the imputed values, so that the record satisfies all the edits and the adjustments are as small as possible. This can be seen as an optimisation problem: minimise the adjustments under the constraint that all edits are satisfied. When the weighted least squares criterion is chosen to measure the discrepancy between the unadjusted and the adjusted values, this problem can be formalised as

$$\boldsymbol{x}_{adj} = \underset{\boldsymbol{x}\in\mathbb{R}^n}{\arg\min}\,(\boldsymbol{x}-\boldsymbol{x}_{unadj})^T\boldsymbol{W}(\boldsymbol{x}-\boldsymbol{x}_{unadj})$$

$$\text{subject to } \boldsymbol{E}\boldsymbol{x}_{adj} \odot \boldsymbol{b}, \tag{9}$$

where it is understood that only the imputed values may be changed; the other elements of $\boldsymbol{x}_{adj}$ remain equal to the corresponding elements of $\boldsymbol{x}_{unadj}$. The matrix $\boldsymbol{W}$ is a positive diagonal matrix with weights that determine the amount of adjustment for each of the variables; adjustments to variables with large weights have more impact on the criterion value and therefore these variables are adjusted less than variables with small weights. For instance, the choice $\boldsymbol{W} = diag(\boldsymbol{x}_{unadj})^{-1}$ leads to minimisation of the squared discrepancies relative to the size of the unadjusted values; see Pannekoek and Zhang (2011) for more details. For a different approach in the context of sequential imputation, see Pannekoek et al. (forthcoming).

3.6. *Selection of Units for Further Treatment*

Automatic treatment cannot be expected to find and repair all important errors and consequently some form of additional manual treatment will be needed. The selection of units for manual treatment is the essential part of selective editing. The goal of this approach is to identify units for which it can be expected that manual treatment has a significant effect on estimates of totals and other parameters of interest and to limit manual review to those units.

An important tool in this selection process is the score function (Latouche and Berthelot 1992; Lawrence and McDavitt 1994; Lawrence and McKenzie 2000; Hedlin 2003) that assigns values to records that measure the expected effect of editing. The record score is usually built up from local scores for a number of important variables. Each local score measures the significance for the variable of concern. Often it can be decomposed into a *risk* component that measures the likelihood of a potential error, and an *influence* component that measures the contribution or impact of that error on the estimated target parameter. The local score for variable j in record i can then be expressed as $s_{ij} = I_{ij} \times R_{ij}$ with I_{ij} the influence component and R_{ij} the risk component for variable j in record i. See, for example, Di Zio and Guarnera in this issue, for an example of a local score function with risk and influence components. A record- or unit-level score is a function of local scores, that is $S_i = f(s_{i1}, \ldots, s_{iJ})$. The measure of risk is commonly based on the deviation of a variable from a reference value, often a historical value or stratum median. Large deviations from the reference value indicate a possible erroneous value and, if it is indeed an error, a large correction.

Since the local score and the record score reflect the occurrence and size of outlying values with respect to the reference values, the score can be seen as a quantitative measure for an aspect of the quality of a record. In this sense it is a verification function (cf. Section 4). Its purpose, however, is selection, and this can be accomplished by comparing the scores with a predetermined threshold value and selecting the units with score values higher than the threshold for manual editing. Alternatively, the units can be ordered with respect to their score values and manual editing can proceed according to this ordering, until some stopping criterion is met.

In practice we also see other, simpler selection functions being applied. The following are some examples.

- A function that identifies units that are "crucial" because they dominate the totals in their branch; selected units will be reviewed manually, whether they contain suspect values or not (selection on influence only).

- A function that selects influential units for which automatic imputation is not considered an accurate treatment because some main variables are missing or obviously incorrect; selected units will be recontacted.
- A function that selects noninfluential units for which automatic imputation is not considered an accurate treatment because some main variables are missing or obviously incorrect; selected units will be treated as unit nonresponse, for instance by weighting techniques in the estimation phase after editing is completed.
- A function that selects units for which an automatic action has failed, for instance, if the error localisation took too much time and the process was stopped without having obtained a solution. Selected units can be treated as unit nonresponse or reviewed manually, depending on their influence.

For some recent theoretical developments in the field of selective editing, see Arbués et al. and Di Zio and Guarnera in this issue.

4. The Data-Editing Process

Much of the complexity in the design of a data-editing system is caused not by mathematical difficulties relating to the underlying methods, but by combining the implementation of those methods into a working process or supporting system. A typical data-editing process consists of a mixture of domain-specific error correction and localisation actions, a number of automated editing steps, and a possibility for manual intervention on selected records. Each part of such a process has its own input, output, and control parameters that influence how it can be combined with other steps to build up a full process.

To design, compare and evaluate data-editing processes it is useful to have a common terminology for the *types of activities* that are instrumental in realising the end result of a data-editing process. In line with Camstra and Renssen (2011) we call these types of activities *statistical functions*. In Subsection 4.1 below we propose a decomposition of the overall data-editing process in a taxonomy of statistical functions that are characterised by the kind of task they perform and the kind of output they produce. The effects of these statistical functions can be evaluated by inspecting their characteristic output.

A statistical function describes *what* type of action is performed but leaves unspecified *how* it is performed. To implement a statistical function for a specific data-editing application (discussed in Subsection 4.2), a method for that function must be specified and configured. It should be noted that the same statistical function can, and often will, be implemented by several methods even within the same application. For instance, the statistical function (or type of task or kind of activity) *record selection* can be implemented by both a score function methodology and by graphical macro-editing.

An actual implementation of a data-editing process can now be seen as a collection of implementations of statistical functions. The overall process can be structured by dividing it into subprocesses or *process steps*, that each implement one or several (but related) statistical functions executed by specified methods. Process steps are application-specific but the statistical functions that they implement are much more general and are used to categorise the kinds of activities implemented by the process steps. The granularity in which a process is divided into process steps is, to an extent, arbitrary. For example, one

may talk about a statistical process as the complete process from gathering input data to publishing results, and divide that process into process steps using the GSBPM (UNECE Secretariat 2009). In that model, data editing occurs as a single step. For our purposes however, it is natural to define a more fine-grained approach.

The choice of methods to be used in the process steps and the order in which the process steps are executed will depend on the properties and requirements of the specific application at hand, but some general considerations regarding these choices are discussed in Subsection 4.3.

4.1. A Taxonomy of Data-Editing Functions

Just like process steps, statistical functions may be separated on several levels of granularity. In Figure 2 we decompose data editing hierarchically, in three levels, into ultimately six low-level statistical functions.

At the first level of the decomposition we distinguish between functions that leave the input data intact (*compute indicator*) and those that alter the input data (*amend values*). At the second level, functions are classified according to their purpose. We distinguish between indicators that are used to verify the data against quality requirements (*verification*) and indicators that are used to separate a record or data set into subsets (*selection*). *Verification* functions are separated further into functions that verify hard (mandatory) edit rules (*rule checking*) and functions that compute softer quality indicators (*compute scores*). The *selection* function allows for different records (*record selection*) or different fields in a record (*field selection*) to be treated differently. There is no separation based on purpose for the *amendment* function; *amendment* functions are only separated into functions that alter observed values (*amend observations*) and functions that alter unit properties (*amend unit properties*) such as classifying (auxiliary) variables. This may be interpreted as a decomposition based on a record-wise or field-wise action.

It should be recognised that there are many other dimensions along which one could separate the types of tasks performed in a data-editing process. For example, Pannekoek

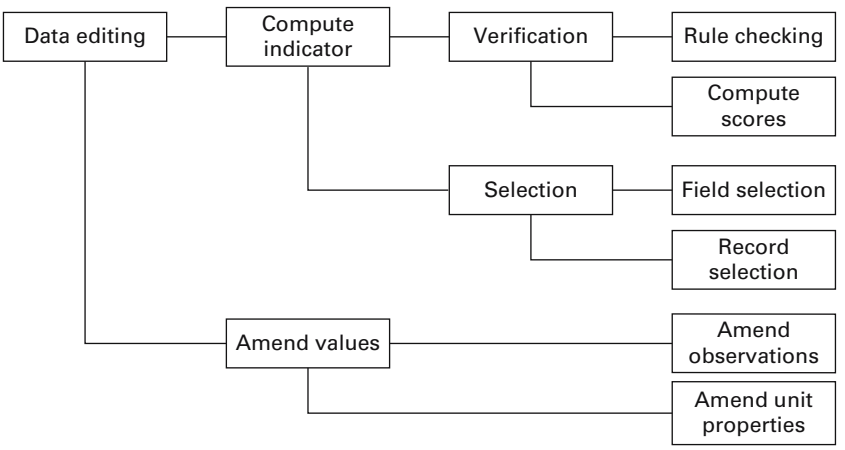

Fig. 2. A taxonomy of data-editing functions. Each data-editing function has its own minimal input-output profile which determines how they may be combined in a data editing process (Table 1)

and Zhang (2012) distinguish between methods that can be performed on a per-record basis (e.g., Fellegi-Holt error localisation, imputation with historical values) and actions that need batch processing (e.g., error localisation by macro-editing, imputation with current means). The point of view we take here is that we wish the taxonomy to abstract from implementation issues. The lowest-level statistical functions defined here allow one to define quality indicators for each function, in terms of their effect on data, performance, expense, and so on, which are independent of the chosen statistical method or implementation thereof. Below, the six lowest-level data-editing functions are discussed in some detail.

Rule checking. This verification function checks, record by record, whether the value combinations in a record are in the allowed region of the space of possible records. Such a task may be done automatically, when the rules and possible reference data are available in a machine-readable format, or manually, by expert review.

Compute scores. The score function computes a quality indicator of a record or field. Examples of score functions are counting the number of missings in a record, determining whether a field contains an outlier, or counting the number of edits violated by a field. The output of score functions is often input for automated selection functions. Score functions are rarely computed manually.

Field selection is used to point out fields in records that need a different treatment than the remaining fields, for example because they are deemed erroneous. Selection may be done manually by expert review, or automatically. Examples of automated methods include detection of unit of measurement errors, and Fellegi and Holt's method for error localisation.

Record selection aims to select records from a data set that need separate processing. This can be done automatically, for example by comparing the value of a score function to a predefined threshold value. Manual record selection is commonly based on macro-editing methods, such as sorting on a score function, reviewing aggregates, and graphical analyses.

Amend observations. This function aims to improve data quality by altering observed values or by filling in missing values. Many automated imputation and adjustment methods exist, some of which have been discussed in Section 3. The amendment function can also be performed manually, for example by data-editing staff who may recontact respondents.

Amend unit properties. This function does not alter the value of observed variables but amends auxiliary properties relating to the observed unit. In business statistics, this function entails tasks like changing erroneous business classification codes (NACE codes) and is often performed manually. Another commonly performed task falling into this category is the adjustment of estimation weights for representative outliers.

Pannekoek and Zhang (2012) and Camstra and Renssen (2011) also proposed a decomposition of statistical functions related to data editing. The former distinguish between the *verification*, *selection* and *amendment* functions, while the latter also distinguish *calculation of score functions*. The taxonomy in the current article further completes the picture by assigning data-editing functions a place in a hierarchy based on clearly defined separating principles (amend or not at the first level and select or verify at the second level).

4.2. *Specification of Data-Editing Functions*

As mentioned in the beginning of section 4, each function in the taxonomy of Figure 2 can be performed with several methodologies, and each methodology may be implemented in several ways. The operationalisation of a function for a specific data-editing process can therefore be made by documenting the input, the output, and the method. Indeed, Camstra and Renssen (2011) propose such a specification model for general statistical functions, shown in Figure 3. In principle, a data-editing process is completely determined once the order of process steps and their specifications are known.

As an example, consider a simple *record selection* function that compares a score value to a threshold value. The input consists of a score value s and a threshold value t, so the data model for the input is $\mathbb{R}^2$. The method specification is the algorithm

$$\mathsf{IF}\,(s > t)\,\mathsf{return}\,(\mathsf{TRUE})\,\mathsf{ELSE}\,\mathsf{return}\,(\mathsf{FALSE}),$$

so the output data model is $\{\mathsf{FALSE}, \mathsf{TRUE}\}$.

The above algorithm is a very simple example of how a *selection* function may be implemented. In our taxonomy, the work of Di Zio and Guarnera and Arbués et al. presented elsewhere in this issue also falls into the category of *record selection* functions, even though the methods described there are much more advanced. The most important commonality between *record selection* functions is the type of output they produce, namely a decision for each record on whether it should be selected or not. Regardless of the method used, such an output can be represented as a boolean vector with the number of records in the data set as dimension. On the input side, any effective *record selection* function will at least need the data to be able to return a reasonable decision vector. At this level of abstraction, even wildly different methods may be compared to support decisions about which method to use in which process step. Indeed, the taxonomy described in this article has been designed with such a purpose in mind.

Just like for the *record selection* function, it is possible to identify a minimal set of input and output parameters for each data-editing function, regardless of the method that implements it. Table 1 denotes this set of minimal in- and output parameters for every low-level statistical function of the taxonomy. Any extra in- or output parameter used in a particular process will be related to the specific method chosen to implement a function. The taxonomy and input-output model presented above make no assumptions about the type of data or type of rule sets. For example, the model leaves undecided whether each data record has the same number of variables, or whether the data have a hierarchical structure (as used in household surveys). Furthermore, there are no assumptions about the

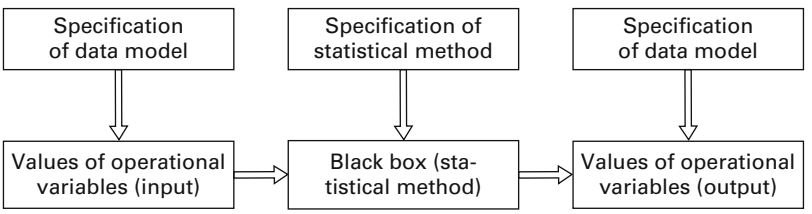

Fig. 3. A model to specify the operationalisation of statistical functions (Camstra and Renssen 2011). Besides statistical data, the values of operational variables include auxiliary information and control parameters at the input side and process metadata and quality indicators at the output side

Table 1. The minimal input and output for data-editing functions. The input data consist of N records, where the number of variables may vary per record. Each record is subject to K rules

Function	Input	Output
Rule checking	Data, rules	$N \times K$ edit failure indicator
Compute scores	Data	N-vector of score values
Field selection	Data, rules	Field selection indicator
Record selection	Data	N-vector of subset indicators
Amend observations	Data	Data
Amend unit properties	Unit properties	Unit properties

type of rules used; they may be numerical, linear, nonlinear, categorical, of mixed type or otherwise specified.

4.3. Combining Process Steps

An overall editing process can be seen as a combination of process steps, each consisting of one or more statistical functions executed by specified methods. The choice of methods that implement these functions, as well as the specifications of parameters or models for these methods, will differ between applications and depend on the data to be edited, availability of auxiliary data, output and timeliness requirements, and so on. Moreover, the order in which process steps will be carried out is also application-dependent. However, some general considerations about the composition of process steps in terms of statistical functions, the order of application and the choice of methods will be outlined below.

A single process step can combine several functions that will always be applied together. For instance, correction of generic and domain-specific systematic errors typically involves the implementation of a *field selection* function by a method that detects a specific systematic error and an *amend observations* function to replace the erroneous value with a corrected one. Since the detection is always followed directly by the correction action specific for the kind of error detected, it is meaningful to combine these two functions into a single process step with data and rules as input and a field selection indicator as well as modified data as output. The indicator reflects the detection part and the modified data the amendment part.

Several process steps will often perform the same statistical function but with different methods. In particular, *amend observations* refers to a large group of process steps that each implement a different method to solve a different problem in (possibly) different data values. An overall process will often include steps that perform the following amendment tasks: correction of generic and domain-specific systematic errors, deductive imputation, model-based imputation and adjustment of imputed values for consistency.

Although the ordering of process steps can differ between applications, there is a logical order for some process steps. For instance, selection for interactive treatment itself can occur at different stages of the editing process, but it is evident that for efficiency reasons such a selection step should always precede the actual manual amendment of values. In addition, if the selection is performed by a score-function methodology, the calculation of scores must precede the selection step. Furthermore, automatic amendment steps will usually start by exhausting the possibilities for solving systematic errors and deductive imputation before approximate solutions by model-based imputation are applied.

Timeliness of the results is an important requirement that influences the choice of methods for the statistical functions. For surveys where the data collection phase extends over a considerable period of time, it is important that the time-consuming manual editing starts as soon as possible, that is, as soon as the first data arrive. Selection for manual editing should then be based on a score function that can be evaluated on a record-by-record basis without the need to wait until all or a large part of the data are available. On the other hand, for administrative data or surveys with a short data collection period, selection for interactive treatment can be done using macro-editing methods that by definition use a large part of the data.

5. Numerical Illustrations

5.1. Introduction

In this section, we illustrate the effects of applying a sequence of automatic and manual editing functions using two real data sets. Both data sets come from regular production processes at Statistics Netherlands. The first example concerns data on Dutch childcare institutions (Subsection 5.2); the second example concerns SBS data on Dutch wholesalers (Subsection 5.3).

For both examples, we have identified the following possible process steps that can be applied during editing.

1. Correction of generic and domain-specific systematic errors:
   (a) Correction rules for falsely negative values
   (b) Correction of uniform thousand errors
   (c) Other direct correction rules
   (d) Correction of simple typing errors
   (e) Correction of sign errors
   (f) Correction of rounding errors
2. Automatic error localisation (under the Fellegi-Holt paradigm)
3. Deductive imputation of missing or discarded values
4. Model-based imputation of missing or discarded values
5. Adjustment of imputed values for consistency
6. Selection for interactive treatment
7. Manual editing (interactive treatment)

The first six numbered steps were treated in Section 3 as part of our overview of automatic editing methods. Step 7 is the only one considered here that requires real-time human input. The other steps can be run automatically once they have been set up. As was suggested in Subsection 4.3, a large number of different editing processes can be obtained by combining some (not necessarily all) of the above process steps, possibly in a different order. In general, different choices will have a different impact on the quality of the output data and on the efficiency of the editing process. This will be illustrated in the examples below.

Some brief remarks on the implementation now follow. All the numerical experiments reported below have been performed in the R statistical environment. Definition and checking of edit rules can be done with the editrules package of De Jonge

and Van der Loo (2012). Typing, sign, and rounding errors can be corrected, while taking edit rules into account, with the deducorrect package of Van der Loo et al. (2011). The deducorrect package also offers functionality to reproducibly apply user-defined domain-specific actions, as discussed in Subsection 3.2. The term "reproducibly" here means that every action performed on the records is automatically logged, while the user can configure the conditional actions independently of the source code defining the data-editing process. Error localisation for numeric, categorical or mixed data can be done with the editrules package. See De Jonge and Van der Loo (2011) for an introduction. Deductive imputation methods are again included in the deducorrect package. See Van der Loo and De Jonge (2011) for a description. For model-based imputation a multivariate regression method is applied, implemented in R. Imputed values are adapted using the rspa package of Van der Loo (2012). The code used for selecting records for manual editing and for repairing thousand errors is not part of any package and has been developed for the purpose of this article.

5.2. *Data on Childcare Institutions*

In this illustration we will show the effects of a sequence of automatic editing functions in terms of the amount of errors detected and the number of resulting amendments to data values. The data used for this example are taken from a census among institutions for child day care in 2008. Apart from questions on specific activities, the questions and the structure of the questionnaire are similar to what is typical for structural business statistics. For this illustration a subset of the census data was used, consisting of 840 records with 45 variables. For these variables 40 hard edit rules were specified, of which 11 are equalities, 27 are non-negativity edits and the remaining two are other inequalities. The edit rules as well as the rules for detecting thousand errors and domain-specific generic errors are subsets of the rules used in production.

We have applied the automatic process steps 1 through 5 listed in Subsection 5.1 to these data. The results are displayed in Table 2. The second column of this table shows the number of changed data values at each process step. In the third column are the numbers of failed edits after each process step, which can be obtained directly from the failed-edits matrix. Some edits cannot be evaluated for some records because the edit contains variables with missing values in that record. The corresponding elements of the failed-edits matrix are then missing; the number of such missing elements is listed in the column *Not evaluated edits*. The number of missing data values is in the last column.

The first line of Table 2 shows that before automatic editing there are, in the whole data set, 258 edit violations and 158 edits that cannot be evaluted because of 124 missing values. As a first automatic step, 9 false minus signs are removed by a simple direct rule for a variable that is not part of any equality edit. Obviously 9 non-negativity edit failures are resolved by this action. The detection of uniform thousand errors is applied within the revenues, costs and results section separately and 17 such errors are found. However, the number of violated edits is increased by one. By looking at the difference between the failed-edits matrix before and after the correction for thousand errors, it appears that the newly failed edit is *Total revenues* − *Total costs* = *Pre-tax result* and that this occurs because a thousand error was detected in the revenues and pre-tax result, but not in the

Table 2. Numbers of values changed, edit violations and missings at each step of a sequence of automatic editing functions

Process step	Changed values	Violated edits	Not eval. edits	Missings
0. None	0	258	158	124
1a. Rules for false minus signs	9	249	158	124
1b. Thousand errors	17	250	158	124
1c. Other direct rules	43	252	158	124
1d. Simple typing errors	53	187	158	124
1e. Sign errors	0	187	158	124
1f. Rounding errors	102	147	158	124
2. Error localisation	215	0	477	339
3. Deductive imputation	161	0	248	178
4. Model-based imputation	178	109	0	0
5. Adjustment of imputed values	144	0	0	0

costs. Records with thousand error corrections that break edit rules should be followed up manually because falsely correcting a thousand error is bound to have influential effects on estimates. The next step concerns the application of other direct rules, resulting in 43 corrections. Again, some of these changes cause edit failures that should be followed up manually, not only to correct the data but also to see how these direct correction rules can be modified so that they become consistent with the edit rules.

We now apply the algorithms for resolving simple typing errors, sign errors and rounding errors discussed in Subsection 3.1.2. There are 53 typing errors detected and corrected, of which 12 appear to be sign errors. These corrections are very effective in removing errors as the number of violated edit rules is reduced by 65. After the correction of sign errors in Step 1a and 1d, the algorithm for more complex sign errors (Step 1e) could not detect any additional sign errors. Rounding errors (Step 1f) are also important since 40 of the edit violations can be explained by such errors and correcting them with the algorithm mentioned in Subsection 3.1.2 prevents the necessity of treating these violations by the computationally intensive error localisation in Step 2. Separating the trivial rounding errors from other, more important, errors also clarifies our picture of the data quality.

At this stage, the possibilities for correction of generic and domain-specific systematic errors are exhausted. The remaining inconsistencies and missing values are resolved by applying steps 2 through 5. Error localisation (Step 2) identifies 215 values that need to be changed in order to be consistent with all edit rules. These values are treated as missing in the following process steps. The increase in missing values also increases the number of not evaluated edits to a great extent. To impute the missing values, deductive imputation (Step 3) is tried first and succeeds in filling in close to half of the missing values with the unique values allowed by the edit rules. For the remaining 178 missing values, the multivariate regression method of Subsection 3.4.2 (Step 4) is applied. These imputed values again result in edit violations. However, contrary to the situation prior to Step 2, the violation of an edit rule is now not caused by a measurement error in some (probably only a few) variables but by the fact that all model-based imputations are only approximations to the real values. Therefore (Step 5) we adjust the imputed values as little as possible and solve the 109 edit violations and a complete and consistent data set results.

5.3. Data on Wholesale

For a second illustration, we consider a data set of 323 records from the Dutch SBS of 2007. The data are on businesses with ten employed persons or more from the sector wholesale in agricultural products and livestock. The survey contains 93 numerical variables. These should conform to 120 linear edits, of which 19 are equalities.

In terms of the possible process steps listed in Subsection 5.1, the editing process that was actually used in production consisted of Steps 1(abc) and 6, followed by Step 7 for the selected records and by Steps 2, 4, and 5 for the rest. Selection for interactive treatment was based on a score function for businesses with less than 100 employed persons. Businesses with 100 employed persons or more were always edited manually. In addition, the model-based imputations in Step 4 were obtained from a linear regression model with one predictor separately for each variable. We use the outcome of this production process as a benchmark.

The second column of Table 3 shows the mean values of twelve key variables in the production-edited data set. The third column shows the corresponding means for the unedited data (ignoring all missing values). Prior to editing, the observed means of all financial variables are much too high, which reflects the presence of thousand errors in the unedited data. Moreover, while the production-edited means satisfy basic accounting rules such as *Total operating revenues* = *Net turnover* + *Other operating revenues* (apart from rounding effects), the unedited means do not.

The above editing process involves a substantial amount of manual editing: the number of records selected for interactive treatment was 142, or 44% of all records (representing about 84% of total net turnover in the production-edited data set). We now look at two different set-ups that involve less manual editing. The first alternative editing process is almost entirely automated. It consists of the above numbered process Steps 1(abcdef) and 2 through 5 (in that order). Step 7 is included as a fall-back to treat records for which automatic error localisation fails. The second alternative process is almost the same, but we add Steps 6 and 7 at the end, with a simple selection mechanism that sends all businesses with 100 employed persons or more to manual editing. Note that both alternative editing processes contain the deductive correction methods 1(def) and a deductive imputation step, which were not used in production. These additional steps are expected to improve the quality of automatic editing.

To compare the outcome of these alternative editing processes to our benchmark, we simulated the results in R. In the implementation of the process steps, we mostly followed the methodology originally used in production.

We only made changes to the model-based imputation and adjustment steps. For model-based imputation, we did not use a separate regression model for each variable but instead simultaneous regression with all variables, as explained in Subsection 3.4.2. For the adjustment step, linear optimisation was used in production, but here we used quadratic optimisation as implemented in the rspa package. Manual editing was simulated by copying the production-edited values. The number of manually edited records under the first alternative strategy was 4 (about 1% of all records, also representing about 1% of total net turnover). Under the second alternative strategy, this number was 34 (about 11% of records, but 55% of total net turnover).

Table 3. Unweighted means of (nonmissing) values of key variables in the SBS wholesale data. The first ten rows contain rounded multiples of 1,000 Euro; the last two rows are in units

Variable	Benchmark	Unedited		Alternative I		Alternative II	
Total operating revenues	59,342	80,151	(+35%)	62,180	(+4.8%)	59,338	(−0.0%)
Net turnover	59,158	79,546	(+34%)	61,652	(+4.2%)	59,157	(−0.0%)
Other operating revenues	185	655	(+255%)	528	(+186%)	182	(−1.7%)
Total operating costs	57,795	77,996	(+35%)	60,314	(+4.4%)	57,793	(−0.0%)
Purchasing costs	52,864	68,703	(+30%)	55,359	(+4.7%)	52,861	(−0.0%)
Depreciations	302	466	(+54%)	303	(+0.2%)	302	(+0.1%)
Personnel costs	2,446	5,641	(+131%)	2,460	(+0.6%)	2,446	(−0.0%)
Other operating costs	2,183	3,258	(+49%)	2,192	(+0.4%)	2,185	(+0.1%)
Operating results	1,547	2,574	(+66%)	1,866	(+21%)	1,545	(−0.1%)
Pre-tax results	1,898	2,670	(+41%)	1,983	(+4.4%)	1,919	(+1.1%)
Employed persons (count)	63.9	64.9	(+1.6%)	64.5	(+1.0%)	64.4	(+0.8%)
Employed persons (FTE)	50.8	49.5	(−2.6%)	47.1	(−7.2%)	47.9	(−5.6%)

The rightmost columns in Table 3 show the means of key variables for both alternative editing strategies. We see that the first alternative yields large differences with respect to the benchmark for several variables. Moreover, with one exception, all differences are positive. Thus it appears that for these data, relying completely on automatic editing does not produce an acceptable result. By contrast, these second alternative yields values that are close to the benchmark for all variables but one. For nine of the twelve key variables, the relative difference is less than one percent. It is interesting to note that automatic editing appears to have an adverse effect on the quality of the variable *Employed persons* (*FTE*). This may be explained by the fact that the hard edits contain relatively little information about this variable: it is only involved in two inequality edits, whereas the other key variables are all involved in at least one equality edit.

The above results suggest that for this data set, some of the manual work could be replaced by automatic editing without affecting the quality of the main output. However, a more thorough analysis would be required before we can draw this conclusion. For one thing, we did not take the sampling design into account. Moreover, other quality indicators are important besides the unweighted means of key variables. The purpose of this analysis was merely to illustrate the effects of different editing strategies on real-world data.

6. Discussion and Conclusions

In this article we have discussed the relation between automated and manual (selective) data editing from three different viewpoints. The first viewpoint we take is that of the source of error. As it turns out, data-editing staff spend considerable time editing data that are not observed survey data. Often, classifying variables from the business register (e.g., NACE codes) have to be altered as well. The source of error (overcoverage) is then not a measurement error of the survey but an error in the population register. The amendments proposed by editors in such cases are usually based on unstructured information such as websites. Moreover, such amendments often have consequences for other statistical processes, for example when a centrally maintained variable (such as the NACE code) needs to be adapted.

The second viewpoint we take is from the current state of the art in automated data editing. The image emerging from the discussion and numerical examples is that established automated methods tend to perform well for the majority of records, provided that hard edit rules have been defined and sufficient structured auxiliary information is available for the estimation of new values. Exceptions include mostly records of large businesses; these usually have a more complex structure than small establishments and data-editing staff often use external unstructured information to repair such records. Obviously, automated methods are better suited for (computationally, mathematically) complex calculations than data-editing staff. On the other hand, data-editing staff are better at judging the violation of soft edit rules, often again by using unstructured auxiliary information.

Thirdly, we have discussed the relationship between manual and automated data editing from the point of view of process design. We have decomposed the data-editing process into several types of tasks (statistical functions), which are independent of how they are implemented: manually or automatically. This allowed us to separate the tasks which are

currently easier to implement manually from those that may be implemented automatically. Here, we find that record selection, possibly supported by macro-editing tools, as well as judging and amending unit properties, are often performed manually.

Table 4 summarises the above discussion. We may conclude that currently, automated methods serve very well to edit observed variables in business survey records of establishments that are not overly complex (large) and are restricted by hard edit rules. Automated methods are not yet suited to repairing records related to large, complex companies, records under soft restrictions or performing amendments based on unstructured data. Those tasks are still mostly performed manually. Of course, manual editing of observed variables of simple (small) units based on structured information is always possible; our point is that here the same quality can often be achieved more efficiently with automated methods.

The decomposition of data editing into different statistical functions given in Section 4 allows one to assess a data-editing process on a task-by-task basis. This leads to a more refined complementation of automatic editing by (selective) manual editing than what emerges from the classical literature on selective editing. Evaluation of the results of automatic editing tasks also enables one to select the best automatic method for a specific task and thus minimise manual actions related to that task. Furthermore, the statistical functions in the decomposition each have their own set of minimal, well-defined inputs and outputs which are independent of the method used to implement the function. This modular approach to data editing offers clear potential for the development of reuseable components, yielding efficiency gains in process design.

To conclude, we see the following research opportunities. First of all, standardised quality aspects of the statistical functions identified in Section 4 should be developed. Such aspects could be, for instance, the fraction of false negatives or positives in the selection of suspicious units or erroneous fields, the predictive accuracy of imputed values obtained by some imputation method or the reduction in bias of estimates due to different amendment functions. This, then, would allow data-editing processes to be compared in a

Table 4. Relative strengths and weaknesses of manual and automated data editing

	Editing mode	
Aspect	Manual	Automated
Variable		
Observed variable	−	+
Unit property	+	−
Edit rules		
Hard edits	−	+
Soft edits	+	−
Use of aux. information		
Structured	−	+
Unstructured	+	−
Type of unit		
Simple	−	+
Complex	+	−

standardised way and paves the way for further development of reusable components based on various methodologies. Secondly, data-editing research should focus on areas where automated data editing is currently less suitable (Table 4). Interesting fields of research are the use of unstructured information to verify and/or amend data and the use of soft edits in automated data editing. Some recent progress in the latter field was made by one of the authors (Scholtus and Göksen 2012; Scholtus 2013). The use of, for example, web-scraping or text-mining techniques in data editing remains largely unexplored.

7. References

Al Hamad, A., Lewis, D., and Silva, P.L.N. (2008). Assessing the Performance of the Thousand Pounds Automatic Editing Procedure at the ONS and the Need for an Alternative Approach. Working Paper No. 21, UN/ECE Work Session on Statistical Data Editing, Vienna. Available at: http://www.unece.org/stats/documents/2008.04.sde.html (accessed October 2013).

Bethlehem, J. (2009). Applied Survey Methods. Wiley series in survey methodology. New York: John Wiley & Sons, Inc.

Boskovitz, A. (2008). Data Editing and Logic: the Covering Set Method from the Perspective of Logic. Ph. D. thesis, Australian National University.

Camstra, A. and Renssen, R. (2011). Standard Process Steps Based on Standard Methods as Part of the Business Architecture. Proceedings of the 58th World Statistical Congress (Session STS044): International Statistical Institute, 1–10. Available at: http://2011.isiproceedings.org/ (accessed October 2013).

Damerau, F. (1964). A Technique for Computer Detection and Correction of Spelling Errors. Communications of the ACM, 7, 171–176.

De Jonge, E. and Van der Loo, M. (2011). Manipulation of Linear Edits and Error Localization with the Editrules Package. Technical Report 201120, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl/en-GB/menu/methoden/onderzoekmethoden/discussionpapers/archief/2011/default.htm (accessed October 2013).

De Jonge, E. and Van der Loo, M. (2012). Editrules: R Package for Parsing and Manipulating of Edit Rules and Error Localization, R package version 2.5. Available at: http://www.cbs.nl/en-GB/menu/methoden/onderzoekmethoden/discussionpapers/archief/2012/default.htm (accessed October 2013).

De Waal, T. and Quere, R. (2003). A Fast and Simple Algorithm for Automatic Editing of Mixed Data. Journal of Official Statistics, 19, 383–402.

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. Wiley handbooks in survey methodology. New York: John Wiley & Sons.

De Waal, T., Pannekoek, J., and Scholtus, S. (2012). The Editing of Statistical Data: Methods and Techniques for the Efficient Detection and Correction of Errors and Missing Values. Wiley Interdisciplinary Reviews: Computational Statistics, 4, 204–210. DOI: http://dx.doi.org/10.1002/wics.1194

Di Zio, M., Guarnera, U., and Luzi, O. (2005). Editing Systematic Unity Measure Errors Through Mixture Modelling. Survey Methodology, 31, 53–63.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71, 17–35.

Granquist, L. and Kovar, J.G. (1997). Editing of Survey Data: How Much is Enough? In Survey Measurement and Process Quality. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwartz, and D. Trewin (eds). Wiley series in probability and statistics. New York: Wiley, 416–435.

Groves, R.M. (1989). Survey Errors and Survey Costs. Wiley series in survey probability and mathematical statistics. New York: John Wiley & Sons, Inc.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistics, 19, 177–199.

Latouche, M. and Berthelot, J.-M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. Journal of Official Statistics, 8, 389–400.

Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earning. Journal of Official Statistics, 10, 437–447.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243–253.

Levenshtein, V.I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Soviet Physics Doklady, 10, 707–710.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data (second Edition). New York: John Wiley & Sons.

Pannekoek, J., Shlomo, N., and de Waal, T. (forthcoming). Calibrated Imputation of Numerical Data Under Linear Edit Restrictions. Annals of Applied Statistics.

Pannekoek, J. and Zhang, L.-C. (2011). Partial (donor) Imputation with Adjustments. Working Paper No. 40, UN/ECE Work Session on Statistical Data Editing, Ljubljana. Available at: http://www.unece.org/stats/documents/2011.05.sde.html (accessed October 2013).

Pannekoek, J. and Zhang, L.-C. (2012). On the General Flow of Editing. Working Paper No. 10, UN/ECE Work Session on Statistical Data Editing, Oslo. Available at: http://www.unece.org/stats/documents/2012.09.sde.html (accessed October 2013).

Scholtus, S. (2009). Automatic Correction of Simple Typing Errors in Numerical Data with Balance Edits. Technical Report 09046, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl/en-GB/menu/methoden/onderzoekmethoden/discussionpapers/archief/2009/default.htm (accessed October 2013).

Scholtus, S. (2011). Algorithms for Correcting Sign Errors and Rounding Errors in Business Survey Data. Journal of Official Statistics, 27, 467–490.

Scholtus, S. (2013). Automatic Editing with Hard and Soft Edits. Survey Methodology, 39, 59–89.

Scholtus, S. and Göksen, S. (2012). Automatic Editing with Hard and Soft Edits – Some First Experiences. Technical Report 201225, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl/en-GB/menu/methoden/onderzoekmethoden/discussionpapers/archief/2012/default.htm (accessed October 2013).

UNECE Secretariat (2009). Generic Statistical Business Process Model version 4.0. Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata.

Van der Loo, M. (2012). rspa: Adapt Numerical Records to Fit (in)Equality Restrictions with the Successive Projection Algorithm. R package version 0.1-1. Available at: http://cran.r-project.org/web/packages/rspa/index.html (accessed October 2013).

Van der Loo, M. and De Jonge, E. (2011). Deductive Imputation with the Deducorrect Package. Technical Report 201126, Statistics Netherlands, The Hague. Available at: http://www.cbs.nl/en-GB/menu/methoden/onderzoekmethoden/discussionpapers/archief/2011/default.htm (accessed October 2013).

Van der Loo, M., De Jonge, E., and Scholtus, S. (2011). Deducorrect: Deductive Correction, Deductive Imputation, and Deterministic Correction. R package version 1.3-1. Available at: http://cran.r-proiect.org/web/packages/deducorrect/index.html (accessed October 2013).

Received January 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 539–555, http://dx.doi.org/10.2478/jos-2013-0039

A Contamination Model for Selective Editing

Marco Di Zio[1] *and Ugo Guarnera*[1]

The aim of selective editing is to identify observations affected by influential errors. A score function based on the impact of the potential error on target estimates is useful to prioritize observations for accurate reviewing. We assume a Gaussian model for true data and an "intermittent" error mechanism such that a proportion of data is contaminated by an additive Gaussian error. In this setting, scores can be related to the expected value of errors affecting data. Consequently, a set of units can be selected such that the expected residual error in data is below a prefixed threshold. In the context of economic surveys when positive variables are analyzed, the method is more realistically applied to logarithms of data instead of data in their original scale. The method is illustrated through an experimental study on real business survey data where contamination is simulated according to error mechanisms frequently encountered in the practical context of economic surveys.

Key words: Statistical data editing; influential errors; finite mixture models; score function.

1. Introduction

Selective editing is based on the idea of looking for observations containing important errors in order to focus the treatment only on them thus reducing the cost of the editing and imputation phase (E&I), while maintaining a desired level of quality of estimates (Granquist 1997; Lawrence and McKenzie 2000; Lawrence and McDavitt 1994). The underlying assumption is that the true values for the selected units can be obtained through follow-up or interactive treatment. In practice, observations are prioritized according to the importance of errors expressed by the values of a score function (Latouche and Berthelot 1992; Hedlin 2003), and units having a value of the score function above a given threshold, are selected for a careful treatment.

The most commonly used methods to determine the scores are based on the difference between observed and predicted values. This difference is composed of the possible measurement error and the prediction error. When only raw data are available, traditional methods do not allow the estimation of these two elements separately, hence scores are not directly related to the expected errors. The consequence is that the value of the selective editing threshold will not be directly interpretable as a level of accuracy of estimates of interest and it will be difficult to find a stopping rule related to the expected level of quality of estimates.

The introduction of a contamination model naturally leads to building a score function as defined in Jäder and Norberg (2005). It is defined in terms of a risk component

[1] ISTAT, Italian National Institute of Statistics, Via Cesare Balbo 16, 00184 Rome, Italy. Emails: dizio@istat.it and guarnera@istat.it

(the probability of being in error) and an influence component (the magnitude of error), and allows the estimation of the expected error associated with each unit. In particular, the contamination normal model is characterized by peculiarities that make it useful for the problems generally treated by selective editing. In fact, it is usually applied to deal with gross errors (see Ghosh-Dastidar and Schafer 2006) and is based on a latent variable addressing the status of error for each observation. The latent variable describes the intermittent nature of the errors generally affecting surveys carried out by National Statical Institutes (NSIs) where in fact only a part of the observations are affected by errors. In order to make the model useful in practice, it is extended to deal with lognormal variables, to manage the presence of auxiliary variables not affected by errors (for instance in the case of administrative variables), and to cope with missing values. As far as incomplete observations are concerned, usual methods may lack of possibility of computing a set of consistent and comparable scores. In our setting, the score is coherently computed by taking into account the relevant marginal distribution obtained from the estimated multivariate distribution. In the proposed approach, the scores can be interpreted as expected errors, and a threshold can be determined such that the expected error of the target estimates due to residual errors left in data is below a predefined value. An algorithm to select the units to be edited is also proposed. Although the contamination model, the score function and the selection algorithm are presented as parts of a unique procedure, they can be used separately in different selective editing strategies.

Some experiments showed that the procedure can be usefully applied even when there are some departures from the model assumptions (Buglielli et al. 2011). It is currently used in some Istat surveys such as the *Building permits survey*, the *Structure of earning surveys*, and the *Information and communication technology survey*.

The selective editing procedure is modularly implemented in an R package named SeleMix (Buglielli and Guarnera 2011) that is available on the R website (http://cran.r-project.org/).

The article is structured as follows. The contamination model is described in Section 2 where, in particular, it is explained how to obtain predictions for each single observation (Subsection 2.1). The algorithm to estimate the model parameters is illustrated in Section 3. The use of the model in presence of missing data is presented in Section 4. Section 5 describes the application of the contamination model in the selective editing setting, and in particular a proposal for a score function and for a selection criterion is given. In Section 6 we present an experimental application to illustrate the approach and to empirically evaluate its properties. Concluding remarks are given in Section 7.

2. True Data Model and Error Mechanism

An important feature of the proposed model is that it explicitly takes into account the fact that only a proportion of survey data are affected by errors, that is, the error mechanism has an intermittent nature. Data may be partitioned in two groups: error-free data and contaminated data, the membership of each unit being unknown. This naturally leads to modelling the observed data through a latent class model, where the latent variable is a binary variable to be interpreted as an error indicator variable. When the interest is focused on the identification of gross errors, one possible approach consists in specifying

a distribution for the observed data as a mixture of two probability distributions corresponding to error-free and contaminated data respectively. This is the approach followed, for instance, by Ghosh-Dastidar and Schafer (2006), that uses the membership posterior probabilities to asses the degree of outlyingness of each observation. In the context of selective editing however, one is mostly interested in identifying errors having high impact on some estimate of interest, rather than in identifying implausible observations. Thus there is the need to estimate the error magnitude. This can be done if the distribution of the "true" unobserved data and the error mechanism are specified separately. In particular, the error mechanism is specified via the conditional distribution of observed data given true data. In the following, the true data model and the error mechanism are described in detail.

We suppose that true unobserved data are independent realizations of p-variate random vectors $\boldsymbol{Y}_i^* = (\boldsymbol{Y}_{i1}^*, \ldots \boldsymbol{Y}_{ip}^*)'$, $i = 1, \ldots, n$, whose distributions are Gaussian with mean vectors $\boldsymbol{\mu}_i$ and common covariance matrix $\boldsymbol{\Sigma}$. Furthermore, it is assumed that on each sampled unit i a (possibly empty) set of q covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iq})'$ is also available and that $\boldsymbol{\mu}_i = \boldsymbol{B}'\boldsymbol{x}_i$, where $\boldsymbol{B}$ is a $q \times p$ matrix of unknown coefficients. The previous hypotheses can be expressed in matrix form as

$$\boldsymbol{Y}^* = \boldsymbol{XB} + \boldsymbol{U} \tag{1}$$

where $\boldsymbol{Y}^*$ is the $n \times p$ true data matrix, $\boldsymbol{X}$ is the $n \times q$ covariate matrix, and $\boldsymbol{U}$ is the $n \times p$ matrix of normal residuals whose rows are independent realizations of Gaussian random vectors with zero mean and covariance matrix $\boldsymbol{\Sigma}$.

Hereafter, the notation $f(\boldsymbol{v})$ will denote the generic marginal probability distribution (or density) for the random variable $\boldsymbol{V}$. Analogously, $f(\boldsymbol{v}, \boldsymbol{w})$ and $f(\boldsymbol{v}|\boldsymbol{w})$ will denote joint and conditional distributions involving variables $\boldsymbol{V}$ and $\boldsymbol{W}$. Thus, for instance, for the ith unit, $f(\boldsymbol{y}_i^*)$ and $f(\boldsymbol{u}_i)$ are the marginal distributions of the true value and of the residual respectively. From the previous assumptions:

$$f(\boldsymbol{y}_i^*) = N(\boldsymbol{y}_i^*; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}), \quad f(\boldsymbol{u}_i) = N(\boldsymbol{u}_i; \boldsymbol{0}, \boldsymbol{\Sigma}), \quad i = 1, \ldots, n, \tag{2}$$

where, as usual, $N(\boldsymbol{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian density with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

We assume that presence of errors in data is governed by n independent Bernoulli random variables $I_i, (i = 1, \ldots, n)$ with parameter π, that is, $I_i = 1$ if an error occurs on unit i and $I_i = 0$ otherwise. Furthermore, given that an error is present on the ith unit (i.e., given the event $\{I_i = 1\}$), its action is described through an additive random noise represented by a p-variate random Gaussian variable $\boldsymbol{\epsilon}_i$ with zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ proportional to $\boldsymbol{\Sigma}$. If $\boldsymbol{Y}$ denotes the data matrix associated with the observed (possibly contaminated) data and $\boldsymbol{\epsilon}$ the error matrix whose ith row is $\boldsymbol{\epsilon}_i'$, we can formally express the error mechanism as:

$$\boldsymbol{Y} = \boldsymbol{Y}^* + \boldsymbol{I\epsilon}, \;\; f(\boldsymbol{\epsilon}_i) = N(\boldsymbol{\epsilon}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}), \;\; \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = (\alpha - 1)\boldsymbol{\Sigma}, \tag{3}$$

where α is a numeric constant greater than 1, and $\boldsymbol{I}$ is a diagonal $n \times n$ matrix whose diagonal elements are the Bernoullian variables $I_1, \ldots, I_n$. Equivalently, we can specify the error model through the conditional distribution:

$$f(\boldsymbol{y}|\boldsymbol{y}^*) = (1 - \pi)\delta(\boldsymbol{y} - \boldsymbol{y}^*) + \pi N(\boldsymbol{y}; \boldsymbol{y}^*, \boldsymbol{\Sigma}_\epsilon). \quad (4)$$

where π (mixing weight) represents the "a priori" probability of contamination and $\delta(\boldsymbol{t}' - \boldsymbol{t})$ is the delta-function with mass at $\boldsymbol{t}$.

In the previous model, the crucial aspect is the intermittent nature of the error implied by the introduction of the Bernoullian variables I_i. Due to this assumption, it is conceptually possible to think of data as partitioned into the two groups of error-free and contaminated data, and to estimate, for each observation, the posterior probability of group membership, i.e., the probability of being error-free or contaminated. This is the key aspect of the proposed approach to selective editing. In fact, as we will see, differently from most selective editing methods, the "suspiciousness" of each observation is naturally incorporated in the model through the posterior probabilities.

Once the true data distribution and the error mechanism have been specified, the distribution of the observed data can also be easily derived through multiplying the true data density by the error density (4), and integrating over $\boldsymbol{y}^*$. The resulting distribution is:

$$f(\boldsymbol{y}_i) = (1 - \pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma}). \quad (5)$$

Expression (5) represents a mixture of two regression models having the same coefficient matrix $\boldsymbol{B}$ and proportional residual variance-covariance matrices. This distribution can be estimated by maximizing the likelihood based on n sample units via an ECM algorithm (see Meng and Rubin 1993). Details are provided in Section 3.

2.1. *Predictions*

The contamination model can be used to obtain predictions or "anticipated values" for true unobserved data. The separate specification of true data model and error model allows, contrarily to the direct specification of the observed data distribution, to derive, for $i = 1, \ldots, n$, the distribution $f(\boldsymbol{y}_i^*|\boldsymbol{y}_i)$ of the true data conditional on the observed data, where we have suppressed the reference to the $\boldsymbol{X}$ variates in the notation. A straightforward application of the Bayes formula provides:

$$f(\boldsymbol{y}_i^*|\boldsymbol{y}_i) = \tau_1(\boldsymbol{y}_i)\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i) + \tau_2(\boldsymbol{y}_i)N(\boldsymbol{y}_i^*; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}) \quad (6)$$

where

$$\tilde{\boldsymbol{\mu}}_i = \frac{(\boldsymbol{y}_i + (\alpha - 1)\boldsymbol{\mu}_i)}{\alpha}; \quad \tilde{\boldsymbol{\Sigma}} = \left(1 - \frac{1}{\alpha}\right)\boldsymbol{\Sigma},$$

$\delta(\boldsymbol{y}_i^* - \boldsymbol{y}_i)$ is the delta function with mass at $\boldsymbol{y}_i$, and $\tau_1(\boldsymbol{y}_i)$, $\tau_2(\boldsymbol{y}_i)$ are the posterior probabilities that a unit with observed values $\boldsymbol{y}_i$ belongs to the correct or erroneous data group respectively:

$$\tau_1(\boldsymbol{y}_i) = Pr(\boldsymbol{y}_i = \boldsymbol{y}_i^*|\boldsymbol{y}_i) = \frac{(1 - \pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{(1 - \pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma})},$$

$$\tau_2(\boldsymbol{y}_i) = Pr(\boldsymbol{y}_i \neq \boldsymbol{y}_i^*|\boldsymbol{y}_i) = 1 - \tau_1(\boldsymbol{y}_i),$$

$$i = 1, \ldots, n.$$

In order to make the meaning of Formula (6) clear, let us consider the univariate case in absence of covariates ($E(Y^*) = \mu$). Let σ^2, σ_ϵ^2 denote the variances of true data and errors respectively, and define $\alpha = \left(\sigma^2 + \sigma_\epsilon^2\right)/\sigma^2$. Then it is easily seen that the mean $\tilde{\mu}_y$ of the second component of the mixture (6) is given by $\left(\sigma_\epsilon^{-2} y + \sigma^{-2}\mu\right)/\left(\sigma^{-2} + \sigma_\epsilon^{-2}\right)$. In other words, given that the observed value $\boldsymbol{y}$ is not correct, the expectation of the corresponding true value is a weighted mean of the observed value y and the unconditioned mean μ with weights proportional to the inverse of the variances σ^2 and σ_ϵ^2 respectively. Moreover, the variance $\tilde{\sigma}^2 = \left(\sigma^{-2} + \sigma_\epsilon^{-2}\right)^{-1}$ is lower than both σ^2 and σ_ϵ^2, that is, the knowledge of the error mechanism reduces the uncertainty about y^* and the knowledge of the true data model reduces the uncertainty about the evaluation of the error $y - y^*$ that actually occurred.

It is natural to define predictions in terms of the conditional expected value $\tilde{\boldsymbol{y}}_i = E(\boldsymbol{y}_i^* | \boldsymbol{y}_i)$. From (6) it follows:

$$\tilde{\boldsymbol{y}}_i = \tau_1(\boldsymbol{y}_i)\boldsymbol{y}_i + \tau_2(\boldsymbol{y}_i)\tilde{\boldsymbol{\mu}}_i, \quad i = 1, \ldots, n. \tag{7}$$

Correspondingly, we can define the expected error as

$$\boldsymbol{y}_i - \tilde{\boldsymbol{y}}_i = \tau_2(\boldsymbol{y}_i)(\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}_i).$$

The last expression makes it natural to interpret τ_2 and $\boldsymbol{y}_i - \tilde{\boldsymbol{\mu}}_i$ as "risk component" and "influence component" respectively to be considered in the score function definition. In practice, parameters involved in expected errors are unknown, and have to be estimated. The algorithm to obtain maximum likelihood estimates (MLE) of the parameters is described in Section 3, and their use in a score function is illustrated in Section 5.

We remark that in the context of economic surveys, when positive variables are analyzed, logarithms of data instead of data in their original scale are often modeled through a Gaussian distribution. The above methodology can be easily adapted to the lognormal case. In this case the error model assumed for data in original scale is multiplicative; more precisely, contaminated data are related to true data by means of the relation

$$\boldsymbol{Z} = \boldsymbol{Z}^* \boldsymbol{e}^{\boldsymbol{\epsilon}}$$

where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_\epsilon)$.

For $i = 1, \ldots, n$, let $\boldsymbol{Y}_i^* = \ln \boldsymbol{Z}_i^*$, $\boldsymbol{Y}_i = \ln \boldsymbol{Z}_i$, where $\boldsymbol{Z}_i^*$ and $\boldsymbol{Z}_i$ represent the variables associated with true and contaminated data respectively, and $\boldsymbol{Y}_i^*, \boldsymbol{Y}_i$ are modeled as previously illustrated (Formulas 2–6). The distribution of $\boldsymbol{Z}_i^*$ given z_i is:

$$f(z_i^* | z_i) = \tau_1(\ln(z_i))\delta(z_i^* - z_i) + \tau_2(\ln(z_i))LN(z_i^*; \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Sigma}}), \tag{8}$$

where $LN(\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the lognormal density with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

3. Estimation

In this section, the algorithm to obtain MLEs of the model parameters is described. The log-likelihood to be maximized is:

$$\sum_{i=1}^{n} \log f_i(\boldsymbol{y}_i),$$

where

$$f_i(\boldsymbol{y}_i) = (1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma}),$$

and $\boldsymbol{\mu}_i = \boldsymbol{B}'\boldsymbol{x}_i$. An ECM algorithm is used and it consists in repeatedly applying, until convergence, the following steps:

E-step

$$\tau_1(\boldsymbol{y}_i) = \frac{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})}{(1-\pi)N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) + \pi N(\boldsymbol{y}_i; \boldsymbol{\mu}_i, \alpha\boldsymbol{\Sigma})}$$
$$\tau_2(\boldsymbol{y}_i) = 1 - \tau_1(\boldsymbol{y}_i)$$
$$i = 1, \ldots, n.$$

CM-step

(M1) *update the mixing weight* (π)

$$\pi = \frac{1}{n}\sum_{i=1}^{n} \tau_2(\boldsymbol{y}_i)$$

(M2) *update regression parameters* (**B**)

$$\boldsymbol{B} = (\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{Y}$$

(M3) *update covariance matrix* ($\boldsymbol{\Sigma}$)

$$\boldsymbol{\Sigma} = \frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})}{n}$$

(M4) *update variance inflation parameter* (α)

$$\alpha = \frac{trace\left\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})'\boldsymbol{\tau}_2^D(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}\right\}}{q\pi}$$

where:

$$\boldsymbol{\Omega} \doteq \boldsymbol{\tau}_1^D + \frac{\boldsymbol{\tau}_2^D}{\alpha},$$

and $\boldsymbol{\tau}_j^D$ denotes the diagonal matrix of which the ith diagonal element is $\tau_j(\boldsymbol{y}_i)$, $j = 1, 2$. Note that, in (M1)–(M4), maximization with respect to model parameters is not simultaneous but conditional on the other parameters remaining fixed. This make the convergence of the algorithm, convergence slower than it would be in a genuine EM algorithm. In order to initialize the algorithm we use as starting points for $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ the

estimates of the corresponding parameters obtained through ordinary linear squares (OLS) based on all data. A random value for π in the interval [0.6, 1] is chosen, and α is initialized with some reasonable value, for instance $\alpha \in [5, 10]$.

In case of log-normal data, the ECM algorithm has to be applied to logarithms of data.

In the following, the MLEs will be denoted by $\hat{\pi}, \hat{\boldsymbol{B}}, \hat{\boldsymbol{\Sigma}}, \hat{\alpha}$. Analogously, $\hat{\tau}_1(\boldsymbol{y}_i)$, $\hat{\tau}_2(\boldsymbol{y}_i)$ and $\hat{\boldsymbol{\mu}}_i$ will denote the estimates of $\tau_1(\boldsymbol{y}_i)$, $\tau_2(\boldsymbol{y}_i)$ and $\boldsymbol{\mu}_i$.

4. Incomplete Data

The previous methodology can be easily extended to situations where observed data are incomplete and the nonresponse mechanism is assumed to be MAR. According to the usual notation for incomplete data, the equality $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i,o}, \boldsymbol{Y}_{i,m})$ means that the random vector $\boldsymbol{Y}_i$ can be partitioned in two subvectors $\boldsymbol{Y}_{i,o}$, $\boldsymbol{Y}_{i,m}$ corresponding to the observed and missing items respectively for the ith unit. The partition induces a similar decomposition for the starred variables: $\boldsymbol{Y}_i^* = (\boldsymbol{Y}_{i,o}^*, \boldsymbol{Y}_{i,m}^*)$. Note that by definition, the $\boldsymbol{Y}^*$ variables are never observed, so that partitioning is determined only by the missing pattern of the contaminated variables. According to the partition of $\boldsymbol{Y}$ and $\boldsymbol{Y}^*$ vectors, we obtain the partition of all relevant vectors and matrices. The matrix $\boldsymbol{\Sigma}$ can be partitioned as:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{om} \\ \boldsymbol{\Sigma}_{mo} & \boldsymbol{\Sigma}_{mm} \end{pmatrix}$$

so that, analogously to the complete data case, we can define matrices $\tilde{\boldsymbol{\Sigma}}_{oo}$ and $\tilde{\boldsymbol{\Sigma}}_{mm}$ as $(1 - 1/\alpha)\boldsymbol{\Sigma}_{oo}$ and $(1 - 1/\alpha)\boldsymbol{\Sigma}_{mm}$ respectively.

In the same manner, for each missing pattern, we can partition the matrix $\boldsymbol{B}$ as $\boldsymbol{B} = [\boldsymbol{B}_o, \boldsymbol{B}_m]$, where the columns of matrices $\boldsymbol{B}_o$ and $\boldsymbol{B}_m$ correspond to observed and missing variables respectively. Furthermore, for $i = 1, \ldots, n$, let:

$$\boldsymbol{\mu}_{i,o} = \boldsymbol{B}_o' \boldsymbol{x}_i; \quad \boldsymbol{\mu}_{i,m} = \boldsymbol{B}_m' \boldsymbol{x}_i; \quad \tilde{\boldsymbol{\mu}}_{i,o} = \frac{(\boldsymbol{y}_{i,o} + (\alpha - 1)\boldsymbol{\mu}_{i,o})}{\alpha}; \quad \tilde{\boldsymbol{\mu}}_{i,m} = \frac{(\boldsymbol{y}_{i,m} + (\alpha - 1)\boldsymbol{\mu}_{i,m})}{\alpha}.$$

Our goal is to estimate, for $i = 1, \ldots, n$, the conditional distribution of $\boldsymbol{Y}_i^*$ given $\boldsymbol{Y}_{i,o}$. We have:

$$f(\boldsymbol{y}_i^* | \boldsymbol{y}_{i,o}) = f(\boldsymbol{y}_{i,o}^*, \boldsymbol{y}_{i,m}^* | \boldsymbol{y}_{i,o}) = \frac{f(\boldsymbol{y}_{i,o} | \boldsymbol{y}_{i,o}^*, \boldsymbol{y}_{i,m}^*) f(\boldsymbol{y}_{i,m}^* | \boldsymbol{y}_{i,o}^*) f(\boldsymbol{y}_{i,o}^*)}{f(\boldsymbol{y}_{i,o})}. \tag{9}$$

From the assumed error model in Formula (3), each observed variable, conditionally on the corresponding true variable, is independent of all other true variables, thus we can rewrite (9) as

$$f(\boldsymbol{y}_i^* | \boldsymbol{y}_{i,o}) = \frac{f(\boldsymbol{y}_{i,o} | \boldsymbol{y}_{i,o}^*) f(\boldsymbol{y}_{i,o}^*)}{f(\boldsymbol{y}_{i,o})} f(\boldsymbol{y}_{i,m}^* | \boldsymbol{y}_{i,o}^*). \tag{10}$$

The fraction in (10) is the conditional density of $\boldsymbol{Y}^*_{i,o}$ given $\boldsymbol{y}_{i,o}$ and can be obtained from Formula (6) of Subsection 2.1:

$$\frac{f(\boldsymbol{y}_{i,o}|\boldsymbol{y}^*_{i,o})f(\boldsymbol{y}^*_{i,o})}{f(\boldsymbol{y}_{i,o})} = f(\boldsymbol{y}^*_{i,o}|\boldsymbol{y}_{i,o}) = \tau_1(\boldsymbol{y}_{i,o})\delta(\boldsymbol{y}^*_{i,o} - \boldsymbol{y}_{i,o}) + \tau_2(\boldsymbol{y}_{i,o})N(\boldsymbol{y}^*_{i,o}; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo}).$$

Thus, we can write:

$$f(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) = \tau_1(\boldsymbol{y}_{i,o})f_1(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) + \tau_2(\boldsymbol{y}_{i,o})f_2(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}),$$

where

$$f_1(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) = \delta(\boldsymbol{y}^*_{i,o} - \boldsymbol{y}_{i,o})f(\boldsymbol{y}^*_{i,m}|\boldsymbol{y}^*_{i,o}) = \delta(\boldsymbol{y}^*_{i,o} - \boldsymbol{y}_{i,o})f(\boldsymbol{y}^*_{i,m}|\boldsymbol{y}_{i,o}) \tag{11}$$

$$f_2(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o}) = N(\boldsymbol{y}^*_{i,o}; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo})f(\boldsymbol{y}^*_{i,m}|\boldsymbol{y}^*_{i,o}). \tag{12}$$

Both conditional densities in (11) and (12) can be obtained from that of a bipartitioned multivariate normal distribution. The density (11) can be directly derived from the true-data distribution (2). The density $f_2(\boldsymbol{y}^*_i|\boldsymbol{y}_{i,o})$ is normal, but the derivation is somewhat more involved. It is thus possible to obtain closed expressions of the expected true values given the observed ones. The adaption of these results to the log-normal distribution is straightforward. All the details are given in the Appendix.

As far as parameter estimation is concerned, we have used the ECM algorithm described in Section 3 on completely observed data. This approach is a suboptimal and could be properly modified in order to take into account also incomplete observations. The adaption of the ECM algorithm is a topic for a future study.

5. Selective Editing and Score Function

The score function is the main tool to prioritize observations according to the impact of errors on target estimates. It is natural to think of the score function as an estimate of the error affecting data. The estimate is generally based on comparing observed with predicted values, taking into account the probability of being in error (suspiciousness). The latter element arises from the implicit assumption that only a certain proportion of data is affected by error, or, from a probabilistic perspective, that each measured value has a certain probability of being erroneous. When the degree of suspiciousness is not taken into account a large proportion of false alarms is expected, as noted in several case studies by Norberg et al. (2010).

Prediction and suspiciousness are usually combined to form a score for a single variable, named local score. An example of local score for the unit i with respect to the variable Y_j when the target quantity to be estimated is the total $t^*_j = \sum_{i=1}^{N} y^*_{ij}$ in a population $\mathcal{P}$ of size N is:

$$S_{ij} = \frac{p_i w_i |y_{ij} - \hat{y}_{ij}|}{t^{ref}_j}$$

where p_i is a degree of suspiciousness, y_{ij} is the observed value of the variable Y_j on the ith unit, $\hat{y}_{ij}$ is the corresponding prediction, w_i is the sampling weight, and t^{ref}_j is a reference estimate of the target parameter t^*_j. A review can be found in De Waal et al. (2011).

When the interest is on more than one variable, the local scores can be combined to form a global score GS_i, examples of global scores are $GS_i = \sum_j S_{ij}$, or $GS_i = \max_j S_{ij}$, see Hedlin (2008).

The global score is used to evaluate the impact on the target estimates of the errors remaining in the unedited observations. To this aim, observations are ordered by their global score and all the units with a score above a threshold value are selected. The threshold should be chosen so that the impact on the target estimates of the errors remaining in the unedited observations is negligible.

The evaluation of the impact of errors remaining in data and so of the threshold is generally done through a simulation study based on raw and edited data from a previous occasion of the same survey (De Waal et al. 2011). This approach is based on the assumption that the edited data can be considered as true data and that the error mechanism and the data distribution are the same in the two survey occasions. Moreover it cannot be applied when raw and edited data from previous occasions of the survey are not available.

In our setting, the introduction of a model allows to define a score function that can be interpreted as an estimate of the expected error of the observation, and consequently the threshold value η can be directly linked to the level of accuracy of the estimates of interest.

The proposed score function for the total t_j^* is based on the relative individual error for the ith unit with respect to the variable Y_j. The latter is defined as the ratio between the (weighted) expected error and the reference estimate t_j^{ref} of the target parameter t_j^*, that is

$$r_{ij} = \frac{w_i(y_{ij} - \hat{y}_{ij})}{t_j^{ref}}, \tag{13}$$

where the prediction $\hat{y}_{ij}$ for the variable Y_j on the ith unit is obtained plugging in the MLE of the parameters in the conditional expectation as expressed in Formula (7). The local score function is defined as

$$S_{ij} = |r_{ij}|. \tag{14}$$

Note that, the estimated expected error is $\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i = \hat{\tau}_2(\boldsymbol{y}_i)(\boldsymbol{y}_i - \hat{\tilde{\boldsymbol{\mu}}}_i)$, that is the product of the probability of being in error, $\hat{\tau}_2$, and the difference $(\boldsymbol{y}_i - \hat{\tilde{\boldsymbol{\mu}}}_i)$ between the observed value and the expectation of the true value conditional on the event that $\boldsymbol{y}_i$ is contaminated. Hence, S_{ij} can be seen as composed of a "risk component" $\hat{\tau}_2(\boldsymbol{y}_i)$ and an "influence component" $w_i(\boldsymbol{y}_i - \hat{\tilde{\boldsymbol{\mu}}}_i)$.

In the next paragraph, an algorithm for the selection of units to be accurately edited is described. For $i = 0, 1, \ldots, n$, let us define R_{ij} as the absolute value of the expected residual relative error for the variable Y_j remaining in data after removing errors in the first i ordered units (when $i = 0$ no observations are selected), that is $R_{ij} = \left|\sum_{k>i}^{n} r_{kj}\right|$. Once an accuracy level (threshold) η is chosen, the selective editing procedure consists of:

1. sorting the observations in descending order according to the value of S_{ij};
2. finding $n_e \equiv n_e(\eta)$ such that $n_e = min\{k^* \in (0, 1, \ldots, n) | R_{kj} < \eta, \ \forall k \geq k^*\}$, that is, selecting the first n_e units such that all the residual errors R_{kj} (for a given j) computed from the $(n_e + 1)$th to the last observation are below η.

This procedure implies that the absolute value of the expected difference between the estimator $\hat{t}_j^e$ computed on edited data and the estimator $\hat{t}_j^*$ computed on true data is below the accuracy level ηt_j^{ref}. Furthermore, $S_{kj} < 2\eta, \ \forall k > n_e$ for each unit not revised, implying that also the error at micro level is bounded.

The algorithm described so far is easily extended to the multivariate case by defining a global score function $GS_i = max_j S_{ij}$. The two-step algorithm is:

1. order the observations with respect to GS_i (decreasing order);
2. find n_e such that $n_e = min\{k^* \in (0, 1, \ldots, n) | max_j R_{kj} < \eta, \ \forall k \geq k^*\}$, that is, select the first n_e units such that all the residual errors R_{kj} computed from the $(n_e + 1)$th to the last observation are below η.

The above accuracy properties are still valid for all the variables. In fact,

$$\left| E\left(\hat{t}_j^e - \hat{t}_j^*\right) \right| < \eta t_j^{ref}, \ j =, 1, \ldots, J$$

and $S_{kj} < 2\eta, \ \forall k > n_e, j = 1, \ldots, J$.

We remark that different values of the parameter η can be set for the analyzed variables in order to take into account their possible different importance.

The reference estimate t_j^{ref} in Formula (13) can be computed by using the predictions $\hat{y}_{ij}$ obtained by the contamination model,

$$\hat{t}_j^{ref} = \sum_{i=1}^{n} w_i \hat{y}_{ij}.$$

As an alternative, reference estimates can be obtained by using data from a previous survey occasion.

6. Application to Real Data

In this section we describe an experimental evaluation based on data from the 2008 Istat *Survey on small and medium enterprises*. The application refers to the subset of enterprises in the Nace Rev2 sections B, C, D and E corresponding to aggregation of economic activities in *Manufacturing, mining and quarrying and other industry*. This group of units ($N = 5{,}399$) has been used as the reference population (U) and for this population the variables *turnover* (X) and *labor cost* (Y) have been used, assuming that the available data are error-free. Errors are artificially introduced into the Y variable according to some error mechanisms frequently met in the context of NSI surveys; they are explicitly described in the next paragraphs. We suppose that the population parameter to be estimated is the total of the variable Y. The variable X is used as a covariate in the contamination model to obtain predictions for Y. The Gaussian contamination model is assumed for log-transformed data, according to Formula (8).

A Monte Carlo study based on 1,000 iterations has been carried out to study the performance of the proposed selective editing strategy. Each iteration of the Monte Carlo experiment consists of the following steps:

1. ***Sampling***
Draw a simple random sample without replacement (srswor) s_a of $n_a = 1{,}000$ observations from the target population U.

2. ***Data contamination***
- Multiply Y values by 1,000 in 1% of data.
- Swap the first two digits of Y values in 2% of data.
- Swap the last two digits of Y values in 2% of data.
- Replace the Y value with the value "1" in 2% of data.

3. ***Model estimation and score computation***
Compute on the logarithm of data the MLEs of the model parameters and use them to calculate the score function (14) for each unit. Order observations accordingly. In order to assess the impact of the risk component τ_2, a score function based only on the influence component $y_i - \hat{\mu}_i$ is also computed.

4. ***Selective editing***
Given the threshold η, the most influential observations n_e are selected according to the procedure described in Section 5. In an alternative experiment, we have selected $n_{\tilde{e}}$ observations according to an analogous procedure where the score function is based only on the influence component, as described in the previous step. The selected units are replaced with the corresponding true values.

5. ***Target estimates***
Compute the Horvitz-Thompson estimates of the variable Y on the true data ($\hat{t}_y^*$), on the corrupted data ($\hat{t}_y$), and on the two sets of edited data, that is, $\hat{t}_y^e$ and $\hat{t}_y^{\tilde{e}}$.

The results are summarized through the empirical relative root mean squared error (RRMSE) and the empirical relative bias (RB) based on the 1,000 Monte Carlo realizations $\hat{t}_y^{*(i)}$, $\hat{t}_y^{(i)}$, $\hat{t}_y^{e(i)}$ and $\hat{t}_y^{\tilde{e}(i)}$ ($i = 1, \ldots, 1{,}000$) of the three estimators in Step 5. The error is to be intended as deviation from the estimate based on true data $\left(\hat{t}_y^{*(i)}\right)$, because we are interested in evaluating the effectiveness of the methods regardless of the sampling error. Thus, for instance, for the estimator $\hat{t}_y^{(i)}$ RRMSE and RB are defined respectively as:

$$RRMSE = \sqrt{\frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\left(\frac{\hat{t}_y^{(i)} - \hat{t}_y^{*(i)}}{\hat{t}_y^{*(i)}}\right)^2}$$

and

$$RB = \frac{1}{1{,}000}\sum_{i=1}^{1{,}000}\frac{\hat{t}_y^{(i)} - \hat{t}_y^{*(i)}}{\hat{t}_y^{*(i)}}.$$

Empirical RRMSE and empirical RB are reported in the 3rd and 4th column of Table (1) according to different threshold levels η. The efficiency of the procedure is measured by comparing the percentage of selected units n_e% with $n_{\tilde{e}}$%, and with the percentage of observations n_{e*}% we would obtain by using true values as predictions, that is, by replacing the expression in Formula (13) with $(y_i - y_i^*)/\hat{t}_y^*$. The average percentage of selected units (n_e%) is also reported in the last column of the table.

Table 1. Empirical RRMSE, empirical RB of the estimates computed on contaminated and edited data, and average percentage of edited units according to the threshold η

η		$\hat{t}_y$	$\hat{t}_y^e$	$\hat{t}_y^{\tilde{e}}$	$n_{e*}\%$	$n_e\%$	$n_{\tilde{e}}\%$
0.05	RRMSE	10.882	0.016	0.011	1.0	1.0	12.1
	RB	9.893	−0.006	0.005	–	–	–
0.01	RRMSE	10.542	0.006	0.001	1.7	2.4	61.0
	RB	9.641	0.002	0.000	–	–	–
0.005	RRMSE	10.910	0.005	0.000	2.4	3.1	71.0
	RB	9.984	0.002	0.000	–	–	–

The impact of errors on the estimates is particularly harmful; in fact the RRMSE computed on observed data ranges from 10.54 to 10.91. After the selective editing procedure, the RRMSE dramatically decreases, and its value is (on average) below the accuracy level required and expressed by η. As far as the efficiency is concerned, the results show that n_e is close to the number of selected observations n_{e*} that would be selected in the ideal situation in which true data were known. Based on the comparison of n_e with $n_{\tilde{e}}$, we can note that not taking into account the risk component leads to the selection of a much higher number of observations.

These results are important because they show that the editing procedure performs satisfactorily even though data clearly do not satisfy the assumptions of the model; in particular the error mechanism is clearly far from the normality assumption.

In order to obtain a picture of some important parameters of the procedure, a single Monte Carlo realization is described in Figure 1 and Figure 2.

In Figure 1(a) outliers and selected observations according to $\eta = 0.01$ are reported on the scatter plot of contaminated log data. An observation is considered an outlier if

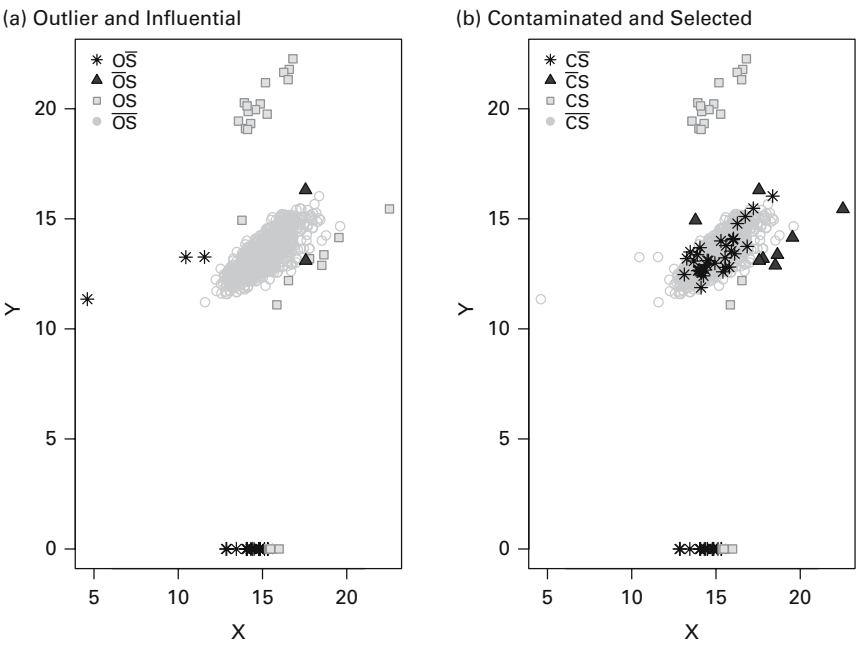

Fig. 1. Outliers, contaminated and selected observations in logdata

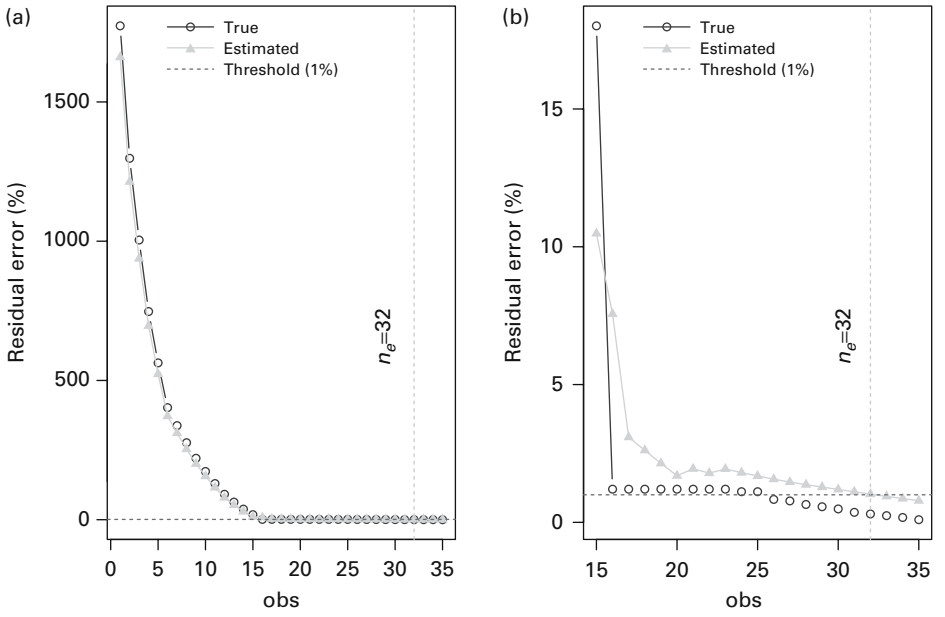

Fig. 2. Estimated vs. true residual error

the estimated conditional probability of being in error $\hat{\tau}_2(\boldsymbol{y}_i)$ is greater than 0.5. Observations classified as outliers and not selected are denoted by $O\bar{S}$, selected and not outliers by $\bar{O}S$, as both outliers and selected by OS. The remaining units that are not selected and not outliers are denoted by $\overline{OS}$.

Figure 1(b) shows contaminated units and selected observations. Observations that are contaminated but not selected are denoted by $C\bar{S}$, not contaminated units that are selected by $\bar{C}S$, contaminated and selected observations by CS. The remaining units that are not selected and not contaminated are denoted by $\overline{CS}$.

The estimated and true residual errors are reported in Figure 2(a) for the first 35 observations, while in Figure 2(b) the same residual errors are reported from the 15th observations onward in order to zoom in and avoid masking scale effects. The horizontal dashed line is the threshold and the vertical dashed line corresponds to the number of selected units in this experiment ($n_e = 32$).

Figure 2(a) and Figure 2(b) show that the estimated residual errors are close to the true residual errors. It is worth noting that the accuracy of the estimate is below the threshold even though many errors are left in the data (see Figure 2), as it is required from a selective editing procedure. As far as the outliers are concerned, it is interesting to note that not all the outliers are considered influential by the procedure, and on the other hand some selected units are not outliers. The distinction is due to the impact of the estimated error on the estimates.

7. Conclusions

In this article a model-based approach to selective editing is proposed. The considered model is referred to in the literature as a contamination model and it is typically used to

detect gross errors. The introduction of a model for both true data and error mechanism makes it possible to define a score function that can be interpreted as an estimate of the error affecting data. This allows the relation between the choice of a threshold for selection of the units to be reviewed and the level of accuracy required for the estimates to be made explicit. According to this peculiarity, an algorithm to select influential errors is proposed.

Since the remaining uncertainty due to the unedited data can be properly estimated under the model-based approach based on latent classes, it is possible to determine a threshold for the score function conditional on the actual sample observations of the current survey. By contrast, traditional methods do not assume an explicit measurement-error model and the threshold value for the score function is usually set based on edited data from previous surveys. Since the error mechanism and the data distribution do not remain exactly the same over time, the remaining uncertainty of the current unedited data can only be heuristically controlled.

The main advantages of the proposed approach are due to the introduction of an explicit model for true data and error mechanism, and of course the limits lie in the validity of the hypothesis on which the model is based. Nevertheless, the experimental studies carried out in this paper suggest that the use of a Gaussian contamination model can be usefully applied also when data and error mechanism deviate from the model assumptions, especially when data are contaminated by gross errors.

An implication of the error model described is that errors on different items are not independent of each other; this means that the intermittence nature of the error is at record level and not at variable level. Further studies should be devoted to study more general models able to encompass this assumption.

The use of edits in such a procedure is an open issue. However, some remarks are needed in this respect. Soft edits such as ratio edits are implicitly taken into account by the procedure, since the analysis of anomalous relationships between variables is the core of the proposed approach. By contrast, it is not easy to treat hard edits consistently in the model, and further analysis should be devoted to this aspect.

The editing described in the article can be classified as "output editing", meaning that a certain amount of data from the current survey is needed to estimate the model. However, it can also be used from an "input editing" perspective, in situations where the model is applied to a previous survey occasion, and the estimated parameters are used to select influential units in the current survey.

Finally, even though the article describes a strategy composed of a latent class model for predicting data and an algorithm to select influential units, they can be used independently of each other. In fact, parameter estimation, computation of predicted values and selection of influential errors are separately implemented in the R package SeleMix.

Appendix

The density in (11),

$$f_1(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o}) = \delta(\boldsymbol{y}_{i,o}^* - \boldsymbol{y}_{i,o}) f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*) = \delta(\boldsymbol{y}_{i,o}^* - \boldsymbol{y}_{i,o}) f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}),$$

is:

$$f_1(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o}) = \delta(\boldsymbol{y}_{i,o}^* - \boldsymbol{y}_{i,o})N(\boldsymbol{y}_{i,m}^*; \boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o}, \boldsymbol{\Sigma}_{m|o}).$$

In the density (12),

$$f_2(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o}) = N(\boldsymbol{y}_{i,o}^*; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo})f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*),$$

the factor $f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*)$ is the true-data conditional distribution corresponding to the missing pattern being considered, and can be derived from the true-data multivariate Gaussian distribution in Formula (2):

$$f(\boldsymbol{y}_{i,m}^*|\boldsymbol{y}_{i,o}^*) = N(\boldsymbol{y}_{i,m}^*; \boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o}^*, \boldsymbol{\Sigma}_{m|o}),$$

where

$$\boldsymbol{\alpha}_{m,i|o} = \boldsymbol{\mu}_{i,m} - \boldsymbol{\beta}_{m|o}\boldsymbol{\mu}_{i,o}, \quad \boldsymbol{\beta}_{m|o} = \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1} \quad \boldsymbol{\Sigma}_{m|o} = \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om}.$$

In order to obtain an explicit expression for the second density $f_2(\boldsymbol{y}_i^*|\boldsymbol{y}_{i,o})$, it suffices to observe that $N(\boldsymbol{y}_{i,o}^*; \tilde{\boldsymbol{\mu}}_{i,o}, \tilde{\boldsymbol{\Sigma}}_{oo})N(\boldsymbol{y}_{i,m}^*; \boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o}^*, \boldsymbol{\Sigma}_{m|o})$ is the factorisation of a multivariate Gaussian density $N(\boldsymbol{y}_{i,o}^*, \boldsymbol{y}_{i,m}^*; \bar{\boldsymbol{\mu}}_i, \bar{\boldsymbol{\Sigma}})$ of which the parameters are:

$$\bar{\boldsymbol{\mu}}_i = [\tilde{\boldsymbol{\mu}}_{i,o}', (\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\tilde{\boldsymbol{\mu}}_{i,o})']', \quad \bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \bar{\boldsymbol{\Sigma}}_{oo} & \bar{\boldsymbol{\Sigma}}_{om} \\ \bar{\boldsymbol{\Sigma}}_{mo} & \bar{\boldsymbol{\Sigma}}_{mm} \end{pmatrix},$$

where:

$$\begin{aligned}
\bar{\boldsymbol{\Sigma}}_{oo} &= \tilde{\boldsymbol{\Sigma}}_{oo} = \frac{\alpha - 1}{\alpha}\boldsymbol{\Sigma}_{oo} \\
\bar{\boldsymbol{\Sigma}}_{mo} &= \bar{\boldsymbol{\Sigma}}_{om}' = \boldsymbol{\beta}_{m|o}\bar{\boldsymbol{\Sigma}}_{oo} = \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\tilde{\boldsymbol{\Sigma}}_{oo} = \frac{\alpha - 1}{\alpha}\boldsymbol{\Sigma}_{mo} \\
\bar{\boldsymbol{\Sigma}}_{mm} &= \boldsymbol{\Sigma}_{m|o} + \bar{\boldsymbol{\Sigma}}_{mo}\tilde{\boldsymbol{\Sigma}}_{oo}^{-1}\bar{\boldsymbol{\Sigma}}_{om} = \\
&= \boldsymbol{\Sigma}_{mm} - \boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om} + \frac{\alpha - 1}{\alpha}\boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om} = \\
&= \boldsymbol{\Sigma}_{mm} - \frac{1}{\alpha}\boldsymbol{\Sigma}_{mo}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{om}.
\end{aligned}$$

From the previous formulas it follows that the expected value of Y_i^* conditional on the observed value $y_{i,o}$ is:

$$E(\boldsymbol{Y}_i^*|\boldsymbol{y}_{i,o}) = \tau_1(\boldsymbol{y}_{i,o})\boldsymbol{E}_{1i} + \tau_2(\boldsymbol{y}_{i,o})\boldsymbol{E}_{2i},$$

where

$$\begin{aligned}
\boldsymbol{E}_{1i} &= [\boldsymbol{y}_{i,o}', (\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o})']' = [\boldsymbol{y}_{i,o}', (\boldsymbol{\mu}_{i,m} + \boldsymbol{\beta}_{m|o}(\boldsymbol{y}_{i,o} - \boldsymbol{\mu}_{i,o}))']', \\
\boldsymbol{E}_{2i} &= [\tilde{\boldsymbol{\mu}}_{i,o}', (\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\tilde{\boldsymbol{\mu}}_{i,o})'] = [\tilde{\boldsymbol{\mu}}_{i,o}', (\boldsymbol{\mu}_{i,m} + \boldsymbol{\beta}_{m|o}(\tilde{\boldsymbol{\mu}}_{i,o} - \boldsymbol{\mu}_{i,o}))']'.
\end{aligned}$$

The case of incomplete log-normal data can also be easily treated, in fact with a slight shift of the notation and letting $\boldsymbol{y}_{i,0} = ln(\boldsymbol{z}_{i,o})$ we have:

$$\boldsymbol{E}(\boldsymbol{Z}_i^*|\boldsymbol{z}_{i,o}) = \tau_1(\ln(\boldsymbol{z}_{i,o}))\boldsymbol{E}_{1i}^L + \tau_2(\ln(\boldsymbol{z}_{i,o}))\boldsymbol{E}_{2i}^L,$$

where:

$$\boldsymbol{E}_{1i}^L = \left[\exp\left(\boldsymbol{y}_{i,o} + \frac{1}{2}\boldsymbol{\Sigma}_{00}^d\right)', \quad \exp\left(\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\boldsymbol{y}_{i,o} + \frac{1}{2}\boldsymbol{\Sigma}_{m|0}^d\right)'\right]$$

$$\boldsymbol{E}_{2i}^L = \left[\exp\left(\tilde{\boldsymbol{\mu}}_{i,o} + \frac{1}{2}\tilde{\boldsymbol{\Sigma}}_{00}^d\right)', \quad \exp\left(\boldsymbol{\alpha}_{m,i|o} + \boldsymbol{\beta}_{m|o}\tilde{\boldsymbol{\mu}}_{i,o} + \frac{1}{2}\tilde{\boldsymbol{\Sigma}}_{m|0}^d\right)'\right],$$

and $\boldsymbol{\Sigma}^d$ denotes the vector of the diagonal elements of the matrix $\boldsymbol{\Sigma}$.

8. References

Buglielli, M.T., Di Zio, M., Guarnera, U., and Pogelli, F.R. (2011). Selective Editing of Business Survey Data Based on Contamination Models: an Experimental Application. NTTS 2011 New Techniques and Technologies for Statistics, Brussels, 22–24 February 2011.

Buglielli, T. and Guarnera, U. (2011). SeleMix: Selective Editing via Mixture models. R package version 0.8.1. Available at: http://cran.r-project.org/web/packages/SeleMix/index.html (accessed October 9, 2013).

De Waal, T., Pannekoek, J., and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. New York: John Wiley and Sons.

Ghosh-Dastidar, B. and Schafer, J.L. (2006). Outlier Detection and Editing Procedures for Continuous Multivariate Data. Journal of Official Statistics, 22, 487–506.

Granquist, L. (1997). The New View on Editing. International Statistical Review, 65, 381–387.

Hedlin, D. (2003). Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics. Journal of Official Statistics, 19, 177–199.

Hedlin, D. (2008). Local and Global Score Functions in Selective Editing. In Proceedings of UN/ECE Work Session on Statistical Data Editing, 21–23 April, Vienna. Available at: http://www.unece.org/fileadmin/DAM/stats/documents/2008/04/sde/wp.31.e.pdf

Jäder, A. and Norberg, A. (2005). A Selective Editing Method Considering both Suspicion and Potential Impact, Developed and Applied to the Swedish Foreign Trade Statistics. In Proceedings of UN/ECE Work Session on Statistical Data Editing, 16–18 May, Ottawa. Available at: http://www.unece.org/stats/documents/2005.05.sde.htm (accessed October 9, 2013).

Latouche, M. and Berthelot, J.M. (1992). Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys. Journal of Official Statistics, 8, 389–400.

Lawrence, D. and McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. Journal of Official Statistics, 10, 437–447.

Lawrence, D. and McKenzie, R. (2000). The General Application of Significance Editing. Journal of Official Statistics, 16, 243–253.

Meng, X.L. and Rubin, D.B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: a General Framework. Biometrika, 80, 267–278.

Norberg, A., Adolfsson, C., Arvidson G., Gidlund, P., and Nordberg, L., (2010). A General Methodology for Selective Data Editing. Stockholm: Statistics Sweden. Available at: http://gauss.stat.su.se/master/statdatabaser/HT10/Literature/SwedishEditingMethods.pdf (accessed October 9, 2013).

Received January 2013
Accepted September 2013

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 557–582, http://dx.doi.org/10.2478/jos-2013-0040

Aspects of Responsive Design with Applications to the Swedish Living Conditions Survey

Peter Lundquist[1] *and Carl-Erik Särndal*[2]

In recent literature on survey nonresponse, new indicators of the quality of the data collection have been proposed. These include indicators of balance and representativity (of the set of respondents) and distance (between respondents and nonrespondents), computed on available auxiliary variables. We use such indicators in conjunction with paradata from the Swedish CATI system to examine the inflow of data (as a function of the call attempt number) for the 2009 Swedish Living Conditions Survey (LCS). We then use the LCS 2009 data file to conduct several "experiments in retrospect". They consist in interventions, at suitable chosen points and driven by the prospects of improved balance and reduced distance. The survey estimates computed on the resulting final response set are likely to be less biased. Cost savings realized by fewer calls can be redirected to enhance quality of other aspects of the survey design.

Key words: Household surveys; nonresponse; auxiliary vector; register variables; stopping rules; balance indicators; representativeness; *R*-indicator.

1. Introduction

Large nonresponse is typical of many sample surveys today. This can be a serious detriment to survey quality. Nonresponse causes systematic error, called bias, in the survey estimates. The purpose of this article is to define and apply new tools, in the spirit of responsive design, to the Swedish Survey of Living Conditions (LCS), so as to improve the data collection for this important survey, which has become affected by high nonresponse in recent years.

An extensive literature is devoted to survey nonresponse and its consequences. In dealing with the problem, statisticians need to consider (a) measures to be taken at the data collection stage, and (b) measures to be taken at the estimation stage.

With the data collection completed, the estimation stage begins, and the statistician's task is to produce estimates that are properly adjusted for the nonresponse bias still remaining, despite efforts to achieve balance or representativity at the data collection stage. The objective at the estimation stage is to achieve the best possible reduction of a nonresponse bias that can never be completely eliminated. One way to do this is by adjustment weighting, through calibration on selected auxiliary variables. Nonresponse

[1] Senior Methodologist, Statistics Sweden, Karlavägen 100, 104 51 Stockholm, Sweden. Email: peter.lundquist@scb.se

[2] Visiting Professor, Statistics Sweden, 70189 Örebro, Sweden and Örebro University. Email: carl.sarndal@scb.se

Acknowledgments: The authors are grateful to three anonymous referees for constructive comments. The results and the opinions expressed in this article are the sole responsibility of the authors.

weighting adjustment has been studied in several publications, including Särndal and Lundström (2005, 2008, 2010), Särndal (2011b).

The focus in this article lies on the data collection. The nonresponse rate measures one aspect of the data collection. It has become increasingly clear that the nonresponse rate by itself is not suitable, or at least not sufficient, for effective monitoring of the data collection. For example, it may be wasteful to continue a data collection according to an unchanging scenario driven primarily by the desire to obtain the highest possible response rate in the end, or to reach, by a costly and unrelenting effort, a predefined rate of response, such as 70% for example.

Wagner (2012 p. 557) expresses the dilemma as follows: "To the extent that response rates are not a good indicator for nonresponse bias, decisions about data-collection activities or post-survey adjustments that are made based on the response rate will be inefficient, biasing or both. Something is needed to fill this gap between response rates – which are known – and nonresponse biases – which are unknown, but are the thing about which we are really concerned." In the typology of data sources in Wagner (2012), the type that describes our approach is "the response indicator and frame data/paradata."

Two important recent concepts with implications for this article are *adaptive design* and *responsive design*. Bethlehem et al. (2011) regard responsive design as a special case of adaptive design.

At the present stage of development, adaptive design appears to refer mainly to situations where treatments applied to sampled elements are identified prior to the start of the data collection, although they may also be revised or modified during the data collection.

Responsive design is an adaptive approach where available information is used to modify the data collection for the remaining cases. The data collection may thus involve two or more phases, with decisions taken underway about subsequent phases. The general objectives of responsive design are formulated in Groves and Heeringa (2006). A number of applications of related approaches have subsequently appeared. Options for responsive design in a Canadian setting are discussed in Mohl and Laflamme (2007) and Laflamme (2009). Work on the development of adaptive designs has been presented for example in Wagner (2008). The present article draws mainly on the ideas of responsive design. Groves and Heeringa (2006) use the term "phase capacity" for "the stable condition of an estimate in a specific design phase". When phase capacity has been reached in a given phase, it is no longer effective to continue data collection in the same mode or phase; there is an incentive to modify the design, if data collection is to be continued at all.

Several directions have emerged in recent years in research on adaptive designs. The question whether a definite relationship exists between nonresponse rates and bias in the estimates is reviewed in Groves (2006). A meta-analysis on nonresponse studies is reported in Peytcheva and Groves (2009). The conclusion, somewhat pessimistic about the bias-reducing effect of demographic auxiliary variables, is that there is no strong evidence that variation in response rates across sample groups can help reduce biases in the study variables.

In the Scandinavian countries, the choice of auxiliary variables is much broader. Indicators for the data collection were developed in Schouten et al. (2009) and in Särndal (2011a). We apply the indicators to an existing data set: that of the 2009 Swedish Living

Conditions Survey (LCS). Our objective is to demonstrate how the indicators work and to suggest improvements for the data collection in future versions on the Swedish LCS.

Stopping rules for the data collection have been studied in Rao et al. (2008) and in Wagner and Raghunathan (2010).

An approach with obvious appeal is to observe changes in survey estimates, for variables that allow this, as a function of making additional contact attempts. It is one of the techniques used in this article. Related to this is the question of whether respondents interviewed early (say, in the first five attempts) differ considerably from those brought in later with respect to important measurable variables. These questions are studied in Peytchev et al. (2009) and in Peytchev et al. (2010). They conclude that focusing on groups with low response probability may not be efficient in some surveys; it may be better to identify those units with the greatest potential to induce bias in the survey variable estimates.

Responsive design may take different forms. One option is to strive for an ultimate set of respondents with measurable and favorable characteristics for the set of respondents. Especially in the later stages of the data collection intervention is permissible in order to realize an ultimate response set that is better balanced, or more representative of the total sample, than if no special effort is made. Recently proposed indicators for balance and representativity are important in this process; they are used in this article to monitor the data collection and to implement changes. Both concepts build on a specified auxiliary vector with values known for the full sample.

The 7th EU Framework Programme funded a project called RISQ, which stands for Representativity Indicators for Survey Quality; on it, see, for example Schouten and Bethlehem (2009). One of its objectives was to develop and study indicators for the *representativity* of survey response. The R-indicator (with R for Representativity) was proposed by Schouten et al. (2009) and further developed in Schouten et al. (2011). One of its uses is in comparing surveys – the same survey in different countries, or different surveys within the same country – with respect to the representativity of the final set of respondents. The statistical concept behind the R-indicator is the variance of the response probabilities, estimated with the aid of auxiliary variables. The motivation is that a small variability of such estimates would suggest a "representative set of respondents".

Indicators based on the concept of a *balanced response set* were developed in Särndal (2011a). The response set is said to be balanced if the means for specified important auxiliary variables are the same or almost the same for the set of respondents as for all those selected in the probability sample. That respondents should be on average like all those sampled is an attractive notion. The balance indicators are computable from the auxiliary variable values available for responding as well as for nonresponding units.

The present article presents general concepts for monitoring the data collection, and they are applied to the 2009 LCS. We describe the survey in Section 3, and we analyze the LCS 2009 data in Section 4. The concepts of balance (of the response set), distance (between respondents and nonrespondents), and representativity are reviewed in Section 5, then applied to the LCS 2009 data. In Section 6 we conduct several "experiments in retrospect" with the LCS 2009 data. These experiments show that balance and distance can be improved by interventions in the data collection with the aid of paradata from Statistics

Sweden's WinDATI system explained below. Implications for the future are discussed in Section 7. The theoretical framework presented in the article is general in scope, applicable to any probability sampling design.

The access to ample auxiliary information is of crucial importance. Statistics Sweden operates in a data-rich survey environment, where high quality administrative registers allow access to many auxiliary variables, particularly for surveys on individuals and households. This also applies to the other Scandinavian countries and the Netherlands. The whole issue of nonresponse adjustment will necessarily present itself in quite a different light in countries where only highly limited auxiliary information is available, say at best a few demographic variables. However, a trend towards increased availability and use of high quality administrative data is evident in many countries.

2. Earlier Experiences at Statistics Sweden

Several earlier studies at Statistics Sweden illustrate that a data collection motivated principally by a desire to achieve the best possible ultimate rate of response is inefficient. They suggest that scarce resources are being spent with little effect on the estimates and little improvement in representativity. Hörngren et al. (2008) and Lundquist and Särndal (2012) summarize several studies of surveys with telephone interviewing of individuals drawn by probability sampling from the Swedish Total Population Register (TPR). We mention them briefly here.

A study of the November 2002 edition of the Swedish Labour Force Survey (LFS) had found that the estimates change very little after the fifth contact attempt. It was concluded that a less elaborate fieldwork strategy, with say four call attempts instead of twelve, could considerably reduce the monthly cost for calls in the LFS.

A study along similar lines was carried out in 2007 for the Household Finances (HF) survey. Estimates were computed successively over the data collection period, as a function of the number of call attempts identified by "WD-events", which are events registered by the data collection instrument WinDATI. The simple expansion estimator (the mean for units having responded up until a given number of attempts) stabilizes at an early stage in the data collection: After about ten call attempts, the estimates change very little. Since the total number of call attempts for a sampled person may exceed 20, there is strong indication that resources are not effectively used. The calibration estimator based on selected auxiliary variables stabilizes even sooner, at around five call attempts.

In a later project, the effect of a follow-up strategy for the HF survey was studied. Low response rates had been observed in the primary data collection for several groups expected to have a high impact on the nonresponse error. However, it was found that the follow-up (the field work following the ordinary data collection) had little effect on the estimates, and that follow-up respondents are not the ones that influence the nonresponse error the most. In the end, the response rate remains disappointingly low for groups already underrepresented in the ordinary data collection.

An earlier study of the LCS had found that the representativity (measured by indicators) changes very little after an early point in the data collection, suggesting that the response set fails to become more similar to the selected sample. In addition, the follow-up appears to have little effect on the estimates. The present article examines the LCS in more depth.

3. The Swedish Living Conditions Survey

The Swedish LCS is a yearly sample survey. It has a long tradition of providing important information about social welfare in Sweden, in particular among different subgroups of Swedish society. It has become increasingly affected by nonresponse. The sample consists of individuals with an age of 16 or above drawn from the Swedish Total Population Register. The data set used in the analysis in this article is a subsample of $n = 8{,}220$ individuals, taken from the actual LCS 2009 sample. This subsample can be regarded as a simple random sample from the population of individuals.

Telephone interviews were conducted by a staff of interviewers using the Swedish CATI-system, WinDATI. All attempts by interviewers to establish contact with a sampled person are registered by WinDATI. Those paradata are important for this article. For every sampled individual, the WinDATI system records a series of events which we refer to as "call attempts". They play an important role in our analysis. The WinDATI events include "call without reply", "busy line", "contact with household member other than the sampled person", and "appointment booking for later contact". When contact and data delivery has occurred, the data collection effort is completed for the sampled person in question. All registered WinDATI events are taken into account in the analysis that follows.

The LCS 2009 ordinary field work lasted five weeks, at the end of which the response rate was 60.4%; for some sampled persons, 30 or more call attempts had then been recorded. This was followed by a three week break during which characteristics of non-interviewed individuals were examined in order to prepare the three week follow-up period, which concluded the data collection. All individuals considered by the survey managers to be potential respondents were included in the follow-up effort, which brought the response rate up to an ultimate 67.4%. However, there was no separate strategy or revised procedure for the follow-up. It followed the same routines as the ordinary field work. Hence, there were no attempts at responsive design where, for example, a follow-up would focus on underrepresented groups, in an objective to improve balance and reduce nonresponse bias.

In addition to these paradata, the information recorded in the LCS 2009 data set includes the response obtained on the survey target variables. In addition, it contains for all 8,220 individuals the values of a number of register variables, some of which we use as auxiliary variables. Three other register variables are used as study variables (y-variables), as explained in Section 4. For these we can compute unbiased estimates, based on the full sample, and compare them with estimates made under nonresponse.

We have chosen here to regard data inflow as a function of the attempt number rather than as a function of time evolved (Day 1, Day 2, and so on) since the start of data collection. Our analysis could have been conducted under the time-evolved perspective instead, with somewhat different results. In a CATI data collection, the attempt number concept is practical and natural.

4. An Analysis of the LCS 2009 Data

Results in this section reinforce the impression from earlier studies at Statistics Sweden that a data collection (including a follow-up) that proceeds according to an essentially unchanging format will produce very little change in the estimates beyond a certain "stability point" reached quite early in the data collection.

In this section, we study the dynamic behavior of survey estimates as the data collection proceeds. We measure the progression of the population total estimates for three variables as a function of the number of the call attempt, defined more precisely as the attempt at which an interviewer made successful contact with a sampled person and data delivery occurred. The three variables are register variables, used here as study variables. Their values are therefore known for all sampled units, not only for responding ones. To let three register variables play the role of study variables restricts to some extent the pool of available auxiliary variables, but it is a price worth paying in order to realize the methodological objectives.

Some notation is needed. The finite population $U = \{1, \ldots, k, \ldots, N\}$ consists of N units indexed $k = 1, 2, \ldots, N$. A probability sample s is drawn from U; in this sampling, unit k has the known inclusion probability $\pi_k = \Pr(k \in s) > 0$ and the known design weight $d_k = 1/\pi_k$. We denote the value of the study variable y as y_k. The target parameter for estimation is the population total $Y = \sum_U y_k$. (A sum $\sum_{k \in A}$ over a set of units $A \subseteq U$ will be written as $\sum_A$.) Normally, the survey involves many study variables and many totals to be estimated.

The response set is the set of units for which the value y_k has been recorded. Since we follow the data collection as a function of the call attempt number, there is a series of successively larger response sets. In a completely rigorous notation, we would denote these increasingly large response sets as $r^{(a)}$, where a refers to "call attempt number", $a = 1, 2, \ldots$, and

$$r^{(1)} \subset r^{(2)} \subset \ldots \subset r^{(a)} \subset \ldots \tag{4.1}$$

But in order to not burden the notation, it is sufficiently clear to let the notation r refer to any one of the increasingly larger response sets. Data collection stops before the expanding r has reached the full probability sample s. The value y_k recorded for $k \in r$ provide, together with auxiliary variable values, the material for estimating the parameter $Y = \sum_U y_k$.

The (design-weighted) survey response rate is

$$P = \sum_r d_k / \sum_s d_k. \tag{4.2}$$

In the context of the LCS, all d_k are equal because of simple random sampling from the Swedish TPR, so P is simply number of individuals responding divided by number in sample, but for more generality, the formulas that follow are expressed in arbitrary design weights d_k. The response probability of unit k, denoted $\theta_k = \Pr(k \in r | k \in s)$, is a conceptually defined, nonrandom, unknown number. The response rate P is an estimate of the (unknown) mean response probability in the population, $\bar{\theta}_U = \sum_U \theta_k / N$.

Auxiliary information is crucial. We denote as $\mathbf{x}_k$ the auxiliary vector value for unit k, assumed available at least for all units $k \in s$, possibly for all $k \in U$. If $J \geq 1$ auxiliary variables are used, then $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$, where x_{jk} is the value for unit k of the jth auxiliary variable, x_j. We consider auxiliary vectors $\mathbf{x}_k$ of a form such that for some constant vector $\boldsymbol{\mu}$ we have $\boldsymbol{\mu}' \mathbf{x}_k = 1$ for all k. This is not a major restriction. Vectors of importance in practice are usually of this kind, such as when $\mathbf{x}_k = (1, x_k)'$ and $\boldsymbol{\mu} = (1, 0)'$.

The LCS 2009 data file analyzed here contains the observed values y_k of a number of study variables ("the y-variables") and the values x_{jk} of a number of auxiliary variables

("the x-variables") which may be either continuous or categorical equal to 1 or 0 to code the presence or the absence of a given trait of unit k. Some of these auxiliary values are obtained from the Swedish Total Population Register, while others are derived by matching from other reliable Swedish administrative registers, using the personal identification number.

We compute estimation weights calibrated on auxiliary information about $\mathbf{x}_k$ for $k \in s$. The weight given to the value y_k observed for $k \in r$ is $d_k m_k$, the product of the sampling weight $d_k = 1/\pi_k$ and the adjustment factor

$$m_k = \left(\sum_s d_k \mathbf{x}_k\right)' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k.$$

Hence the resulting calibration estimator is

$$\hat{Y}_{CAL} = \sum_r d_k m_k y_k. \tag{4.3}$$

The weights $d_k m_k$ are constructed to deliver unbiased estimates for the variables in the auxiliary vector, as expressed by the calibration equation $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$. For any $\mathbf{x}$-vector and any response set r, the mean adjustment factor is $\sum_r d_k m_k / \sum_r d_k = 1/P$. Consequently, when the data collection progresses and r gets increasingly larger, the increasing proportion P of observed sample units in the Estimator (4.3) is correspondingly matched by a decreasing average mean adjustment factor $1/P$. But it is the composition of the response set r, the particular units that are in r at any given point, that determines the more or less pronounced bias of $\hat{Y}_{CAL}$. We want "the right kind of units" to be in r in the end.

For the theory behind calibration for nonresponse, see, for example Särndal and Lundström (2005). Calibration will generally reduce the nonresponse bias, and quite considerably if the auxiliary vector is powerful, but without eliminating it entirely. At Statistics Sweden, many potential auxiliary variables are typically available for the estimation. The question then arises about the best choice among those. Indicators for this purpose are given in Särndal and Lundström (2008, 2010).

Remark: The calibrated weights in (4.3) use an auxiliary vector $\mathbf{x}_k$ known for the sample units. In practice at Statistics Sweden the calibration estimates ordinarily draw on auxiliary information at two levels: at the population level, transmitted by a vector $\mathbf{x}_k^*$, and at the sample level, transmitted by a vector $\mathbf{x}_k^\circ$. The population total $\sum_U \mathbf{x}_k^*$ is known, while $\sum_U \mathbf{x}_k^\circ$ is unknown but estimated without bias by $\sum_s d_k \mathbf{x}_k^\circ$, which helps the reduction of nonresponse bias. The auxiliary vector is $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^\circ \end{pmatrix}$, and to benefit from the potential for reduced variance when $\sum_U \mathbf{x}_k^*$ is known, the weights are calibrated to satisfy $\sum_r w_k \mathbf{x}_k = \mathbf{X} = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^\circ \end{pmatrix}$. The published survey estimate is $\hat{Y}_{CAL}^* = \sum_r w_k y_k$ with weights $w_k = d_k \left\{ \mathbf{X}' \left(\sum_r d_k \mathbf{x}_k \mathbf{x}'_k \right)^{-1} \mathbf{x}_k \right\}$. But for the purposes of this article, it is deemed appropriate to use $\hat{Y}_{CAL} = \sum_r d_k m_k y_k$ in (4.3) with weights calibrated as $\sum_r d_k m_k \mathbf{x}_k = \sum_s d_k \mathbf{x}_k$.

In our analysis of the LCS, the adjustment factors m_k in (4.3) are computed on an auxiliary vector $\mathbf{x}_k$ of dimension eight considered suitable for monitoring the estimates over the course of the data collection and composed of the following categorical auxiliary variables: *Phone access* (equaling 1 for a person with accessible phone number; 0 otherwise), *Education level* (equaling 1 if high; 0 otherwise), *Age group* (four zero/one coded groups; age brackets -24, $25-64$, $65-74$, $75+$ years); *Property ownership* (equaling 1 for a property owner; 0 otherwise); *Country of origin* (equaling 1 if born in Sweden; 0 otherwise). We refer to this vector as the *standard* x-*vector* (to distinguish it from the *experimental* x-*vector* needed in Section 6). These variables are a subset of those used to produce the published calibration estimates in the LCS 2009.

The variable *Property ownership* equals one for a person identified in the property tax register as having paid taxes on real estate property owned. The variable *Phone* equals one for a person whose phone number is available and ready to be used at the very beginning of the data collection period. All persons in Sweden have access to a phone, whether a landline or cell phone. When the sample of persons has been drawn from the Swedish Total Population Register, it is matched to the phone register and, if found, the number is noted. "Found" or "not found" defines the dichotomous variable *Phone*. For different reasons, not all phone numbers are in the phone register. About 90% of the needed phone numbers are found. Before the start and during the field work, the interviewers try to trace the persons with phone numbers as yet missing using various sources, for example the internet. In this manner, telephone numbers are found and can be used for about one third of those with initial value zero on the *Phone* variable.

In Tables 1 and 2 (as in later tables), the entries for "Attempt number" a (where $a = 1, 2, 3 \ldots$) are computed on the union of the sets of persons having responded at attempts 1, 2, $\ldots$, a, as expressed by (4.1). Not all call attempts are shown in the tables, but changes for deleted rows are minor. The entries for "End ordinary field work" are computed on the respondents at the end of the five week ordinary data collection period; "Final" is based on the total response recorded at the end of the follow-up period.

The three register variables used here as study variables are: *Sickness insurance benefits* (for simplicity called *Benefits*, a categorical variable equaling 1 for a recipient of such benefits; 0 otherwise), *Income* (a continuous variable including employment as well as retirement income), and *Employed* (a categorical variable equaling 1 for an employed person; 0 otherwise). We chose these three register variables because they are central aspects of living conditions as studied in the LCS. The use of these register variables as study variables meets a methodological objective: We can follow the progression of estimators of interest, and study the benefits of calibrated weighting.

For the three register variables, y_k is available for $k \in s$, and we can for comparison compute the unbiased full sample (Horvitz-Thompson) estimate

$$\hat{Y}_{FUL} = \sum_s d_k y_k. \tag{4.4}$$

The computable percentage relative difference between $\hat{Y}_{FUL}$ (unbiased) and $\hat{Y}_{CAL}$ (biased to some extent) is

$$RDF_{CAL} = 100 \cdot (\hat{Y}_{CAL} - \hat{Y}_{FUL})/\hat{Y}_{FUL}. \tag{4.5}$$

The calibration estimator generated by the primitive auxiliary vector, $\mathbf{x}_k = 1$ for all units k serves as a benchmark; it is the expansion estimator given by

$$\hat{Y}_{EXP} = \Big(\sum_s d_k\Big)\Big(\sum_r d_k y_k\Big)/\Big(\sum_r d_k\Big). \tag{4.6}$$

Its often large relative deviation from the unbiased $\hat{Y}_{FUL}$ is

$$RDF_{EXP} = 100 \cdot (\hat{Y}_{EXP} - \hat{Y}_{FUL})/\hat{Y}_{FUL}. \tag{4.7}$$

Table 1 shows RDF_{CAL} (computed with the standard **x**-vector) and RDF_{EXP} for the three variables and for a number of steps in the LCS 2009 data collection. We note that:

- The numerically important changes in RDF_{CAL} and RDF_{EXP} occur early in the series of attempts because of important data inflows. From around attempt five onwards, both follow quite a stable pattern; later changes are small, moving in smooth continuous fashion. The changes are necessarily minute when the data collection has gone on for some time, because small amounts of new data are added to a substantial

Table 1. The LCS 2009 data collection: Progression of the response rate P (in per cent) and of RDF for three selected register variables. The calibration estimator is based on the standard x-vector explained in this section

Attempt number	$100 \times P$	Benefits		Income		Employed	
		RDF_{CAL}	RDF_{EXP}	RDF_{CAL}	RDF_{EXP}	RDF_{CAL}	RDF_{EXP}
1	12.8	10.5	−10.0	−0.05	0.3	−1.3	−9.0
2	24.6	3.3	−13.9	−1.1	0.4	−2.0	−8.1
3	32.8	1.6	−12.1	−0.4	1.6	0.2	−4.7
4	39.6	2.7	−10.1	0.2	2.9	0.4	−2.4
5	44.3	3.7	−7.2	0.7	3.6	1.1	−1.1
6	47.8	2.7	−7.0	1.2	4.5	1.7	0.4
7	50.9	1.6	−7.3	2.1	5.5	2.5	1.6
8	53.0	1.0	−7.4	2.4	6.2	2.4	2.3
9	54.6	0.2	−8.0	2.8	6.4	2.6	2.5
10	55.7	0.2	−8.0	2.8	6.6	2.6	2.8
11	56.8	−0.5	−8.5	2.7	6.5	2.6	3.0
12	57.7	0.1	−7.9	3.0	6.8	2.5	3.1
13	58.3	−0.3	−8.0	3.0	6.9	2.7	3.4
14	58.7	−0.1	−7.7	3.0	6.9	2.7	3.6
15	59.1	−0.5	−8.0	3.1	7.1	2.8	3.8
⋮							
20	60.1	−0.5	−7.7	3.4	7.5	3.0	4.1
End ord. fieldwork	60.4	−0.9	−7.9	3.3	7.4	2.9	4.2
				Follow-up			
1	61.4	−1.0	−8.0	3.3	7.1	2.9	4.1
2	62.6	−1.6	−8.2	3.1	6.7	3.0	3.9
3	63.8	−2.5	−9.2	3.0	6.7	3.2	4.2
4	64.6	−2.8	−9.3	3.1	6.7	3.3	4.3
5	65.3	−2.7	−9.0	3.1	6.8	3.1	4.3
⋮							
10	66.8	−2.9	−8.9	2.9	6.7	3.0	4.5
Final	67.4	−3.6	−9.4	2.9	6.7	3.1	4.8

body of existing data. Every new contact attempt brings progressively smaller amounts of new information. For this we use the word "stabilization", as evidenced in Table 1 for *RDF* and in later tables for other statistics.

- For all three study variables, RDF_{CAL} is small, in fact near zero, early in the data collection. For example, for the *Benefits* variable, RDF_{CAL} hovers around zero in the range from 9 to 14 call attempts. The other two variables have near-zero RDF_{CAL} even earlier in the data collection. Nevertheless, at the very end (the row "Final"), the value of RDF_{CAL} is large, -3.6%, 2.9%, and 3.1% respectively. The LCS with its unchanging data collection plan does not result in small estimation error.
- The large departures from the unbiased estimate $\hat{Y}_{FUL}$, signaled by high values of RDF_{EXP}, indicate that the LCS 2009 data collection ends up with a markedly skewed response. In most of the steps in Table 1, RDF_{EXP} is greater than RDF_{CAL}. Hence the auxiliary information that $\hat{Y}_{CAL} = \sum_r d_k m_k y_k$ can draw on is valuable for reducing the departure from the unbiased estimate for all three variables, although the end result is short of satisfactory. The difference between RDF_{EXP} and RDF_{CAL} is less pronounced for *Employed*; for this variable, the auxiliary vector is less effective.

5. Further Tools: Indicators of Balance, Distance and Representativity

In the theoretical first part of this section, we explain several indicators designed for monitoring the data collection. The indicators reflect well known statistical concepts. Later in the section we illustrate the indicators numerically by computing them on the LCS 2009 data. Furthermore, in Section 6, those indicators will be used in an experiment with the LCS 2009 data, whereby we intervene "in retrospect" in the data collection process, aiming to achieve a better balanced, or more representative, ultimate response set than if no action were taken. The indicators can be computed from the auxiliary variable values, known for both respondents and nonrespondents.

We distinguish three types of concept from which to construct an indicator: (i) Balance (of the response set, for selected auxiliary variables), (ii) Distance (between respondents and nonrespondents), and (iii) Variability of (estimated) response probabilities. All depend on the idea of imbalance now to be defined. Desirable features are high balance, low distance and small variability of the response probabilities. We use the indicators to observe the dynamic pattern as the data collection unfolds and to allow interventions to be made at suitable points.

All the indicators rely on an auxiliary vector, denoted in general by $\mathbf{x}$ with value $\mathbf{x}_k = (x_{1k}, \ldots, x_{jk}, \ldots, x_{Jk})'$ known for the units $k \in s$ (or possibly for $k \in U$). The dimension J is arbitrary. For the j:th auxiliary variable x_j, with value x_{jk} for unit k, we compute the difference $D_j = \bar{x}_{jr} - \bar{x}_{js}$ between the respondent mean, $\bar{x}_{jr} = \sum_r d_k x_{jk} / \sum_r d_k$, and the full sample mean, $\bar{x}_{js} = \sum_s d_k x_{jk} / \sum_s d_k$. If $D_j = 0$ for all J auxiliary variables, then r is a *perfectly balanced* response set. In vector form, $\mathbf{D} = (D_1, \ldots, D_j, \ldots, D_J)' = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s$ with vector mean $\bar{\mathbf{x}}_r = \sum_r d_k \mathbf{x}_k / \sum_r d_k$ for the respondents and $\bar{\mathbf{x}}_s = \sum_s d_k \mathbf{x}_k / \sum_s d_k$ for the full sample. Under perfect balance, $\mathbf{D} = \mathbf{0}$, the zero vector.

We must seek balance on the auxiliary variables, because unlike real study variables, they are individually known for the full sample (or for the whole population). What benefit can we expect from balancing on a chosen vector $\mathbf{x}_k$? Is there reason to expect that balance on the

x-vector will produce, if not perfect, at least good balance for the *y*-variables in the survey, in particular for the highly important *y*-variable? Let us consider these questions. The concept of balance refers to the equality of response set mean and full sample mean. We have **x**-vector balance if $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$. We can strive to come close to this during data collection. The desirable goal of *y*-variable balance is expressed as $\bar{y}_r = \bar{y}_s$. Whether or not this comes close to being satisfied for a real study variable *y* will never be known. But balancing on a chosen **x**-vector can bring us closer. Using the property $\boldsymbol{\mu}'\mathbf{x}_k = 1$ for all *k*, we express the difference $\bar{y}_r - \bar{y}_s$ (which we would like to be zero or close to zero) as a sum of two terms,

$$\bar{y}_r - \bar{y}_s = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r + (\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s \tag{5.1}$$

where $\mathbf{b}_r = \left(\sum_r d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_r d_k \mathbf{x}_k y_k$ and $\mathbf{b}_s = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \sum_s d_k \mathbf{x}_k y_k$ are linear regression coefficients (regressing *y* on **x**), for the response set and for the full sample respectively. That $\mathbf{b}_r$ and $\mathbf{b}_s$ may differ is an expression of a dilemma well known in regression analysis: Non-random selection of cases causes biased regression. The computable first term of (5.1), $(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\mathbf{b}_r$, is zero if the **x**-balance $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_s$ is realized. This by itself does not imply that the second (not computable) term $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ is zero or small. But often that term, which is what remains of the difference $\bar{y}_r - \bar{y}_s$ after complete balance on the **x**-vector, is smaller than what that difference would be in the absence of any balancing. One situation where the second term is small is when *y* is well explained by the **x**-vector, so that $y_k \approx \boldsymbol{\beta}'\mathbf{x}_k$ and therefore $\mathbf{b}_r \approx \mathbf{b}_s$. In other words, if the response is balanced for a vector $\mathbf{x}_k$ highly related to the study variable *y*, then we are close to *y*-variable balance. Another condition under which the second term $(\mathbf{b}_r - \mathbf{b}_s)'\bar{\mathbf{x}}_s$ is small occurs if the data collection can be directed to yield a response set *r* that is an essentially random subset of *s*. In many situations there is a strong incentive to seek balance on a suitable **x**-vector, because it will likely bring us closer to *y*-variable balance. Multiplying by $\hat{N} = \sum_s d_k$ shows Equation (5.1) in a different light:

$$\hat{Y}_{EXP} - \hat{Y}_{FUL} = (\hat{Y}_{EXP} - \hat{Y}_{CAL}) + (\hat{Y}_{CAL} - \hat{Y}_{FUL}).$$

Here the computable difference $\hat{Y}_{EXP} - \hat{Y}_{CAL}$ is the adjustment we apply to the primitive estimate $\hat{Y}_{EXP}$ to arrive at the improved estimate $\hat{Y}_{CAL}$. The term $\hat{Y}_{CAL} - \hat{Y}_{FUL}$, unknown for a real *y*-variable, is not zero, but may be small compared with the adjustment $\hat{Y}_{EXP} - \hat{Y}_{CAL}$. There is no choice of **x**-vector that will completely eliminate the nonresponse error $\hat{Y}_{CAL} - \hat{Y}_{FUL}$. Some bias always remains after calibration. Expressed differently, there exists no **x**-vector that realizes missing at random, given **x**.

Normally in practice, $\mathbf{D} \neq \mathbf{0}$, suggesting departure from balance. We transform the multivariate **D** into a suitable univariate measure of *imbalance*, for the given survey outcome (*s*, *r*) and the given composition of $\mathbf{x}_k$. The imbalance is a quadratic form in **D** defined as

$$\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s)'\boldsymbol{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s) \tag{5.2}$$

with weighting matrix $\boldsymbol{\Sigma}_s = \sum_s d_k \mathbf{x}_k \mathbf{x}_k' / \sum_s d_k$. Increased mean differences D_j tend to increase the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$. Interposing the inverse of the weighting matrix permits an upper bound to be stated on the imbalance: For any outcome (*s*, *r*) and any composition of $\mathbf{x}_k$ we have $0 \leq \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} \leq Q - 1$ with $Q = 1/P$ (see Särndal 2011a). For most data

encountered in practice, $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ is not a large number, often 0.3 or less. As the data collection unfolds and the response rate P gets larger, one often finds that $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ decreases, because $\bar{\mathbf{x}}_r$ moves closer to $\bar{\mathbf{x}}_s$ when the response r grows toward the full sample s, although the question depends also on what particular units happen to be in the set r at a given moment.

Balance is imbalance with a reversed sign. We use two indicators of balance, measured on the unit interval scale and such that the value "1" implies perfect balance. The first is

$$BI_1 = 1 - \sqrt{\frac{\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}}{Q-1}}. \tag{5.3}$$

It follows from $0 \leq \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} \leq Q-1$ that $0 \leq BI_1 \leq 1$. Because $P(1-P) \leq 1/4$, an alternative indicator also contained in the unit interval is

$$BI_2 = 1 - 2P\sqrt{\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}}. \tag{5.4}$$

For most data, $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ does not come near $Q-1$. It is not a sharp upper bound. Consequently, both indicators transmit an inflated notion of balance, often greater than 0.8 for both BI_1 and BI_2. The lower portion of the unit interval is not effectively used. The notion of distance now to be discussed is less subject to this criticism.

Our concept of distance, which contrasts respondents with nonrespondents for the chosen $\mathbf{x}$-vector, is a transformation of mean difference vector, $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr}$, where $nr = s - r$ is the nonresponse set with mean $\bar{\mathbf{x}}_{nr} = \sum_{s-r} d_k\mathbf{x}_k / \sum_{s-r} d_k$. This distance is

$$dist_{r|nr} = [(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})'\mathbf{\Sigma}_s^{-1}(\bar{\mathbf{x}}_r - \bar{\boldsymbol{x}}_{nr})]^{1/2}. \tag{5.5}$$

If respondents and nonrespondents agree on average for every variable in the $\mathbf{x}$-vector, then $\bar{\mathbf{x}}_r = \bar{\mathbf{x}}_{nr}$ and $dist_{r|nr} = 0$. From (5.2) and the equation $\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s = (1-P)(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})$ follows that $dist_{r|nr}$ is a simple transformation of the imbalance $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$:

$$dist_{r|nr} = \frac{1}{1-P}(\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D})^{1/2} \tag{5.6}$$

From $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} \leq Q-1$ follows that $dist_{r|nr} \leq 1/\sqrt{P(1-P)}$. Thus for nonresponse in the range 20% to 50%, $dist_{r|nr}$ can never exceed 2.5. But for data encountered in practice $dist_{r|nr}$ is normally much lower, rarely exceeding 0.6. One reason is that the upper bound covers any vector composition $\mathbf{x}_k$ and even the most extreme response outcome r that can occur for the given sample s. The measure $dist_{r|nr}$ reacts distinctly but smoothly to the steps in the data collection and is a more expressive indicator than BI_1 or BI_2, which tend to concentrate in the upper quarter of the unit interval. For example, in Table 3 (Subsection 6.2), $dist_{r|nr}$ roughly doubles from 0.23 at the beginning to 0.47 at the end of the data collection, while BI_1 only moves from 0.85 to 0.72.

Simple relationships between $dist_{r|nr}$ and the balance indicators follow from (5.3), (5.4) and (5.6):

$$BI_1 = 1 - \sqrt{P(1-P)} \times dist_{r|nr} \quad ; \quad BI_2 = 1 - 2P(1-P) \times dist_{r|nr}. \tag{5.7}$$

The principal tools for the empirical work reported later are the balance indicator BI_1 and the distance $dist_{r|nr}$. It is important to follow their progression as the response set r expands

and P increases. Increasing balance and decreasing distance are signs of a satisfactory data collection. But the undesirable opposite can happen, as in an empirical illustration that follows. The level of the indicators depend on the choice of $\mathbf{x}_k$, notably on the number of x-variables in $\mathbf{x}_k$ – it is harder to obtain balance on more variables – and on their relationship with response. The ultimate response set r should have high balance and low distance; however, it is hard to formulate definite ultimate target values for BI_1 and $dist_{r|nr}$, because of their strong dependence on $\mathbf{x}_k$.

If $dist_{r|nr}$ decreases when P increases, then the balance, as measured by BI_1 or by BI_2, may or may not improve, that is, get larger. If the distance $dist_{r|nr}$ increases when P increases towards 50%, then the balance, measured by BI_1 or by BI_2, will necessarily deteriorate.

The third concept behind an indicator for the data collection is the variability of (estimated) response probabilities. It was used in the RISQ project mentioned earlier. The resulting indicators are called R-indicators (with R for representativity); see, for example Schouten et al. (2009). Let $\hat{\theta}_k$ be the estimated response probability for unit $k \in s$. Their variance is

$$S^2_{\hat{\theta}s} = \sum_s d_k(\hat{\theta}_k - \bar{\hat{\theta}}_s)^2 / \sum_s d_k \tag{5.8}$$

where $\bar{\hat{\theta}}_s = \sum_s d_k \hat{\theta}_k / \sum_s d_k$. The R-indicator is defined as

$$R = 1 - 2S_{\hat{\theta}s} \tag{5.9}$$

The rationale behind the construction (5.9) is that if the data collection can be directed to reduce variability in the estimated response probabilities, then the representativity of the response set, measured by R, is said to be improved.

For the chosen specification $\mathbf{x}_k$, the estimates $\hat{\theta}_k = f(\mathbf{x}_k'\hat{\boldsymbol{\beta}})$ can be obtained via different link functions. For the linear response function, using weighted least squares, we determine $\boldsymbol{\beta}$ to minimize $\sum_s d_k(I_k - \mathbf{x}_k'\boldsymbol{\beta})^2$, where I_k is the response indicator, $I_k = 1$ for $k \in r$ and $I_k = 0$ for $k \in s - r$. As a result, $\hat{\boldsymbol{\beta}} = \left(\sum_s d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1}\left(\sum_r d_k \mathbf{x}_k\right)$, and for $k \in s$, the response probability estimate is $\hat{\theta}_k = t_k$ with $t_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}}$. If we denote by S^2_{ts} the variance (5.8) computed with $\hat{\theta}_k = t_k$, the R-indicator (5.9) is $R = 1 - 2S_{ts}$. Because $S^2_{ts} = P^2 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$, the R-indicator for the linear response function is equal to the balance measure (5.4): $1 - 2S_{ts} = BI_2$.

Schouten et al. (2009) closely examine the case where $\hat{\theta}_k$ is obtained through a logistic response function. By logistic regression fit, we obtain first $\hat{\boldsymbol{\beta}}$, then $\hat{\theta}_{k,\log} = \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})/[1 + \exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})]$ for $k \in s$. Their variance, denoted $S^2_{\hat{\theta}\log,s}$, is computed in the manner of (5.8) with $\hat{\theta}_k = \hat{\theta}_{k,\log}$, and the resulting logistic R-indicator is

$$R = 1 - 2S_{\hat{\theta}\log,s}. \tag{5.10}$$

Bethlehem et al. (2011) consider a bias-adjusted R-indicator, reflecting a desire to reduce the bias that (5.10) may have when viewed as an estimate of a corresponding population quantity. With our data, (5.10) differed negligibly from its bias-corrected counterpart, not shown in Table 2.

A noteworthy property of the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is its simple relation to the coefficient of variation of the estimates $\hat{\theta}_k = t_k$ for $k \in s$. Because $\bar{t}_s = \sum_s d_k t_k / \sum_s d_k = P$ we have

$$cv_{ts} = S_{ts}/\bar{t}_s = (\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2}. \tag{5.11}$$

Often close in value to cv_{ts} is the coefficient of variation of the adjustment factors m_k in the estimator $\hat{Y} = \sum_r d_k m_k y_k$. It can be shown that

$$cv_{mr} = S_{mr}/\bar{m}_r = (\mathbf{D}'\boldsymbol{\Sigma}_r^{-1}\mathbf{D})^{1/2}$$

where $\boldsymbol{\Sigma}_r = \sum_r d_k \mathbf{x}_k \mathbf{x}_k' / \sum_r d_k$; $S_{mr} = \left[\sum_r d_k (m_k - \bar{m}_r)^2 / \sum_r d_k\right]^{1/2}$ and $\bar{m}_r = \sum_r d_k m_k / \sum_r d_k = 1/P$. The two coefficients of variation differ only in the inverted weighting matrix: $\boldsymbol{\Sigma}_s^{-1}$ in the former, Σ_r^{-1} in the latter. The statistic cv_{mr} is used in

Table 2. The LCS 2009 data collection: Progression of the response rate P (in per cent), balance BI_1 and BI_2, logistic R–indicator (5.10), distance $dist_{r|nr}$, and square root of imbalance $(\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2} = cv_{ts}$. Computations based on the standard x-vector explained in Section 4

Attempt number	$100 \times P$	Balance		*R*-indicator *formula (5.10)*	Distance $dist_{r\|nr}$	Sqrt. imbalance cv_{ts}
		BI_1	BI_2			
1	12.8	0.855	0.904	0.902	0.433	0.378
2	24.6	0.802	0.829	0.829	0.460	0.347
3	32.8	0.779	0.793	0.794	0.470	0.316
4	39.6	0.770	0.775	0.780	0.471	0.285
5	44.3	0.767	0.769	0.775	0.469	0.261
6	47.8	0.763	0.763	0.770	0.475	0.248
7	50.9	0.756	0.756	0.763	0.488	0.240
8	53.0	0.751	0.752	0.758	0.499	0.234
9	54.6	0.750	0.752	0.757	0.501	0.227
10	55.7	0.748	0.749	0.756	0.508	0.225
11	56.8	0.746	0.749	0.754	0.512	0.221
12	57.7	0.747	0.750	0.756	0.513	0.217
13	58.3	0.744	0.748	0.754	0.519	0.217
14	58.7	0.742	0.746	0.753	0.523	0.216
15	59.1	0.741	0.745	0.752	0.527	0.215
⋮						
20	60.1	0.737	0.743	0.751	0.536	0.214
End ordinary	60.4	0.738	0.744	0.752	0.536	0.212
			Follow-up			
1	61.4	0.736	0.743	0.751	0.542	0.210
2	62.6	0.734	0.742	0.750	0.550	0.206
3	63.8	0.730	0.741	0.748	0.561	0.203
4	64.6	0.728	0.740	0.747	0.569	0.201
5	65.3	0.727	0.740	0.748	0.573	0.199
⋮						
10	66.8	0.719	0.736	0.742	0.596	0.198
Final	67.4	0.717	0.735	0.742	0.603	0.197

selecting auxiliary variables at the estimation stage, as in Särndal and Lundström (2010) and Särndal (2011b).

Table 2 shows the progression during the LCS 2009 data collection of BI_1, BI_2 and the logistic R-indicator (5.10), all viewed as functions of the call attempt number. The three measures are numerically close, but all are subject to the criticism that they fail – here and in other applications – to effectively use the whole unit interval. Often around 0.8 or higher, they seldom fall below 0.7. Here, as for most other data, the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ does not come near its upper bound $Q - 1$, so (5.3) and the related (5.4) fall predominantly into the upper end of the unit interval. Because BI_1, BI_2 and the logistic R-indicator (5.10) tell essentially the same story, we focus in the following on BI_1. Table 2 also shows the progression of the distance $dist_{r|nr}$ between respondents and nonrespondents.

The desired pattern of reduced distance and increased balance does not happen for the LCS 2009 data collection. Instead Table 2 shows that the balance indicators and the distance $dist_{r|nr}$ go the wrong way: The balance decreases; the distance $dist_{r|nr}$ gets larger. Thus Table 2 reinforces the message already conveyed in Table 1 of a weakness in the LCS data collection. It raises the question of whether the ordinary field work should proceed as long as it currently does, instead of ending after say ten attempts. The follow-up does not bring improvement; the indicators continue in the wrong direction.

Also shown in Table 2 is $(\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D})^{1/2} = cv_{ts}$ defined in (5.11). Here the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ goes from an initial value 0.14 to a final value 0.04, a decrease explained in large part by the increasing proportion P, which makes $\bar{\mathbf{x}}_r$ move closer to $\bar{\mathbf{x}}_s$ and $\mathbf{D}$ closer to the zero vector.

6. Experimental Data Collection Strategies

6.1. Auxiliary Vector for the Experiments

There is strong evidence that realizing a predefined "respectable" overall response rate should no longer be accorded the same dominant importance in future renditions of the LCS. It is hard to justify a costly effort for a possible five per cent greater ultimate response rate unless accompanied by concrete measures of quality in the response set, such as progressively better balance and closeness of respondents to nonrespondents. As Tables 1 and 2 have shown, those features are lacking in the LCS 2009 data collection.

This section presents the results of three "experiments in retrospect" carried out with the existing LCS 2009 data file. We cannot add more data, but we can delete data from that file to show the effects of different interventions in the data collection, in particular the trend in the balance indicators BI_1 and BI_2, and in the distance $dist_{r|nr}$. Increasing balance and decreasing distance are features we hope to find. Our experiments consist in treating data collection as terminated, at suitably chosen points in the data inflow, for sample groups with relatively high response. For example, it might stop in some groups after a suitable number of call attempts because realistic expectations for the response have already been met, whereas for the other groups, data collection would continue for some time yet before stopping, and for remaining low-responding groups it would continue until the very end of the data collection period. We refer to the points where stopping occurs as *intervention points*.

In this manner we delete data in the existing LCS 2009 data file pretending that data collection has been terminated at given points for relatively high-responding sample subgroups. In other words, for those groups, we sacrifice some data y_k that were in reality available beyond the intervention points. The imbalance measured by $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ plays a key role in the analysis. It determines the balance measures BI_1 and BI_2 given respectively by (5.3) and (5.4) and the distance $dist_{r|nr}$ given by (5.6). An aim for the data collection should be to reduce the differences D_j that make up the vector $\mathbf{D} = \bar{\mathbf{x}}_r - \bar{\mathbf{x}}_s = (D_1, \ldots, D_j, \ldots, D_J)'$.

The imbalance $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ has a particularly transparent expression when the vector $\mathbf{x}_k$ is defined by J mutually exclusive and exhaustive traits or characteristics, for example when "Age" is defined by, say, $J = 3$ traits, Young, Middle-aged and Elderly. But usually in practice, several categorical variables are crossed to define a set of mutually exclusive and exhaustive groups. The trait of unit k is then uniquely coded by the J-vector $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{jk}, \ldots, \gamma_{Jk})' = (0, \ldots, 1, \ldots, 0)'$ (with a single entry "1"), or equivalently by the J-vector $\mathbf{x}_k = (1, \gamma_{1k}, \ldots, \gamma_{J-1,k})'$, where $\gamma_{jk} = 1$ if k has the trait j and $\gamma_{jk} = 0$ otherwise. Denote by s_j the (non-empty) subset of the sample s consisting of the units k with the trait j, and let r_j be the corresponding responding subset of the whole response set r; $r_j \subseteq s_j$. For trait j, denote by $W_{js} = \sum_{s_j} d_k / \sum_s d_k$ that trait's share of the whole sample s. Then the imbalance is a sum of non-negative terms expressed as

$$\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = \sum_{j=1}^{J} C_j \tag{6.1}$$

with

$$C_j = W_{js} \times \left(\frac{P_j}{P} - 1\right)^2 \tag{6.2}$$

where $P_j = \sum_{r_j} d_k / \sum_{s_j} d_k$ is the response rate for the jth group and P is the overall response rate given by (4.2). We call $(P_j - P)/P$ the *response rate differential* for the jth group. Together, the J differentials $(P_j - P)/P$ describe the state of the response for the set of groups at any given point in the data collection. The differentials are positive, negative or zero. If all are zero, the imbalance is zero, and the balance is perfect for the chosen $\mathbf{x}$-vector: $BI_1 = BI_2 = 1$. The differentials change continuously during data collection and can be substantially different, although experience shows that they are seldom greater in absolute value than 0.3. At any given point in the data collection, their weighted average is zero: $\sum_{j=1}^{J} W_{js} \times ((P_j/P) - 1) = 0$. The imbalance (6.1) is therefore the variance over the J groups of the differentials $(P_j - P)/P$. If the maximum $|P_j - P|/P$ over the J groups equals, say, 0.5, it follows that $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} \leq 0.25$ and that $(\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D})^{1/2} = cv_{ts} \leq 0.5$. If all J response rates P_j are equal then $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D} = 0$.

The analysis in Sections 4 and 5 was based on the *standard* $\mathbf{x}$-*vector*, close to the auxiliary vector used to produce the calibration estimates for LCS 2009. Here we choose a more appropriate vector that identifies a set of particularly important sample subgroups. Using this *experimental* $\mathbf{x}$-*vector* we carry out three "experiments in retrospect" on the LCS 2009 data, each based on an *experimental data collection strategy* defined by one or more intervention points and a stopping rule for each intervention point. This vector points

out membership in one of $J = 8$ mutually exclusive and exhaustive sample groups. Every intervention point marks a change in the data collection. The stopping rule is formulated in terms of a predefined target response rate for each group, so that data collection will be deemed terminated at a given intervention point for groups having at that point reached the specified response rate. This is a simple form of responsive design, made possible by the categorical nature of the experimental vector.

The experimental **x**-vector is defined by the crossing of three dichotomous auxiliary variables: *Education level* (high, not high), *Property ownership* (owner, non-owner), *Country of origin* (Sweden, other). There are $J = 2^3 = 8$ mutually exclusive and exhaustive groups coded by the experimental x-vector $\mathbf{x}_k = (\gamma_{1k}, \ldots, \gamma_{8k})'$, where $\gamma_{jk} = 1$ if k belongs to group j and $\gamma_{jk} = 0$ otherwise. Group membership, and hence the value $\mathbf{x}_k$, is known auxiliary information for all $k \in s$. Those eight groups are important to monitor because experience has shown their response rates to be considerably different and indeed strikingly low for some, as Table 5 confirms.

6.2. *The Actual LCS 2009 Data Collection Analyzed with the Experimental x-vector*

By the *actual LCS data collection* we mean the data collection as actually carried out, with all the contact attempts and realized responses, resulting in the actual LCS 2009 response set. We compare it in Subsection 6.3 and 6.4 with three experimental data collections where the LCS 2009 data set is censored by stopping rules for data collection in certain sample subgroups.

To put the experiments in their proper light, we analyze first the actual LCS 2009 data collection in the light of the experimental **x**-vector defined in Subsection 6.1. Summary results are shown in Tables 3 and 4.

As Table 1 showed for the standard **x**-vector, Table 3 shows that RDF_{CAL} (with weights now calibrated on the experimental **x**-vector) does not terminate at desirable near-zero levels. The balance, measured by BI_1 and BI_2, decreases as the data collection proceeds, and the distance $dist_{r|nr}$ increases. This again indicates an inefficiency in the 2009 data collection, with its predefined unchangeable format.

Table 4 shows the progression over the LCS 2009 data collection of the eight terms C_j defined by (6.2), whose total $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is given in the bottom line. Both are multiplied by 100. A low variability in the C_j is a goal, because if all C_j are equal in the end, the imbalance is zero. The groups in lines 1 and 8 stand out, but for different reasons. In both cases, C_j remains high from attempt 5 (where the data collection has gained a certain stability) until the very end. High values of C_j also prevail throughout for lines 4, 5 and 6. The very low-responding line 1 group, *education not high, non-owner, foreign origin*, has a large negative response differential $(P_j - P)/P$, and although $100 \times C_j$ decreases somewhat from 1.44 at attempt 5 to a final value of 1.18, the decrease is much weaker than desired. A negative response differential and a large C_j also characterizes the line 5 group. By contrast, distinctly positive response differentials $(P_j - P)/P$ characterize lines 4, 6 and 8. Most prominent of these is the line 8 group, *high education, property owner, Swedish origin*, for which $100 \times C_j$ decreases somewhat, from 0.58 at attempt 5 to a final value of 0.44, but less than one would like to see.

Table 3. The actual LCS 2009 data collection: Progression of RFD_{CAL} (for three study variables), BI_1 and $dist_{r|nr}$. Computations based on the experimental x-vector of dimension eight defined in Subsection 6.1

		RDF_{CAL}				
Attempt number	$100 \times P$	Benefits	Income	Employed	BI_1	$dist_{r\|nr}$
1	12.8	−8.4	−2.7	−10.2	0.922	0.233
2	24.6	−13.2	−3.2	−9.7	0.887	0.263
3	32.8	−11.5	−2.3	−6.3	0.867	0.283
4	39.6	−8.5	−1.5	−4.4	0.850	0.306
5	44.3	−5.8	−0.5	−3.0	0.846	0.310
⋮						
8	53.0	−5.7	1.7	0.2	0.812	0.377
⋮						
12	57.7	−6.1	2.5	1.2	0.805	0.394
⋮						
20	60.1	−6.0	3.1	2.2	0.795	0.418
⋮						
End ordinary field work	60.4	−6.2	3.1	2.3	0.796	0.417
			Follow-up			
1	61.4	−6.2	3.0	2.3	0.796	0.418
⋮						
4	64.6	−7.9	2.8	2.6	0.792	0.435
⋮						
Final	67.4	−7.9	2.9	3.1	0.779	0.471

In Subsections 6.3 and 6.4 we contrast these results on the group factors C_j with results from three experimental strategies obtained through interventions in the LCS 2009 data base. All three use the experimental **x**-vector of dimension $J = 2^3 = 8$ as defined in Subsection 6.1, but they differ in the points of intervention and in the stopping rules.

Table 4. The actual LCS 2009 data collection: values of the eight terms C_j of $\mathbf{D}^t \boldsymbol{\Sigma}_s^{-1} \mathbf{D}$ (both multiplied by 100). Computations based on the experimental x-vector defined in Subsection 6.1

Group characteristic			$100 \times C_j$						
			Ord. field work attempt				Follow-up attempt		
Education	Property ownership	Origin	1	5	12	End	1	4	Final
Not high	Non-owner	Abroad	1.49	1.44	1.26	1.23	1.25	1.16	1.18
Not high	Non-owner	Sweden	0.00	0.06	0.11	0.11	0.08	0.07	0.07
Not high	Owner	Abroad	0.06	0.01	0.00	0.00	0.00	0.00	0.00
Not high	Owner	Sweden	0.72	0.24	0.21	0.19	0.17	0.17	0.18
High	Non-owner	Abroad	1.28	0.39	0.29	0.26	0.25	0.23	0.22
High	Non-owner	Sweden	0.11	0.26	0.25	0.24	0.21	0.20	0.23
High	Owner	Abroad	0.18	0.01	0.03	0.03	0.03	0.02	0.04
High	Owner	Sweden	0.29	0.58	0.64	0.66	0.62	0.53	0.44
		$100 \times \mathbf{D}^t \boldsymbol{\Sigma}_s^{-1} \mathbf{D}$	4.13	2.99	2.78	2.72	2.61	2.37	2.36

Table 5. Response rate P (in per cent) at three points in the actual LCS 2009 data collection for the eight groups formed by the experimental x-vector

Group characteristic			Response rate P (per cent)			
Education	Property ownership	Origin	Attempt 12 ordinary	Attempt 2 follow-up	Final	Individuals in sample
Not high	Non-owner	Abroad	37.5	41.8	44.6	847
Not high	Non-owner	Sweden	54.6	59.8	64.6	3210
Not high	Owner	Abroad	58.5	62.3	66.8	171
Not high	Owner	Sweden	63.0	67.6	73.2	2036
High	Non-owner	Abroad	39.4	44.9	48.7	236
High	Non-owner	Sweden	66.8	71.6	77.6	816
High	Owner	Abroad	68.1	73.6	81.9	72
High	Owner	Sweden	72.2	77.4	81.5	832
		Total	57.7	62.6	67.4	8220

6.3. *Experimental Strategy 1 and Its Results*

We define Experimental Strategy 1 to have two intervention points, Attempt 12 of the ordinary data collection (point 1), and Attempt 2 of the follow-up (point 2); the stopping rule is to declare data collection terminated (so that no further y-values are taken into account) in a group that has realized at least 65% response. Table 5 shows the response rates for the actual LCS 2009 data at the two intermediate points and at the very end

Table 6. Experimental strategy 1: the eight terms C_j of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (multiplied by 100), the response rate P (in per cent), the balance BI_1, and the distance $dist_{r|nr}$, computed on the experimental x-vector at three points in the data collection

Group characteristic			Value of $100 \times C_j$ at		
Education	Property ownership	Origin	Attempt 12 ordinary	Attempt 2 follow-up	Final
Not high	Non-owner	Abroad	1.26	1.06	0.94
Not high	Non-owner	Sweden	0.11	0.03	0.00
Not high	Owner	Abroad	0.00	0.00	0.00
Not high	Owner	Sweden	0.21	0.24	0.08
High	Non-owner	Abroad	0.29	0.21	0.16
High	Non-owner	Sweden	0.25	0.07	0.02
High	Owner	Abroad	0.03	0.01	0.00
High	Owner	Sweden	0.64	0.31	0.17
		$100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$	2.78	1.93	1.39
		$100 \times P$	57.7	61.5	63.9
		BI_1	0.805	0.824	0.843
		$dist_{r\|nr}$	0.394	0.361	0.326

(Final). It follows that the Strategy 1 data collection is deemed terminated at point 1 for the groups in lines 6, 7 and 8, and at point 2 for the group in line 4, while remaining groups continue until the very end. For the low-responding line 1 and 5 groups, the final response rate is still far from 65%.

For the data collection of Strategy 1, Table 6 shows the progression of the terms C_j and their total $\mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ (both multiplied by 100). Data collection has ended at point 1 for the relatively high-responding groups in lines 6, 7 and 8. The ensuing marked decrease in C_j for lines 6 and 8 occurs because the response differential $(P_j - P)/P$ drops when the increasing P draws nearer the unchanging P_j. The low-responding group 1 accounts for the largest $100 \times C_j$. It drops from 1.26 at point 1 but still ends at a fairly high 0.94. Both P_1 and P increase; they are getting closer, and $|P_1 - P|/P$ is reduced, but not enough. Although only two interventions are used in Strategy 1, the imbalance $100 \times \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ is greatly reduced, from 2.78 at first intervention to 1.39 at the end. As Table 6 also shows, both balance and distance now go in the desired directions: The balance BI_1 increases from 0.805 to 0.843 and the distance $dist_{r|nr}$ decreases from 0.394 to 0.326. The ultimate response rate for Strategy 1 is 63.9%, as compared with 67.4% in the actual LCS 2009 data collection.

6.4. *Experimental Strategies 2 and 3*

Experimental strategies 2 and 3 use sharpened stopping rules for the data collection in the eight groups defined by the experimental **x**-vector defined in Subsection 6.1. The objective is to attempt to confirm the expectation that still better balance can be achieved.

Strategy 2 is defined to declare data collection terminated (in the ordinary data collection or in the follow-up) for a group as soon as its response has reached 60%. The resulting five intervention points are shown in Table 7: Five groups terminate at four different points in the ordinary data collection, and one group terminates at follow-up attempt 3. The low-responding line 1 and line 5 groups continue to the end, but still do not come near 60% response.

Table 8 shows the terms $100 \times C_j$, which sum to the total imbalance $100 \times \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$. Compared with Strategy 1 in Table 6, we see that Strategy 2 brings improvement in that for all but the line 1 group, C_j is reduced to near-zero levels at the end (the column Final). The column total $100 \times \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$ is reduced markedly from 3.07 at first intervention to a final value of 0.82, considerably lower than the final value 1.39 for Strategy 1. As a result, the balance improves markedly, and the distance $dist_{r|nr}$ is reduced in the end to 0.220, as compared with 0.326 for Strategy 1.

The stopping rule for experimental strategy 3 is to declare data collection terminated for a group whose response rate has reached 50%. For this more stringent rule, data collection terminates with still fewer attempts than in Strategy 2. The improvement in the indicators becomes further pronounced, giving better balance and decreased distance compared with Strategies 1 and 2. The distance $dist_{r|nr}$ now ends at 0.089, as compared with final values of 0.220 in Strategy 2 and 0.326 in Strategy 1. Strategy 3 in the end leaves very little variation in the response differentials, which explains the low value 0.20 of $100 \times \mathbf{D}'\mathbf{\Sigma}_s^{-1}\mathbf{D}$.

Table 7. Experimental strategies 2 and 3: termination points for data collection and response rate (in per cent) at termination for the eight groups

Group			Strategy 2		Strategy 3	
Education	Property ownership	Origin	Termination point	Response rate at termination	Termination point	Response rate at termination
Not high	Non-owner	Abroad	Final	44.6	Final	44.6
Not high	Non-owner	Sweden	Att. 3 follow-up	61.0	Att. 8 ordinary	50.0
Not high	Owner	Abroad	Att. 15 ordinary	60.2	Att. 7 ordinary	51.5
Not high	Owner	Sweden	Att. 9 ordinary	60.0	Att. 6 ordinary	52.6
High	Non-owner	Abroad	Final	48.7	Final	48.7
High	Non-owner	Sweden	Att. 7 ordinary	60.2	Att. 5 ordinary	50.5
High	Owner	Abroad	Att. 8 ordinary	62.5	Att. 6 ordinary	52.8
High	Owner	Sweden	Att. 7 ordinary	63.5	Att. 4 ordinary	50.1
		Entire data collection		58.9		50.3

Table 8. Experimental strategy 2: the eight terms C_j of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (multiplied by 100), the balance BI_1, and the distance $dist_{r|nr}$, computed on the experimental x-vector at six points in the data collection. Column "7 ord." refers to "Attempt 7 in the ordinary data collection"; analogous for other columns

Group			Value of $100 \times Cj$ at data collection point					
Education	Property ownership	Origin	7 ord.	8 ord.	9 ord.	15 ord.	3 fol.-up	Final
Not high	Non-owner	Abroad	1.39	1.40	1.29	0.99	0.78	0.60
Not high	Non-owner	Sweden	0.12	0.09	0.05	0.00	0.07	0.05
Not high	Owner	Abroad	0.00	0.00	0.00	0.01	0.00	0.00
Not high	Owner	Sweden	0.25	0.33	0.33	0.13	0.02	0.01
High	Non-owner	Abroad	0.35	0.31	0.31	0.22	0.15	0.09
High	Non-owner	Sweden	0.33	0.21	0.14	0.05	0.01	0.00
High	Owner	Abroad	0.02	0.03	0.02	0.01	0.00	0.00
High	Owner	Sweden	0.61	0.44	0.33	0.18	0.07	0.06
		$100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$	3.07	2.81	2.49	1.59	1.09	0.82
		$100 \times P$	50.9	52.5	53.8	56.0	58.6	58.9
		BI_1	0.822	0.824	0.830	0.858	0.876	0.892
		$dist_{r\|nr}$	0.357	0.353	0.341	0.287	0.252	0.220

Table 9. Experimental strategy 3: the eight terms C_j of $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ (multiplied by 100), the balance BI_1, and the distance $dist_{r|nr}$, computed on the experimental x-vector at six points in the data collection. Column headed "4 ord." refers to "Attempt 4 in the ordinary data collection"; analogous for other columns

Group			Value of $100 \times C_j$ at data collection point					
Education	Property ownership	Origin	4 ord.	5 ord.	6 ord.	7 ord.	8 ord.	Final
Not high	Non-owner	Abroad	1.51	1.39	1.23	1.10	1.05	0.13
Not high	Non-owner	Sweden	0.05	0.03	0.02	0.00	0.03	0.00
Not high	Owner	Abroad	0.01	0.00	0.00	0.01	0.01	0.00
Not high	Owner	Sweden	0.26	0.30	0.46	0.25	0.16	0.05
High	Non-owner	Abroad	0.59	0.38	0.35	0.27	0.22	0.00
High	Non-owner	Sweden	0.27	0.30	0.12	0.06	0.03	0.01
High	Owner	Abroad	0.00	0.01	0.02	0.01	0.01	0.00
High	Owner	Sweden	0.72	0.21	0.07	0.02	0.01	0.00
		$100 \times \mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$	3.42	2.62	2.26	1.73	1.51	0.20
		$100 \times P$	39.6	43.8	46.4	47.8	48.7	50.3
		BI_1	0.850	0.857	0.860	0.874	0.880	0.955
		$dist_{r\|nr}$	0.306	0.288	0.281	0.252	0.240	0.089

6.5. *A Comparison of the Data Collection Strategies*

All three experimental strategies use interventions in the LCS 2009 data, with successively more stringent stopping rules, as described in Subsections 6.3 and 6.4. Table 10 summarizes the experiments and compares them with the actual LCS 2009 data collection (with no interventions). For comparability, at the end of the respective data collections the entries in Table 10 are computed on the standard auxiliary vector defined in Section 4, which resembles the one used to produce the LCS estimates in 2009. The entries for the actual LCS 2009 data collection (the first line) are taken from the bottom line, "Final", in Tables 1 and 2.

Table 10. The three experimental strategies compared with the actual LCS 2009 data collection; response rate P (in per cent), RDF_{CAL}, BI_1, $dist_{r|nr}$ and reduction (in per cent) of the number of call attempts. Computations based on the standard x-vector explained in Section 4

End data collection	$100 \times P$	RDF_{CAL}			BI_1	$dist_{r\|nr}$	*Reduction in %*
		Benefits	Income	Employment			
Actual LCS 2009	67.4	−3.6	2.9	3.1	0.717	0.603	0.0
Strategy 1	63.9	−1.6	2.7	3.0	0.765	0.489	8.2
Strategy 2	58.9	−1.2	2.6	3.2	0.787	0.433	20.2
Strategy 3	50.3	1.0	1.0	2.3	0.808	0.383	36.4

Table 10 shows that each experimental strategy improves on the preceding one. The relative deviation RDF_{CAL} is reduced in each step for all three register variables used as study variables (if we disregard a slightly higher value for the variable *Employment* in Strategy 2). For *Income* and *Employment*, the major reduction in RDF_{CAL} occurs in the step from Strategy 2 to Strategy 3.

Both the balance and the response-to-nonresponse distance improve in each step. The distance $dist_{r|nr}$ drops from 0.603 to 0.383. The balance BI_1 increases from 0.717 to 0.808, the greatest step occurring from the actual LCS 2009 to Experimental Strategy 1. The balance shown in Table 10 is lower than the balance for the corresponding experimental strategy in Tables 6, 8 and 9. This is because the **x**-vectors are different; it is harder to achieve high balance for a more extensive vector.

A striking benefit from the experimental strategies is an implicit reduction of data collection cost through significantly fewer call attempts. To reach the 67.4% response in the complete 2009 LCS data collection, 53,258 attempts were used, but to reach the 63.9% response in experimental Strategy 1, only 48,883 attempts are used, a reduction of 8.2%. The reduction in call attempts is even more striking for the other two experimental strategies: 20.2% for Strategy 2 and 36.4% for Strategy 3. In practice, such cost savings should be used to improve other aspects of the survey design; one could for example afford a larger size sample s to begin with.

7. Discussion and Implications for the Future

The concepts proposed in this article, more specifically those presented in Sections 4 and 5, are general in scope and can be applied to a variety of sample surveys. We have chosen the 2009 Swedish Living Conditions Survey as an instrument to illustrate the use of these concepts, which are also being tested and evaluated in other surveys at Statistics Sweden.

In Section 5 we introduced the concept of imbalance, defined mathematically by the quadratic form $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$, Formula (5.2). The imbalance, a function of the chosen auxiliary vector $\mathbf{x}_k$, determines important tools presented in Section 5: The *balance of the set of respondents* and the *distance between respondents and nonrespondents*. Signs of a good data collection are increasing balance and decreasing distance during the course of the data collection.

In this article we have used these tools and paradata from the Swedish CATI-system to examine the data collection in the 2009 Swedish Living Conditions Survey. Earlier studies at Statistics Sweden had cast doubt on the merits of conducting a follow-up in the LCS; our

results in Sections 4 and 5 confirm those earlier findings. In Table 1 we studied the changes in the estimates for three register variables as the data collection progresses. The follow-up does not produce the improvement one would hope for.

Table 2 shows that the balance indicators BI_1 and BI_2 have a decreasing trend over the course of the LCS 2009 data collection. Contrary to reasonable expectations, the set of respondents is thus less balanced after the follow-up than at the end of the ordinary fieldwork. Furthermore, in Table 2 the distance $dist_{r|nr}$ shows an increasing rather than a decreasing trend as the data collection unfolds. This adds to earlier doubts about the efficiency of the current LCS data collection.

When the auxiliary vector $\mathbf{x}_k$ codifies membership in one of J mutually exclusive and exhaustive sample subgroups, the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ is particularly transparent: It is a sum of non-negative terms, $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D} = \sum_{j=1}^{J} C_j$, where C_j is the contribution to imbalance of the j:th group. This representation allows us to focus on each specific group in the data collection. Problematic groups are those for which C_j remains high throughout the data collection. This happens in the LCS 2009 data collection, as illustrated in Table 4. To reduce imbalance, one should direct the data collection so that all group contributions C_j are small in the end.

Section 6 described three experiments carried out by interventions in the LCS 2009 data file. A set of eight important sample subgroups was defined, and data collection was deemed terminated when subgroup response meets specified levels. These experiments, summarized in Table 10, showed that appropriate interventions in the data collection can bring considerable improvement – increased balance, reduced distance – compared with the actual LCS data collection. The cost savings realized by fewer call attempts might instead be used to improve other aspects of survey quality.

To use the conclusions from these experiments in practice, we must anticipate a "reasonable expectations" response rate to be used as a stopping rule for data collection in a group. In a regularly repeated survey such prior information is usually available, but this may not be the case in a survey carried out for the first time. But an assessment would be necessary.

In practice, whether or not responsive design has been used, nonresponse weighting adjustment for nonresponse will necessarily take place at the estimation stage. Balancing does not guarantee that the nonresponse bias is eliminated. The question then arises whether one could just as well delay the use of some of the auxiliary variables – those chosen to inform the responsive design – until the estimation stage, where they, usually together with other auxiliary variables, will determine the calibrated adjustment weights. In our opinion, although the responsive data collection can be an advantage, it does not eliminate the need for efficient calibrated weighting at the estimation stage. The question needs to be addressed in future research.

An issue in the LCS survey is the frame over-coverage; certain sample subgroups contain highly mobile people, some of whom may no longer reside in the country. It is clear that groups with chronically low response rate deserve particular attention in the data collection. In particular, improvements are needed to reach immigrants and younger persons whose style of living and interest in the survey may differ substantially compared with a majority of the population. Future savings could be realized by transferring interviewer effort from "easy-to-reach" respondents to the more problematic groups.

A central question is the choice of auxiliary variables to enter into the vector $\mathbf{x}_k$ that determines the imbalance $\mathbf{D}'\boldsymbol{\Sigma}_s^{-1}\mathbf{D}$ used to direct the data collection. This question needs to be addressed further in the future. Auxiliary variables are used first during the data collection and then with a somewhat different perspective at the estimation stage. In the data collection, the selected auxiliary variables serve to monitor the balance of the response set and the distance $dist_{r|nr}$ between respondents and nonrespondents. At the data collection stage, the auxiliary vector should thus be one that lends itself well to contrasting respondents with nonrespondents. At the estimation stage, on the other hand, the auxiliary vector serves to yield the most accurate estimates, particularly for the most important survey variables, and this vector is likely to contain more variables than the one used in monitoring the data collection.

8. References

Bethlehem, J., Cobben, F., and Schouten, B. (2011). Handbook of Nonresponse in Household Surveys. New York: Wiley.

Groves, R. (2006). Research Synthesis: Nonresponse Rates and Nonresponse Bias in Household Surveys. Public Opinion Quarterly, 70, 646–675. DOI: http://www.dx.doi.org/10.1093/poq/nfl033

Groves, R.M. and Heeringa, S.G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. Journal of the Royal Statistical Society: Series A, 169, 439–457. DOI: http://www.dx.doi.org/10.1111/j.1467-985X.2006.00423.x

Hörngren, J., Lundquist, P., and Westling, S. (2008). Effects of Number of Call Attempts on Nonresponse Rates and Nonresponse Bias – Result from Some Case Studies at Statistics Sweden. Proceedings 24th International Methodology Symposium, Statistics Canada, Session 17, catalogue no. 11-522-XIE. Available at: http://www5.statcan.gc.ca. (accessed September 19, 2013).

Laflamme, F. (2009). Experiences in Assessing, Monitoring and Controlling Survey Productivity and Costs at Statistics Canada. Proceedings of the 57th Session of the International Statistical Institute, South Africa. (August 16–22). Available at: http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/0049.pdf (accessed October 11, 2013).

Lundquist, P. and Särndal, C.-E. (2012). Aspects of Responsive Design for the Swedish Living Conditions Survey. R&D report 2012:1, Statistics Sweden. Available at: www.scb.se. (accessed September 19, 2013).

Mohl, C. and Laflamme, F. (2007). Research and Responsive Design Options for Survey Data Collection at Statistics Canada. Proceedings of the American Statistical Association, Section on Survey Research Methods. (July 29–August 2) Available at: https://www.amstat.org/sections/SRMS/Proceedings/y2007/Files/JSM2007-000421.pdf (accessed October 11, 2013).

Peytchev A, Baxter, R.K., and Carley-Baxter, L.R. (2009). Not All Survey Effort is Equal. Reduction of Nonresponse Bias and Nonresponse Error. Public Opinion Quarterly, 73, 785–806. DOI: http://www.dx.doi.org/10.1093/poq/nfp037

Peytchev, A., Riley, S., Rosen, J., Murphy, J., and Lindblad, M. (2010). Reduction of Nonresponse Bias in Surveys Through Case Prioritization. Survey Research Methods, 4, 21–29.

Peytcheva, E. and Groves, R.M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence on Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193–201.

Rao, R.S., Glickman, M.E., and Glynn, R.J. (2008). Stopping Rules for Surveys with Multiple Waves of Nonrespondent Follow-Up. Statistics in Medicine, 27, 2196–2213. DOI: http://www.dx.doi.org/10.1002/sim.3063

Schouten, B. and Bethlehem, J. (2009). Representativeness Indicator for Measuring and Enhancing the Composition of Survey Response. RISQ work package 8, deliverable 9. Available at: http://www.risq-project.eu/. (accessed September 19, 2013).

Schouten, B., Cobben, F., and Bethlehem, J. (2009). Indicators for the Representativeness of Survey Response. Survey Methodology, 35, 101–113.

Schouten, B., Shlomo, N., and Skinner, C. (2011). Indicators for Monitoring and Improving Representativeness of Response. Journal of Official Statistics, 27, 231–253.

Särndal, C.-E. (2011a). Dealing with Survey Nonresponse in Data Collection, in Estimation. Journal of Official Statistics, 27, 1–21.

Särndal, C.-E. (2011b). Three Factors to Signal Nonresponse Bias, with Applications to Categorical Auxiliary Variables. International Statistical Review, 79, 233–254. DOI: http://www.dx.doi.org/10.1111/j.1751-5823.2011.00142.x.

Särndal, C.-E. and Lundström, S. (2005). Estimations in Surveys with Nonresponse. New York: Wiley.

Särndal, C.-E. and Lundström, S. (2008). Assessing Auxiliary Vectors for Control of Nonresponse Bias in the Calibration Estimator. Journal of Official Statistics, 24, 251–260.

Särndal, C.-E. and Lundström, S. (2010). Design for Estimation: Identifying Auxiliary Vectors to Reduce Nonresponse Bias. Survey Methodology, 36, 131–144.

Wagner, J. (2008). Adaptive Survey Design to Reduce Nonresponse Bias. Ph.D. thesis, University of Michigan, Ann Arbor. Available at: http://www.google.se/-books?hl=sv&lr=&id=ilVbF2qPlTgC&oi=fnd&pg=PR3&dq=Wagner+(2008)+Adaptive+survey+design+to+reduce+nonresponse+bias&ots=iqKD3_BnEQ&sig=cRtrqb-zLogBQERLCHdBiuhyNu2k&redir_esc=y#v=onepage&q=Wagner%20(2008)%20Adaptive%20survey%20design%20to%20reduce%20nonresponse%20bias&f=-false (accessed October 11, 2013).

Wagner, J. (2012). Research Synthesis: A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. Public Opinion Quarterly, 76, 555–575. DOI: http://www.dx.doi.org/10.1093/poq/nfs032

Wagner, J. and Raghunathan, T.E. (2010). A New Stopping Rule for Surveys. Statistics in Medicine, 29, 1014–1024, DOI: http://www.dx.doi.org/10.1002/sim.3834

Received September 2011
Revised December 2012
Accepted June 2013

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 583–607, http://dx.doi.org/10.2478/jos-2013-0041

Utilising Expert Opinion to Improve the Measurement of International Migration in Europe

Arkadiusz Wiśniowski[1], *Jakub Bijak*[1], *Solveig Christiansen*[2], *Jonathan J. Forster*[1], *Nico Keilman*[2], *James Raymer*[1,3], *and Peter W.F. Smith*[1]

In this article, we first discuss the need to augment reported flows of international migration in Europe with additional knowledge gained from experts on measurement, quality and coverage. Second, we present our method for eliciting this information. Third, we describe how this information is converted into prior distributions for subsequent use in a Bayesian model for estimating migration flows amongst countries in the European Union (EU) and European Free Trade Association (EFTA). The article concludes with an assessment of the importance of expert information and a discussion of lessons learned from the elicitation process.

Key words: Bayesian modelling; elicitation; expert knowledge; migration statistics; Delphi Survey.

1. Introduction

To fully understand the causes and consequences of international movements in Europe, researchers and policy makers need to overcome the limitations of the various data sources, including inconsistencies in data availability, quality and collection mechanisms. For example, in 2007, Germany reported receiving 15,515 migrants from Spain, whereas Spain only reported sending 3,601 migrants to Germany. From this single example, many questions arise: Why are the two numbers so different? How accurate are the data provided by the two countries? Could measurement be responsible for some of the difference? In this article, we describe our attempt to answer these questions by collecting information from experts on migration data.

[1] Southampton Statistical Sciences Research Institute, University of Southampton, University Road, Southampton SO17 1BJ, UK. Emails: a.wisniowski@soton.ac.uk, j.bijak@soton.ac.uk, j.raymer@soton.ac.uk, j.j.forster@soton.ac.uk and p.w.smith@soton.ac.uk

[2] Department of Economics, University of Oslo, P.O. Box 1095 Blindern, N-0317 Oslo, Norway. Emails: s.g.christiansen@econ.uio.no and n.w.keilman@econ.uio.no

[3] Australian Demographic and Social Research Institute, Australian National University, Coombs Building, Canberra AT 0200, Australia. Email: j.raymer@soton.ac.uk

Acknowledgments: This research is part of the integrated Modelling of European Migration (IMEM) project funded by the New Opportunities for Research Funding Agency Co-opertion Europe (NORFACE). We gratefully acknowledge the help of the following persons, who acted as an expert, tested the pilot survey questionnaire, or contributed otherwise to the development of the questionnaire: Guy J. Abel, Corrado Bonifazi, Harri Cruijsen, Frank Heins, Michael Jandl, John Kelly, Ewa Kępińska, Dorota Kupiszewska, Marek Kupiszewski, Giampaolo Lanzieri, João Peixoto, Nicolas Perrin, Michel Poulain, and Rob van der Erf. It should be stressed that all persons contributed in their own personal capacity to the project. We also wish to acknowledge comments received from reviewers and from the editor of this journal, which greatly improved the presentation of this article.

This information is gathered for use as prior inputs into a Bayesian model for harmonising and estimating international migration flows amongst the 31 countries in the European Union (EU) and the European Free Trade Association (EFTA) (Raymer et al. 2013).

Bayesian statistical methods are particularly adept at handling data from different sources and are ideal for situations in which some of the data are inadequate or missing. Additional expert information can be included in the form of prior distributions reflecting expert beliefs and judgements. The resulting estimates are then based on posterior distributions, which combine these expert beliefs with other available information, including all relevant data sources and covariates. The posterior distributions can also be used to quantify uncertainty in the estimates, providing the users, such as governments and planning agencies, with valuable additional information to design their policies directed at supplying particular social services or at influencing levels of migration (Bijak and Wiśniowski 2010).

The structure of this article is as follows. First, we describe the underlying conceptual framework for harmonising and estimating flows of international migration within Europe. Second, we outline our approach for eliciting information from experts concerning the characteristics of the reported statistics on flows. Third, we present our methodology for translating this expert information into informative prior distributions for subsequent use in the model for migration flows. We illustrate the method with an application to a European migration flow matrix for 2002–2008. The article ends with an assessment of the importance of expert information and a discussion of lessons learned from the elicitation process, followed by some conclusions.

2. A Conceptual Framework for Modelling Migration

There have been several attempts to harmonise international migration flow statistics in Europe. Poulain (1993) developed a constrained optimisation procedure to minimise the differences between two origin-destination migration flow tables representing sending and receiving country reported statistics. His 'correction factor' method has been extended more recently by Poulain and Dal (2008), Abel (2010) and De Beer et al. (2010). Van der Erf and Van der Gaag (2007) and DeWaard et al. (2012) developed iterative hierarchical procedures to allow countries providing better data to have more weight in the estimation. Finally, Nowok (2010) proposed a probabilistic framework for harmonising international migration statistics (see also Nowok and Willekens 2011). Our approach to harmonising migration flows differs from these works by the emphasis on modelling the measurement aspects of the reported statistics and by providing measures of uncertainty. In this section, we introduce the underlying conceptual framework for estimating migration flows in Europe, which has been developed as a Bayesian model in (Raymer et al. 2013). In the following section, we turn to the main focus of this article: the elicitation of expert judgements.

The framework we have developed permits expert opinion to be combined with the data on migration flows and covariate information to strengthen the inference. The approach also facilitates the combination of multiple data sources, with their differing levels of error, as well as prior information about the structures of migration processes, into a single

prediction with associated measures of uncertainty. Given the substantial inconsistencies in reported statistics on international migration flows in Europe (Poulain et al. 2006), the elicitation of expert opinion concerning various aspects thereof is critical for the success of the whole modelling exercise.

In terms of measurement, true flows are assumed to be consistent with the United Nations (1998) recommendation for long-term international migration:

> A person who moves to a country other than that of his or her usual residence for a period of at least a year (12 months), so that the country of residence effectively becomes his or her new country of usual residence. From the perspective of the country of departure, the person will be a long-term emigrant and from that of the country of arrival, the person will be a long-term immigrant (United Nations 1998, p. 18).

Place of 'usual residence' is defined as

> The country in which a person lives, that is to say, the country in which he or she has a place to live where he or she normally spends the daily period of rest. Temporary travel abroad for purposes of recreation, holiday, visits to friends and relatives, business, medical treatment or religious pilgrimage does not change a person's country of usual residence (United Nations 1998, p. 17).

Finally, the United Nations definition we have adopted includes undocumented (irregular) migrants. In practice, the migration statistics in most countries do not cover undocumented migrants (for obvious reasons). Thus, one of the aims of the presented approach is to use expert judgement to address the levels of this aspect of migration.

Our approach to measuring migration takes into account four aspects assumed to be independent: (i) accuracy of data collection system, (ii) duration criteria used to qualify migrants that differ from the twelve months in the UN definition, (iii) undercount and (iv) coverage of migrants. Let z^k_{ijt} denote the counts (flows) from country i to country j during year t reported by country k, either the sending $k = i$ or receiving $k = j$. The interest of this research is to estimate y_{ijt} – the true unknown flow of migration from country i to country j in year t. It includes migration flows to and from the rest of world. Note that for each y_{ijt} there are potentially two reported flows: z^i_{ijt} and z^j_{ijt}.

We assume that the observed data z reflect the true flows y, distorted by the above mentioned deficiencies of the migration statistics, that is

$$z^k_{ijt} = y_{ijt} \times dur_k \times und_k \times cov_k \times err^k_{ijt}. \tag{1}$$

The variance of the general error term err^k_{ijt} measures the accuracy of the data collection system for country k. It informs the end users of the outcomes of this study on the quality of the data and measurement mechanisms utilised to collect the data. The number of parameters required to capture differences in accuracy depends on our typology of collection systems, and their relative ability to capture migration flows, regardless of definition and coverage. Here, we distinguish three types of systems: (1) interlinked population registers in the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden), which exchange migration information; (2) other good-quality registers (The Netherlands, Germany, Austria, Belgium, Switzerland, and immigration in Spain) and

(3) less reliable registers and survey-based systems (Poland, Bulgaria, Estonia, Lithuania, Latvia, Italy, Slovenia, Slovakia, Romania, the Czech Republic, Greece, Hungary, Liechtenstein, Malta, France, Luxembourg, Portugal, United Kingdom, Cyprus, Ireland, and emigration from Spain). Our typology of accuracy is based on reports from the MIMOSA project (Kupiszewska and Wiśniowski 2009; Van der Erf 2009) and our own assessment of the data quality in Europe.

The duration parameter dur_k reflects the difference between the duration of stay criterion adopted by the country k data collection system and the baseline twelve-month criterion of the UN. For example, if a given country uses a six-month criterion, the number of true migrants (i.e., residing for twelve months or more) should be smaller than the reported number of migrants, independent of the other measurement deficiencies. Note that in practice the duration is intended or planned rather than actual.

We interpret the undercount parameter und_k as a fraction of the true flow that is captured by the data collection system in a given country. We propose two classifications here. In both of them, we work with two levels of undercount. The first one distinguishes between intra-European flows and those to and from the rest of the world. In the second one, we classify some countries as having high undercount and others as having low undercount; see Section 5 for details. The latter classification of countries with low or high undercount is based on our own assessment, as well as reports from the various projects (Poulain et al. 2006; Kupiszewska and Wiśniowski 2009; Van der Erf 2009).

The country-specific error parameters cov_k reflect the discrepancies between the observed data and the true flows that are not captured by the more general undercount parameters. These often include certain subgroups, such as international students or refugees, in the reported migration flows (Poulain et al. 2006; Kupiszewska and Wiśniowski 2009). Furthermore, we assume these parameters to lie between zero and one and interpret them as the differences in coverage with respect to the United Nations definition of migration. Given that the coverage parameters are country-specific, we assume that they measure the proportions of migration covered in relation to the true flows. For the Nordic countries and the Netherlands, these parameters are constrained to one, that is, we assume that there are no coverage errors for these countries. This assumption ensures identifiability of the parameters. For the rest of the countries, we use noninformative prior distributions. We considered the elicitation of the country-specific prior densities infeasible for the scale of our project. This approach would require at least five experts for each of the 31 countries under study. Also, since the coverage aspect of the measurement model did not utilise expert judgements, it is not discussed further in this article.

3. Obtaining Expert Information

The approach described in Section 2 requires prior information on the quality of data sources, differences in various aspects of measurement and covariates used to predict missing data. In this case, external expert judgement was sought only on the data and measurement aspects of the underlying migration flows. The experts in data collection systems were asked to rate the credibility they give to different types of migration data collected from different types of collection mechanisms (e.g., survey versus register), and to compare sending country data (i.e., emigration flows) with receiving country data

(i.e., immigration flows). Experts were also asked about the bias (e.g., systematic undercount) in the reported migration flow statistics. Each expert was asked to give us a set of values concerning certain parameters, which we then converted into probability distributions. The totality of resulting expert opinions was subsequently combined into a single set of distributions, allowing for the introduction of yet another source of uncertainty, related to the heterogeneity of experts.

To facilitate the elicitation of expert judgements, a two-stage process was used within a Delphi survey framework, whereby the expert opinions were allowed to be informed and influenced by other experts' views. This process provided a convenient avenue for the exchange of opinions and views as well as for clarifying any ambiguities as to the underlying concepts and ideas.

The elicitation of expert opinion to construct probability distributions has a long history (O'Hagan et al. 2006). In general, the acquisition of such information is a very difficult task (Kadane and Wolfson 1998). Asking an expert to draw a distribution would assume he or she has a statistical background or require us to provide such training. In our study, we could not guarantee all experts had a statistical background and did not have the time or resources to provide training. As a result and based on the feedback we received from pretesting the questionnaire, we had to limit the use of statistical terms, such as 'quantile', 'distribution', 'variance' and 'precision'. For this reason, we followed the elicitation guidelines of O'Hagan (1998) and O'Hagan et al. (2006), as well as an example of elicitation of opinion from 'non-statisticians' in Szreder and Osiewalski (1992).

From our heterogeneous group of experts, we sought basic information on particular values associated with the measurement of migration flows, which we then converted into probability distributions that could be used in our computations. After the first Delphi round, experts were provided with the densities resulting from our interpretation and

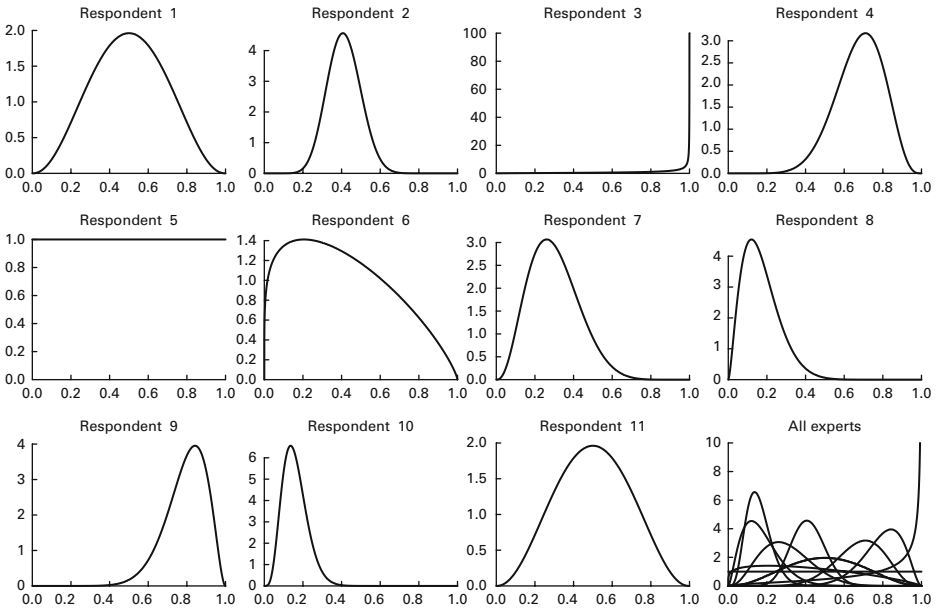

Fig. 1. Selected graphical representations of expert answers from Round 1: Undercount of emigration

parametrisation of their answers (see Figure 1 and Section 4), as well as the anonymous results from other experts in the study. This allowed them to reconsider and revise their opinions.

When formulating questions, it is important to prevent respondents from being overconfident in their opinions. For example, questions about means or medians may lead to anchoring the answer and lowering the uncertainty about the tails of the distribution (Kadane and Wolfson 1998; Rowe and Wright 2001). To avoid this problem, we constructed questions that focused on ranges of values with direct interpretations and the certainty about these ranges. Each certainty could then be interpreted as a probability that a given parameter lies within a specified range.

Experts were free to select the upper and the lower bounds of the intervals. There is an extensive literature on the issue of fixed versus variable interval bounds; see, for example, Kadane and Wolfson (1998), Garthwaite et al. (2005) or Dey and Liu (2007) for reviews. One problem with preselected intervals is that uncertainty may vary across individuals in complex ways, and hence it is difficult to find an optimal design of a preselected interval. On the other hand, lower and upper quantiles (often used in preselected intervals) have the advantage that they can be assessed by a method of bisection, as described in Garthwaite et al. (2005). From the literature on fixed and preselected intervals they also concluded that there is conflicting evidence as to which method performs better.

In one of the questions to our experts, we asked about their subjective probability concerning the accuracy of the data collection system (see Subsection 4.3). As pointed out in the literature, elicitation of probabilities is a difficult task. The perception of probability may vary depending on the formulation of the question, for example, odds ratios tend to be more extreme than the probability specified within a range [0, 1] (Goodwin and Wright 1998). Another issue is viewing uncertainty in terms of frequencies rather than subjective probabilities (Gigerenzer 1994; Kadane and Wolfson 1998) and forgetting about the context of an event under consideration. Hence, in the formulation of our question, we followed the advice of Gigerenzer (1994) of asking about proportions and providing the context of the subject.

3.1. Delphi Technique

The Delphi technique is a method used to obtain information from a group of experts in order to make judgements and forecasts when extensive or reliable data in the field of enquiry are not available (Rowe and Wright 1999). It was first developed by the RAND Corporation for US military use in the 1950s. More recently, and in the context of international migration in Europe, this technique was applied to (i) forecast migration between Central and Western Europe after the fall of communism (Drbohlav 1996), (ii) the MIGIWE (Migration and Irregular Work in Europe) project to gain information on irregular foreign employment in Austria following the 5th Enlargement of the EU (Jandl et al. 2007) and (iii) the IDEA (Mediterranean and Eastern European Countries as new immigration destinations in the European Union) project to augment forecasting models for seven European countries (Wiśniowski and Bijak 2009; Bijak and Wiśniowski 2010).

In a Delphi survey, the elicitation of expert opinions takes the form of an anonymous questionnaire with multiple rounds, where the experts report their subjective beliefs on the

topics in question. Between rounds, experts are provided with feedback on the answers in the preceding round, including qualitative arguments in support of various views. The experts then complete the next round of the survey where they are free to alter their previous answers in light of the new information provided by the feedback.

According to Rowe and Wright (2001), the Delphi technique is most reliable when there are between five and 20 respondents who are experts in the field of enquiry and when there is heterogeneity among the experts. The questions should be sufficiently comprehensive to contain the relevant information but not cause information overload. The final round answers are usually weighted equally. Past evaluations have shown that the answers from the final round Delphi surveys are more accurate than other approaches using only one expert, focus groups or single-round questionnaires. By using an anonymous questionnaire instead of a group meeting, one avoids group pressure and the domination of the group by some individuals. The Delphi method may also lead to better results because the experts think more carefully when responding when they know that their answers will be given as feedback to other experts.

3.2. Constructing the Questionnaire

For our project, the elicitation process consisted of two rounds (hereafter Round 1 and Round 2) and involved eleven external experts. We selected the experts from among those international colleagues who we thought would be knowledgeable about the measurement of international migration in several countries. The online questionnaire was pretested by an additional two external experts and two of our team members. The survey was preceded by an invitation letter, in which the aim of the project and the purpose of the questionnaire were explained. The experts were asked to give their opinion about how specific measurements of international migration deviate from the benchmark of the United Nations definition of a long-term migrant (see Section 2).

The Round 1 questionnaire included a definition of a long-term migrant according to the United Nations definition discussed above plus 14 questions grouped into four sections. Each section contained a specific set of closed questions and an open question, in which experts were allowed to express their comments or arguments related to their answers. In all questions, experts were asked to provide their answers in terms of percentages, and to state how certain they were about their answers, that is, 50%, 75%, 90%, 95% or Other. The first three sections of the questionnaire were restricted to intra-EU/EFTA migrants, while the fourth section concerned migration between the EU/EFTA countries and the rest of the world. Finally, the experts were also allowed to provide general comments or suggestions, as well as to ask questions of their own. The full questionnaire is available for download at [http://www.imem.cpc.ac.uk].

The undercount of migration between EU and EFTA countries and from or to the rest of the world was the focus of Section A (Questions 1–3) and Section D (Questions 12–14) of the questionnaire respectively. Here, experts were asked to provide their judgements and uncertainty regarding the lowest and highest percentages of the possible undercount of emigration and immigration in the published statistics. To do this, the experts needed to consider a nonspecific, hypothetical European country with a good population register and migration definitions corresponding exactly with the United Nations (1998)

recommendation. In other words, the experts were asked to think of migration collection systems rather than specific country experiences.

The focus of Section B of the questionnaire (Questions 4–6) concerned the duration of stay criteria included in the definition of migration. In Europe, different timing criteria are used by different countries and these questions aimed at assessing how this might affect the relative levels of reported migration. Thus, in Question 4, experts were asked how much, in percentage terms, the level of migration would be for a duration of stay criterion of six months instead of twelve months. Question 5 asked for the difference between three- and six-month criteria.

Finally, the questions in Section C were aimed at obtaining information about the accuracy of population registers in measuring migration. Experts were asked to consider registers in which there was no systematic bias and with random factors being the main source of error. In Questions 7 to 11, experts were asked to provide their beliefs and certainty regarding published statistics falling within an interval from minus 5% to plus 5% compared to the true total level of emigration and immigration.

All eleven respondents from Round 1 took part in Round 2 of the survey. Of these, nine chose to change their answers to one or more of questions in Round 2. Further information about the changes in the experts' opinions between the two rounds can be found in the following section. The questionnaire in Round 2 consisted of the same set of questions as in Round 1. It also contained anonymised answers from Round 1 and the arguments used to support the various views, including the underlying reasons for different assessments. The experts also had the option to look at graphical representations of their individual answers, examples of which are shown in Figure 1. Details on how these distributions were compiled are provided in Subsection 4.1.

4. Translating the Expert Information into Prior Distributions

In this section, we explain how the opinions and judgements obtained in the first and second round of the Delphi survey were translated into prior distributions for the parameters introduced in Section 2. The parameters in question are used to address undercount, duration of stay and accuracy of measured migration flows.

The construction of prior densities based on expert answers was a three-step process. First, having obtained the raw answers to a given question about some parameter θ, we identified a distribution, that, in our opinion, reflected the expert judgements about the θ most appropriately. Second, we constructed a prior density $f_i(\theta)$ for each expert i, $i = 1, \ldots, n$. Third, we combined the individual densities into a single prior density:

$$P(\theta) \sim \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) \tag{2}$$

We chose to have an equally-weighted opinion pool because it allowed us to have a simple, robust and general method for aggregating expert knowledge. Aggregation methods based on weighting, such as that of Cooke (1991), require a separate elicitation round in which each expert is asked about a particular variable, of which the real value is known to the facilitator but not to the expert. In our situation, we did not know the real values of any of the parameters. Therefore, we assigned equal weights to the experts. The

equal weights also allowed the different and sometimes opposing assessments to be fed into the estimation model. Smoothing techniques or fitting a parametric distribution to the expert answers, for example, would have reduced the amount of information provided by the experts. Another option, which could be explored in the future work, would be to perform Bayesian model averaging over models with each single expert prior distribution as a separate input. For a discussion about the benefits and consequences of the various ways expert opinions can be combined, we refer the reader to Clemen and Winkler (1990) and O'Hagan et al. (2006).

4.1. Undercount of Emigration and Immigration

4.1.1. Method for Constructing the Prior Density

In the first and fourth section of the Delphi questionnaire, experts were asked to provide answers to the following question about undercount of migration within Europe and to and from the rest of world. In the preamble to the question on undercount, the reference to the baseline UN definition was made. The question was formulated as:

> [. . .] Consider a European country with a good population register, e.g., Sweden or Finland, that has fully adopted the UN definition. Because migrants do not always have sufficient incentives to report their moves to the relevant authorities, migration statistics are often lower than the true total level. For immigrants this difference is thought to be smaller than for emigrants.
>
> (a) *By how many per cent do you expect that emigration (or immigration) flows are undercounted in the published statistics, as compared to the true total level of emigration (immigration)? Please provide a range in percentages.*
>
> (b) *Approximately how certain are you that the true undercount will lie within the range that you provided above?*

Let P_1 and P_2 denote the lower and upper percentages stated by an expert about undercount and c denote the certainty about the range (P_1, P_2). The underlying assumption regarding undercount is that $P \in [0, 1] \times 100\%$, which is

$$(1 - P)y = z, \tag{3}$$

where y are true flows and z are reported flows. Then $(1 - P)$ can be interpreted as a fraction of the true flow which is captured in the reported data. A couple of the answers provided by experts in the first round were not meaningful, suggesting some difficulties were experienced in interpreting the questions. We addressed this issue in the Round 2 questionnaire (see the following section).

To convert the experts' answers into prior distributions for the parameters, we first had to identify which probability distributions would both accurately reflect experts' beliefs and work well with the underlying conceptual framework introduced in Section 2. We considered three densities: piecewise uniform, logit-normal and beta. These densities were chosen because they could be constrained to values between zero and one and they were flexible in terms of shapes. Besides, as opposed to truncated distributions such as normal or log-normal, their parameters could be easily calculated.

Table 1. Experts answers to question 1 – undercount of emigration

Respondent	1	2	3	4
Lowest percentage, P_1	20	30	50	4
Highest percentage, P_2	80	50	90	8
Certainty, c	90	75	90	5

To illustrate the differences between various densities, consider four answers of the experts to Question 1 set out in Table 1. For example, Respondent 2 believes that the emigration flows in the published statistics are undercounted by 30% to 50% with a probability of 75%. Respondent 4, on the other hand, believes that the reported flows of emigrants are only 4% to 8% too low, which represents a very precise range, but his or her certainty is only 5%. It should be intuitive that the wider the range of undercount, the larger the certainty should be. Note that in Round 1 of the Delphi survey, almost all answers were consistent with this rule. For the questions concerning undercount, only one expert indicated relatively large range with a small level of certainty. This led to some computational and interpretation problems.

For the case of the piecewise uniform densities, the computation was straightforward. We assumed that the certainty level c provided by a given respondent corresponded with the probability mass between P_1 and P_2. The remainder, $(1 - c)$, was proportionally distributed between $[0, P_1]$ and $[P_1, 1]$. Thus the quantiles of the resulting piecewise uniform density were

$$q_1 = \frac{(1-c)P_1}{1 + P_1 - P_2} \quad \text{and} \quad q_2 = \frac{(1-c)(1-P_2)}{1 + P_1 - P_2}. \tag{4}$$

The resulting piecewise uniform densities, after transformation into undercount using Equation (3), are presented in the first row of Figure 2.

In the case of the logit-normal density, it was assumed that

$$\begin{cases} \mu + \sigma\,\Phi^{-1}(q_1) = \dfrac{\log(P_1)}{1 - \log(P_1)} \\ \mu + \sigma\,\Phi^{-1}(q_2) = \dfrac{\log(P_2)}{1 - \log(P_2)} \end{cases} \tag{5}$$

where μ and σ are expected value and standard deviation of the underlying normal density and Φ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution. Two specifications of q_1 were considered. In the first one, the probability mass c lies between P_1 and P_2 and the remainder, $(1 - c)$, symmetrically distributed between $[0, P_1]$ and $[P_2, 1]$:

$$q_1 = \frac{1-c}{2} \quad \text{and} \quad q_2 = \frac{1+c}{2} \tag{6}$$

The second specification is based on quantiles as in the piecewise uniform approach, as given by Equation (4). The resulting densities for these two approaches, after

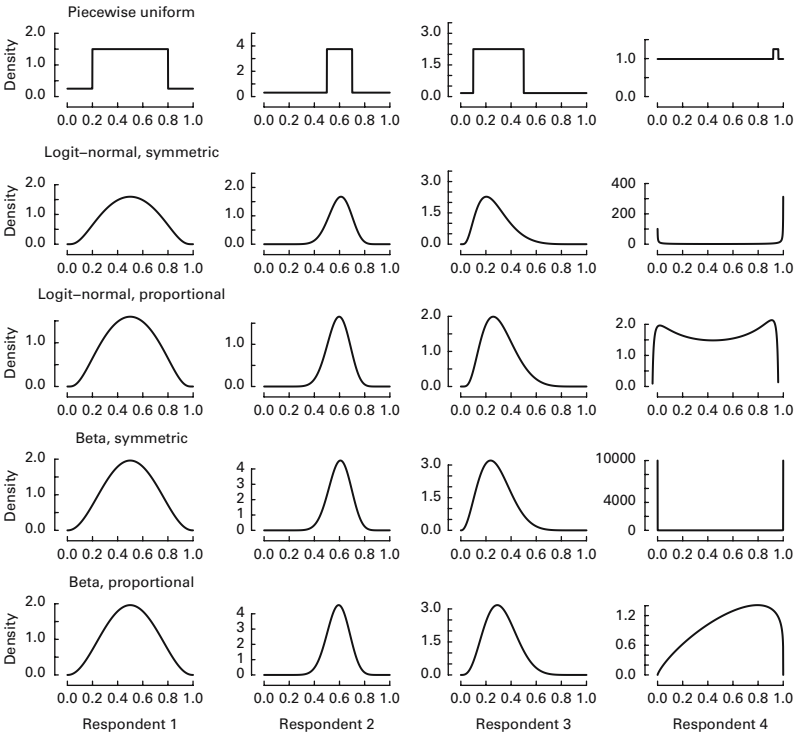

Fig. 2. Densities for four experts with various specifications

transformation using Equation (3), are presented in second and third row of Figure 2 respectively.

Finally, two sets of quantiles were also considered for the beta distribution. The parameters α and β of the beta density were computed by solving a set of two equations:

$$\begin{cases} F_b^{-1}(P_1, \alpha, \beta) = q_1 \\ F_b^{-1}(P_2, \alpha, \beta) = q_2 \end{cases}, \tag{7}$$

where F_b^{-1} is an inverse cumulative distribution function of the beta distribution. This was achieved by finding roots of the following expression:

$$\sum_{i=1}^{2} \left[F_b^{-1}(P_i, \alpha, \beta) - q_i\right]^2, \tag{8}$$

where q_1 and q_2 were either proportionally (4) or symmetrically (6) distributed. Vector $(\alpha_0 = 1, \beta_0 = 1)$ was used as a starting point for this algorithm. The densities obtained for the four example experts are presented in Figure 2 in the fourth and fifth rows for symmetric and proportional quantiles respectively.

From all of the approaches considered to translate and represent the subjective expert opinions, the beta density with proportional quantiles was ultimately chosen. Piecewise uniform was rejected because it produced relatively crude results. The logit-normal and

beta distributions with symmetric quantiles also tended to yield unintuitive shapes, especially in cases where experts assigned more certainty to regions close to zero or 100% undercount. Such a case is represented by Respondent 4 in Figure 2. Both symmetric approaches (logit-normal and beta in rows 2 and 4, respectively) are bimodal with most of the probability mass assigned close to zero and one, which was considered to be a rather implausible representation of an expert's opinion. The proportional logit-normal approach also resulted in a bimodal density and was rejected (depending on relative sizes of μ and σ, the logit-normal distribution has one or two modes; see Johnson 1949, pp. 158–159).

4.1.2. Feedback to Experts and Round 2 Questionnaire

As mentioned in Subsection 3.2, the second round of the Delphi survey included anonymised answers from the first round, together with arguments used to support the views and reasoning of various experts. Besides this feedback, we also took advantage of Round 2 to ensure a shared understanding of all underlying concepts among the participants. For example, in Round 1, a few of the experts gave answers to some of the questions on undercount which lay outside the 0–100% range, making interpretation difficult in terms of Equation (3). This suggests that the undercount was understood as 'how many times larger are the true flows, in comparison to the reported data', that is,

$$y = (1 + \alpha)z \tag{9}$$

where y and z are the true flows and reported data, respectively, and α denotes magnitude of how many times the true flows are larger than the reported data. Hence, if an expert provided at least one number α falling outside of a range [0, 1], both answers were treated according to the interpretation implied in Equation (9) and recomputed to be $P = 1 - 1/(1 + \alpha)$, where P is the undercount factor as in Equation (3). Those experts who in Round 1 had provided answers outside the 0–100% range were contacted to confirm that our interpretation of their answers was correct. In Round 2, it was specifically stressed for some of the questions that the answer must lie in the interval 0–100%.

4.1.3. Expert Answers and Resulting Prior Densities

The answers provided by the experts to the question on undercount of emigrants within EU and EFTA countries, converted into proportions, are presented in Table 2. For the

Table 2. Experts' answers concerning undercount of emigrants

Resp.	1	2	3	4	5	6	7	8	9	10	11
Round 1											
P_1	0.20	0.30	0.00	0.50	0.10	0.04	0.10	0.01	0.80	0.05	0.20
P_2	0.80	0.50	10.00	0.90	0.30	0.08	0.40	0.30	0.95	0.20	0.80
c	0.90	0.75	0.50	0.90	0.20	0.05	0.75	0.90	0.50	0.75	0.90
Round 2											
P_1	0.25	0.30	0.10	0.50	0.10	0.04	0.20	0.01	0.50	0.50	0.30
P_2	0.75	0.50	1.00	0.70	0.30	0.08	0.50	0.50	0.75	0.90	0.90
c	0.90	0.75	0.50	0.75	0.50	0.05	0.50	0.90	0.75	0.90	0.90

Resp. – Respondent, P_1 – Lowest proportion, P_2 – Highest proportion, c – Certainty.

emigration undercount we observe that two respondents did not change their opinions between two rounds of the study, while three increased their confidence. Some of the experts provided wide percentage spans with large confidence (e.g., Respondents 1, 4, 10, 11), while others gave a comparatively narrow range with lower certainty (Respondents 2, 6 and 9). Respondent 3 provided a percentage range exceeding the envisaged 0–100% range with a relatively small confidence. Hence, we interpreted it as the undercount given in Equation (9) and transformed it accordingly. In the Round 2 answers, we observe that only two experts lowered their certainty.

In Figure 3 and Figure 4, we present the Round 1 and Round 2 expert opinions regarding factors $(1 - P)$, that is, the parameters und_k which capture the emigration and immigration undercount, respectively, transformed into beta densities with proportional quantiles. The individual curves were used to construct mixed prior densities (bold curves in Figure 3 and Figure 4) for the und_k parameters.

The prior density for emigration undercount, based on answers from Round 1 (bold curve in the left plot of Figure 3), is weakly informative in the sense that there is no clear region of undercount that would be indicated by the majority of experts. The resulting density has four modes. Mean undercount is 52%, with a standard deviation of 27%. The corresponding Round 2 prior density is unimodal, with a mean of 56% and a standard deviation of 22%. Unimodality and lower spread in the second round suggests there has been some convergence of the answers.

Comparing the prior densities of the immigration undercount answers with those of emigration, we observe a shift of the probability mass from the region of a very high undercount (near zero) to the values suggested by the majority of experts, that is around 60–80%. The Round 1 prior density mean is 68% with standard deviation of 25%; in the second round these values changed to 72% and 18%. Again, the three modes of the Round 1 prior were replaced by a unimodal density in Round 2, which is a sign of convergence in judgements.

The overall large standard deviation and a relatively 'flat' shape of the distribution of the mixture densities reflects the heterogeneity of expert judgements about the undercount. It may also stem from different experiences of the experts with migration statistics. That is, their opinions may have been based on the systems known best to them or on their lack of knowledge regarding other systems.

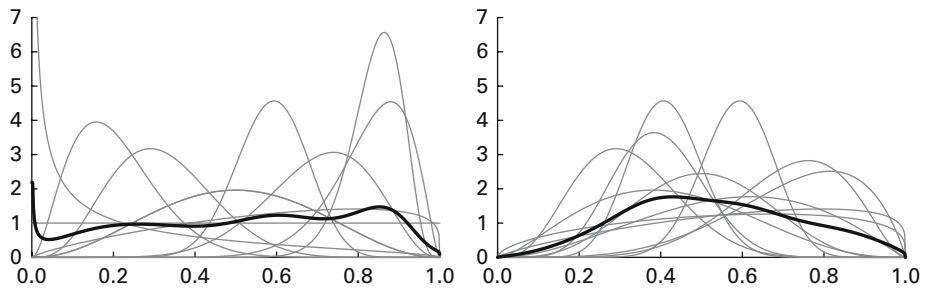

Fig. 3. Expert answers transfomed to densities for undercount of emigrants parameter, Round 1 (left) and Round 2 (right)

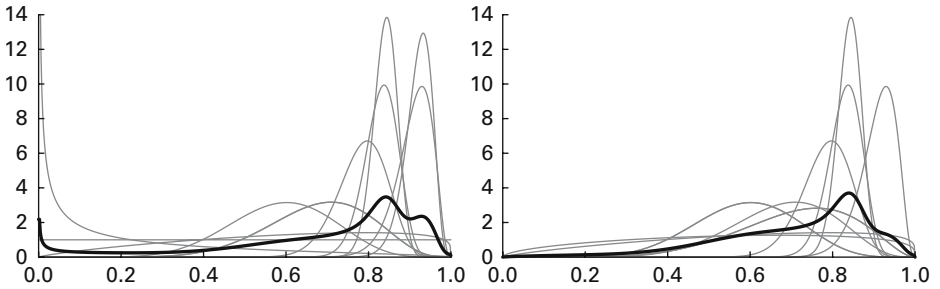

Fig. 4. Expert answers transfomed to densities for undercount of immigrants parameter, Round 1 (left) and Round 2 (right)

As shown in Figure 5 and Figure 6, the expert assessments of the undercount of emigration to and immigration from the rest of the world are more ambiguous than for intra-European migration. Four experts stood by their Round 1 answers in Round 2 and two reduced their confidence and changed the undercount range.

Consensus among experts concerning the undercount of rest of world flows was not reached. Respondents pointed out that the data on non-EU citizens are in general better captured due to more requirements for them than the data on nationals or other EU citizens. This would reduce the undercount. On the other hand, including the undocumented migrants in our estimates has had a reverse effect and blurs its evaluation.

4.2. Overcount Due to Different Duration of Stay Criteria

4.2.1. Method for Constructing the Prior Density

The duration of stay parameters capture the effects of different timing definitions used to qualify migrants. We assume that, in the presence of no undercount and the same accuracy, the shorter the duration measure, the greater the number of migrants:

$$y_p < y_{12} < y_6 < y_3 < y_0, \tag{10}$$

where the subscripts of the true flow y denote the durations with p = permanent, 12 = twelve months, 6 = six months, 3 = three months and 0 = no time limit. For

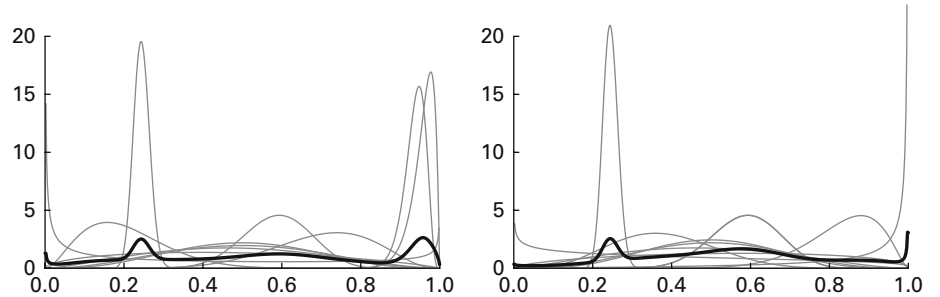

Fig. 5. Expert answers transformed to densities for undercount of emigrants to rest of world parameter, Round 1 (left) and Round 2 (right)

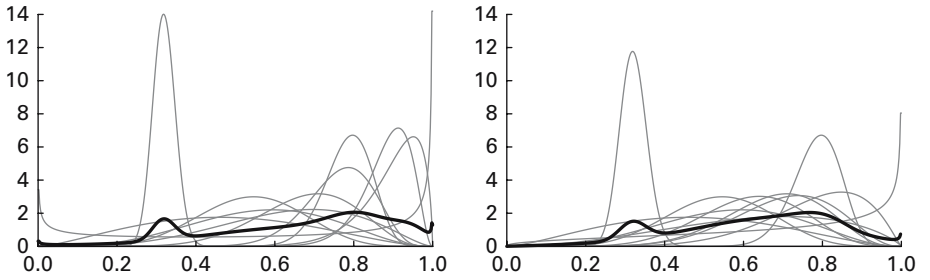

Fig. 6. Expert answers transformed to densities for undercount of immigrants to rest of world parameter, Round 1 (left) and Round 2 (right)

simplicity, we suppress country and time subscripts. Our benchmark criterion was twelve months, following the United Nations (1998) definition described in Subsection 3.2. The overcount of the number of migrants, due to the different duration criterion in the reported data z, can be expressed by a factor dur_s in the equation

$$z = dur_s \times y_{12},$$

where s denotes the applied duration criterion, that is $s \in \{0, 3, 6, 12, p\}$.

The question in the Delphi study about the overcount was introduced after the question concerning the undercount. In the preamble it was pointed out that the undercount did *not play a role in here*. It was formulated as follows:

> [. . .] Consider a European country that uses a 12-month criterion. Now imagine that the six-month criterion is used instead. With this new criterion, more persons are considered migrants compared to the previous criterion.
>
> (a) *By how many per cent do you expect that the level of migration with the SIX (THREE) MONTH criterion is higher than with the twelve (SIX) MONTH criterion? Please provide a range in percentages.*
>
> (b) *Approximately how certain are you that the true value will lie within the range that you provided above?*

The experts were asked to provide lower and upper percentages of the overcount, denoted by P_1 and P_2, as well as their level of certainty about the range (P_1, P_2). The percentage $P > 0$ provided by experts represents the duration overcount in the following way:

$$y_\alpha = (1 + P)y_b, \tag{11}$$

where α denotes a shorter duration criterion than b. The overcount due to using a six-month criterion instead of a twelve-month criterion is captured by $1 + P = \exp(d_3)$, where $d_3 > 0$ is an auxiliary variable, so that $y_6 = \exp(d_3)y_{12}$. Similarly, the overcount of migrants measured using a three-month criterion compared to a six-month criterion is $\exp(d_2)$, $d_2 > 0$, which can be expressed as $y_3 = \exp(d_2)y_6$. Thus the effect of using a

three-month criterion compared to a twelve-month criterion is $y_3 = \exp(d_2 + d_3)y_{12}$. For permanent duration the relevant scaling factor is $y_P = \exp(-d_4)y_{12}$, where $d_4 > 0$. These formulations led to the following constraints imposed on the duration parameters dur_s, $s \in \{0, 3, 6, p\}$:

$$\begin{aligned} dur_0 &= \exp(d_1 + d_2 + d_3), \\ dur_3 &= \exp(d_2 + d_3), \\ dur_6 &= \exp(d_3), \\ dur_p &= \exp(-d_4). \end{aligned} \tag{12}$$

We further assume that each d_l, $l = 1, 2, 3, 4$, follows a log-normal distribution. Then the parameters of each expert-specific density for d_l can be calculated by solving the following set of equations:

$$\begin{cases} \mu + \sigma\, \Phi^{-1}(1/2 + c/2) = \log\log(1 + P_1) \\ \mu - \sigma\, \Phi^{-1}(1/2 + c/2) = \log\log(1 + P_2) \end{cases}, \tag{13}$$

where μ and σ are the expected value and standard deviation respectively of the underlying normal density, c is the elicited certainty level, and Φ^{-1} denotes the inverse cumulative distribution function of the standard normal distribution.

The comparisons of the 'permanent' and twelve-month criterion, as well as the three months with 'no time limit', were elicited from the migration experts during a workshop organised by the authors. This workshop brought together academics and persons responsible for migration data at national and international institutions, including some of the experts from the Delphi study. For elicitation, the same approach and formulation of the questions were used but the number of experts was 24 instead of eleven. Here we present the results only of the original Delphi questionnaire, as it is consistent with the other questions on undercount and accuracy.

4.2.2. Expert Answers and Resulting Prior Densities

The representations of individual expert answers concerning the overcount of migration due to different duration of stay criteria are presented in Figure 7 and Figure 8 for six months versus twelve months and three months versus six months respectively on the

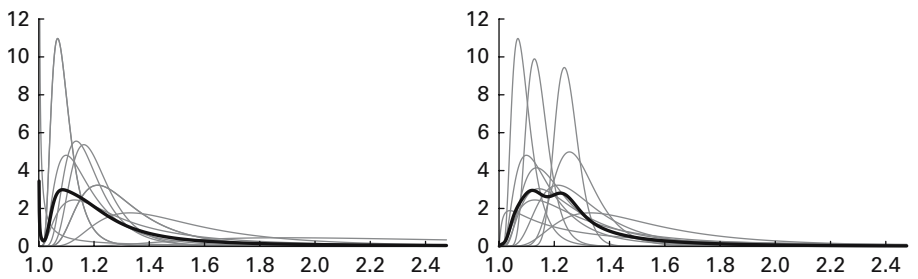

Fig. 7. Expert answers transformed to densities for duration overcount exp(d_3), 6 months versus 12 months, Round 1 (left) and Round 2 (right)

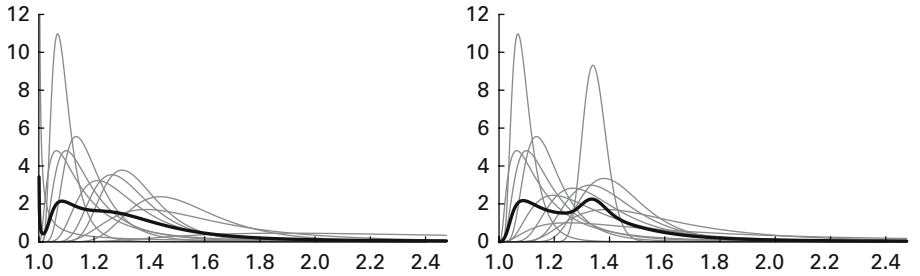

Fig. 8. Expert answers transformed to densities for duration overcount exp(d_2), 3 months versus 6 months, Round 1 (left) and Round 2 (right)

linear scale. In other words, the curves represent the expert answers translated into densities for parameters exp(d_l) and not the overcount factors dur_s.

When we compare the mixture prior densities (bold curves in Figure 7 and Figure 8) resulting from two rounds of questions about the overcount due to different duration criteria, we observe two important changes between Round 1 and Round 2 of the Delphi survey. In both the twelve month to six month and six month to three month comparisons, the expert whose answer contributed to the mode at 0% changed his or her judgement. The mixture is a heavy-tailed distribution because Respondent 3 provided a comparatively small confidence in the answers. Here, the number of migrants captured by the data collection system with six months duration of stay criterion is expected to be 10–30% larger than with the twelve-month criterion. Experts were more uncertain and ambiguous about the difference between the three- and six-month criteria.

4.3. Accuracy

4.3.1. Method for Constructing the Prior Density

The question regarding accuracy of data collection appeared to be the most challenging for the experts to answer. It was asked for in the third section of the Delphi questionnaire. In the preamble to the question, it was explained that accuracy should be assessed assuming there were no biases in the measurement, that is, it was independent from the undercount and duration issues.

> [. . .] Consider a European country with a population register in which there is no systematic bias in the measurement of migration. In this case, we may expect random factors, for instance administrative errors in the processing of the data, to affect the level of migration that is actually measured.
>
> (a) *For EMIGRATION (IMMIGRATION), how probable do you think it is that the published statistics are within an interval from minus 5% to plus 5% compared to the true total level of emigration? (If it helps, think of how often the annual published statistics are within this interval during a period of 100 years). Please provide a range in percentages.*
>
> (b) *Approximately how certain are you that the true value will lie within the range that you provided above?*

The interpretation of the question in brackets was provided to help respondents understand the notion of accuracy and provide a context of the range of minus 5% to plus 5%.

To transform experts' answers into prior densities for the precision of the random terms in the measurement equations, consider a simplified equation for the observed data z and true flows y:

$$z = y \times \xi, \tag{14}$$

where ξ denotes an error term. On the logarithmic scale, ξ is normally distributed with mean zero and precision τ. Given the $\pm 5\%$ deviation from the true level of migration and two probabilities of such an event provided by the experts, P_1 and P_2, it follows that

$$P_i = \Phi[\log(1.05)\sqrt{\tau_i}] - \Phi[\log(0.95)\sqrt{\tau_i}], \quad i = 1, 2. \tag{15}$$

Using the approximation $\log(1.05) \approx -\log(0.95) \approx 0.05$, we simplify the above equation into

$$P_i = 2\Phi(0.05\sqrt{\tau_i}) - 1, \quad i = 1, 2. \tag{16}$$

Then the precision τ_i is computed as

$$\tau_i = 400\left[\Phi^{-1}\left(\frac{P_i + 1}{2}\right)\right]^2, \; i = 1, 2. \tag{17}$$

For expert-specific distribution of τ_i a gamma $\mathcal{G}(\alpha, r)$ density is assumed. Parametrisation of the gamma distribution throughout this article is such that the expected value is α/r and the variance is α/r^2. We can estimate the parameters α and r by solving the following set of equations:

$$\begin{cases} F_g^{-1}(P_1, \alpha, r) = q_1 \\ F_g^{-1}(P_2, \alpha, r) = q_2 \end{cases}, \tag{18}$$

where F_g^{-1} is an inverse cumulative distribution of the gamma distribution. This is achieved by finding the roots of the expression:

$$\sum_{i=1}^{2}\left[F_g^{-1}(P_i, \alpha, r) - q_i\right]^2, \tag{19}$$

where

$$q_1 = \frac{(1-c)P_1}{1 + P_1 - P_2} \quad \text{and} \quad q_2 = \frac{(1-c)(1-P_2)}{1 + P_1 - P_2}$$

For the cases where experts provided zero or 100% probabilities, this formula cannot be used because it has no unique solution. To overcome such answers, we replaced zeros with 0.01% and 100% with 99.99%.

To find starting point values for the optimising algorithm a log-normal approximation was used, with parameters μ and σ calculated as

$$\sigma = \frac{\log(\tau_2) - \log(\tau_1)}{\Phi^{-1}(1 - q_2) - \Phi^{-1}(q_1)} \tag{20}$$

and

$$\mu = \log(\tau_2) - \sigma\,\Phi^{-1}(1 - q_2). \tag{21}$$

Then, the expected value and the variance of the approximating log-normal density were computed as follows:

$$E(\tau) = \exp(\mu + \sigma^2/2)$$
$$\mathrm{Var}(\tau) = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$$

Finally, we solved the basic equations $E(\tau) = \alpha/r$ and $\mathrm{Var}(\tau) = \alpha/r^2$ for α and r to obtain the starting point values.

4.3.2. Expert Answers and Resulting Prior Densities

In Figures 9 and 10, the graphical representations of expert answers for emigration and immigration respectively are shown. For clarity, we present the densities for the expected proportion of observations with less than 5% error, as was requested in the question, rather than the gamma densities for the precision τ. The bold curves represent mixtures of the experts' single densities. In terms of results, we observe that in both Round 1 and Round 2, the experts' answers were diversified. About a third of all experts provided low probabilities suggesting that the measurement of both emigration and immigration is rather poor, while the rest of experts stated that the data collection systems are mostly accurate with probabilities higher than 50%. This heterogeneity could stem from the different backgrounds and experiences with various data collection systems in Europe.

Although experts perceived the measurement of immigration to be more accurate than emigration, their opinions were far from unanimous. For example, one of the experts, having seen the results of Round 1, reduced his or her level of confidence in Round 2. In general, we observed some convergence in opinion for the accuracy of immigration.

5. Importance of Expert Information

As described in Subsection 4.1.3, the elicited prior densities for undercount were varied and uncertain. In our process of assessment, we came to the conclusion that our original specification for the undercount parameters had likely created some confusion amongst the experts related to the difficulty in distinguishing undercount amongst intra-European flows

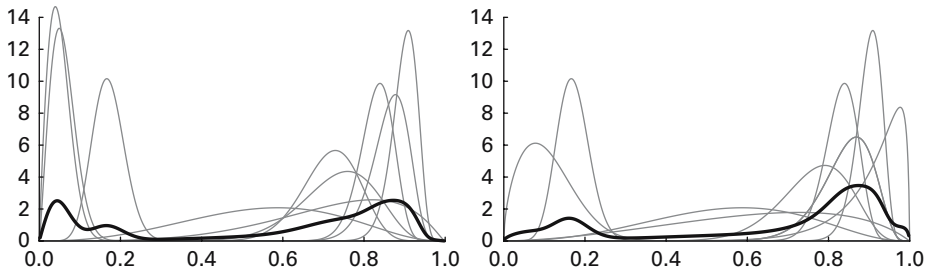

Fig. 9. Expert answers transformed to densities for accuracy of emigration measurement, Round 1 (left) and Round 2 (right)

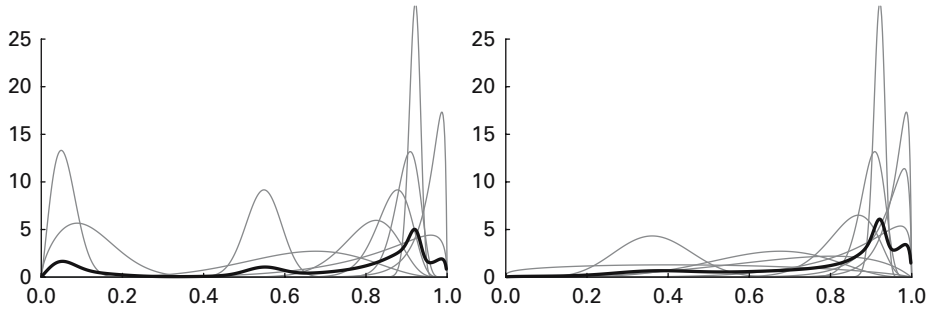

Fig. 10. Expert answers transformed to densities for accuracy of immigration measurement, Round 1 (left) and Round 2 (right)

and flows to and from rest of the world. Moreover, by running the model in (Raymer et al. 2013), we found that the prior densities for undercount led to inflated medians and very wide posterior distributions of the estimated migration flows. This was especially noticeable for countries with reliable population registers, such as Sweden, Norway and the Netherlands.

As a result of our assessment, we considered a different specification for the undercount parameters. Rather than making a distinction between intra-European flows and flows to and from the rest of the world, an expert within our project grouped the countries into two categories: low and high undercount. The opinions for this new specification were also provided by this person. The answers in terms of P_1 and P_2 in Equation (3) were as follows:

- Low undercount countries: The Netherlands, Sweden, Finland, Norway, Denmark, Germany, Iceland, Austria, Belgium, United Kingdom, Cyprus, Ireland, Italy, France, Luxembourg, Switzerland, and immigration to Spain.
  - Emigration: undercount of 20–30% with 60% certainty.
  - Immigration: undercount of 5–15% with 75% certainty.
- High undercount countries: Bulgaria, Estonia, Lithuania, Latvia, Poland, Slovenia, Slovakia, Romania, the Czech Republic, Greece, Hungary, Liechtenstein, Malta, Portugal, and emigration from Spain.
  - Emigration: undercount of 50–60% with 60% certainty.
  - Immigration: undercount of 25–35% with 60% certainty.

This information was then used to construct the prior densities in the same way as described in Subsection 4.1 and resulted in posterior distributions reflecting the assessed differences in the quality of the available data.

We also investigated whether expert opinion on undercount could be removed from the model in two ways. First, we replaced the expert-based prior densities with noninformative uniform prior densities for parameters constrained between zero and one. While we were able to obtain some information concerning the differences between the high category and low category undercount, the level could not be determined purely from the data. Second, we replaced the expert-based prior densities with the noninformative prior densities and assumed all countries had the same level of undercount. In this case, the estimation algorithm did not converge.

The expert-based duration of stay prior densities were examined by keeping the constraints in Equation (10) the same and assuming weakly informative prior densities for the duration parameters in the model described in (Raymer et al. 2013). As it was mentioned in Subsection 4.2, information about the 'no time limit' and 'permanent' criteria was elicited from participants in a workshop organised by the authors. The answers were then transformed into densities following the method outlined in Subsection 4.2. We found that the outcomes were moderately sensitive to the prior densities for the duration of stay parameters. In particular, for the countries with no time limit criterion, the estimated migration flows were lower by only 6–9%, for the three-month criterion, the model with weakly informative prior densities yielded slightly larger estimates (by 4–5%), whereas for the six-month, twelve-month and permanent duration, the differences were smaller than 2%. For individual flows between countries, the differences were seldom larger than $\pm 5\%$, except for countries not providing data for flows from or to the rest of the world. Here, the differences oscillated around $\pm 10-15\%$. Finally, the uncertainty of the flow estimates was unaffected by using weakly informative prior densities.

To assess the sensitivity of the results to the expert-based prior densities for accuracy, we analysed the model in (Raymer et al. 2013) using weakly informative prior densities. The classification of accuracies of the data collection systems in countries remained the same as described in Section 2. In general, this sensitivity analysis showed that the expert-based prior densities, which reflected lack of consensus among experts about accuracy of the data collection, produced nearly the same patterns as when weakly informative prior densities were assumed. This outcome confirms the difficulty of assessing the accuracy of data collection systems.

6. Lessons Learned

As was mentioned in the literature review, elicitation of subjective opinions is a difficult task. Hence, retrospective reflections on the process as well as lessons learned during it can be as valuable as the results themselves. What did this project teach us about elicitation of expert opinion? We mention four points.

First, in our initial analyses of undercount we found that the results are sensitive to the way we specified prior densities, as reported in Section 5. The reason for this problem is not entirely clear. One explanation could be that there is very little information about migration flows to and from Europe, and experts were very uncertain about the undercount, much more so than for intra-European flows. The fact that we found stable results by reformulating the model and distinguishing between two broad categories of countries (rather than distinguishing between intra-European flows and flows to and from the rest of the world) gives some support to this explanation. Therefore, a general lesson is that it may be useful to combine extremely uncertain parameters with ones that are more certain.

Second, the notion of 'undercount of migration flows' expressed as a percentage turned out to have different meanings for different experts. In the first round one of the questions was *By how many per cent do you expect that emigration flows are undercounted in the published statistics, as compared to the true total level of emigration? Please provide a range in percentages*. The idea was that an undercount of 40%, say, reflects a situation

where the published number is 40% lower than the real (unknown) flow. But some experts gave answers that exceeded 100%. We contacted them to verify that their interpretation of an undercount of 200%, say, was as follows: The true flow is three times as large as the reported flow. In Round 2, we improved the wording of the questions on undercount. This example shows that our pilot survey was too limited (two team members and two external experts). Moreover, the testing round could have included various formulations of questions about probabilities (odds, probability, percentage or real example), which would allow us and the experts to check their consistency.

Third, the formulation of questions lacked information about the complement of the range provided by the expert. For the undercount, we did not explain to the experts that the complement of the certainty c, that is $1 - c$, is distributed to the values of the undercount outside the specified interval (but inside the interval [0, 1]). Hence, the probability mass expressed in terms of c lacked context (Gigerenzer 1994; O'Hagan et al. 2006). On the other hand, we did not want to overwhelm the experts with too detailed questions. One option here could have been to ask for a judgement, such as *During last 10 years, how many times did the reported statistics fall into the specified interval?*, rather than confidence. This question would violate the assumption of exchangeability of events (as measurement in a given year is unique) but would provide a context for experts and possibly a clearer interpretation of certainty.

A fourth general lesson is that one should be careful in selecting the experts, in particular when it comes to experience with and knowledge of probabilities and uncertainty. Indeed, we had considerable problems (fortunately in the pilot survey) to convince the experts that subjective probabilities are useful information for our assessment of migration flows. During the first and the second Delphi rounds we were in close contact with two more experts who appeared to be sceptical of the task. Some of these problems might have been avoided had we included in our introductory letter a clear explanation of the two types of uncertainty: epistemic uncertainty (lack of knowledge) and aleatory uncertainty (randomness); see Jenkinson (2005). We should have also emphasised the importance of the explanations and views behind experts' judgements.

7. Conclusion

In situations where data are inconsistent and weak, the inclusion of expert judgements is essential for improving the estimation and for reflecting uncertainty. In our research on modelling migration flows (see Raymer et al. 2013 and http://www.imem.cpc.ac.uk), we sought to provide the best possible estimates and measures of uncertainty based on available data, covariate information and expert judgements. These three pieces of information subsequently can be integrated into a single model for providing harmonised estimates of migration flows amongst 31 countries in the EU and EFTA from 2002 to 2008.

In this article, we have described our methodology for obtaining expert information on migration data to supplement reported flows and covariate information. Our implementation of this methodology was the first attempt at eliciting and quantifying opinions on various aspects of the migration data collection systems. As a result, we obtained a valuable assessment of the data on migration flows. From the varying opinions on the undercount, we can conclude that the data collection systems are expected to

capture about a half of emigrants in Europe and around 60–90% of immigrants. We learned about the likely effects of different duration of stay criteria used to record migration flows, for example, the differences in reported figures between a six-month definition and twelve-month definition. Finally, the largest ambiguity concerns the assessment of the accuracy. The only conclusion that can be drawn in that respect is that the experts expect immigration to be measured with greater precision than emigration.

After two rounds of the Delphi survey, we found that experts often disagreed on the various measurement aspects of migration. The feedback from the first round did not lead to significant changes in their opinions. However, we did not aim at convergence, as this could lead to an artificial reduction of uncertainty. Moreover, we believe that due to the heterogeneity of expert judgements expressed in the survey, the results are an important assessment of the problematic quality of the data collection systems across Europe. Nonetheless, elicitation and quantification of the expert knowledge on the data collection mechanisms in Europe is desired, especially in the context set out by the Regulation (EC) No. 862/2007 of the European Parliament and of the Council of July 11, 2007. According to the Regulation, countries in the EU are required to provide statistics on migration based on the harmonised definition of a migrant to Eurostat. The Regulation allows for use of well-documented scientific estimation and modelling methods to compile statistics on migration. Expert knowledge expressed in terms of probability distributions, as described in this article, can provide an important input to models for harmonising migration data. It also helps to understand the data collection mechanisms applied in Europe and the differences among them, as well as to assess the quality of the data produced.

8. References

Abel, G.J. (2010). Estimation of International Migration Flow Tables in Europe. Journal of the Royal Statistical Society, Series A (Statistics in Society), 173, 797–825. DOI: http://www.dx.doi.org/10.1111/j.1467-985X.2009.00636.x

Bijak, J. and Wiśniowski, A. (2010). Bayesian Forecasting of Immigration to Selected European Countries by Using Expert Knowledge. Journal of the Royal Statistical Society, Series A, 173, 775–796. DOI: http://www.dx.doi.org/10.1111/j.1467.985x.2009.00635.x

Clemen, R.T. and Winkler, R.L. (1990). Unanimity and Compromise Among Probability Forecasters. Management Science, 36, 767–779.

Cooke, R.M. (1991). Experts in Uncertainty: Opinion and Subjective Probability in Science. New York: Oxford University Press.

De Beer, J., Raymer, J., Van der Erf, R., and Van Wissen, L. (2010). Overcoming the Problems of Inconsistent Migration Data: A New Method Applied to Flows in Europe. European Journal of Population, 26, 459–481.

DeWaard, J., Kim, K., and Raymer, J. (2012). Migration Systems in Europe: Evidence from Harmonized Flow Data. Demography, 49, 1307–1333.

Dey, D.K. and Liu, J. (2007). A Quantitative Study of Quantile Based Direct Prior Elicitation from Expert Opinion. Bayesian Analysis, 2, 137–166. DOI: http://www.dx.doi.org/10.1214/07-BA206

Drbohlav, D. (1996). The Probable Future of European East-West International Migration-Selected Aspects. In Central Europe after the Fall of the Iron Curtain; Geopolitical Perspectives, Spatial Patterns and Trends, F.W. Carter, P. Jordan, and V. Rey (eds). Frankfurt: Lang, 269–296.

Garthwaite, P., Kadane, J.B., and O'Hagan, A. (2005). Statistical Methods for Eliciting Proability Distributions. Journal of the American Statistical Association, 100, 680–700. DOI: http://www.dx.doi.org/10.1198/016214505000000105

Gigerenzer, G. (1994). Why the Distinction Between the Single Event Probabilities and Frequencies is Important for Psychology (and vice-versa). In Subjective Probability, G. Wright, P. Ayton (eds). Chichester: John Wiley, 129–161.

Goodwin, P. and Wright, G. (1998). Decision Analysis For Management Judgement (2nd Edition). Chichester: John Wiley.

Jandl, M., Hollomey, C., and Stepien, A. (2007). Migration and Irregular Work in Austria. Results of a Delphi-Study. International Migration Papers 90. International Labour Office; International Centre for Migration Policy Development. Geneva: ILO.

Jenkinson, D. (2005). The Elicitation of Probabilities: A Review of the Statistical Literature. Department of Probability and Statistics, University of Sheffield, Sheffield UK.

Johnson, N.L. (1949). Systems of Frequency Curves Generated by Methods of Translation. Biometrica, 36, 149–176.

Kadane, J.B. and Wolfson, L.J. (1998). Experiences in Elicitation. The Statistician, 47, 3–19. DOI: http://www.dx.doi.org/10.1111/1467-9884.00113

Kupiszewska, D. and Wiśniowski, A. (2009). Availability of Statistical Data on Migration and Migrant Population and Potential Supplementary Sources for Data Estimation. MIMOSA Deliverable 9.1 A Report, Netherlands Interdisciplinary Demographic Institute, The Hague. Available at: http://mimosa.gedap.be/Documents/Mimosa_2009.pdf (accessed November 2012).

Nowok, B. (2010). Harmonization by Simulation: A Contribution to Comparable International Migration Statistics in Europe. Amsterdam: Rozenberg Publishers.

Nowok, B. and Willekens, F. (2011). A Probabilistic Framework for Harmonisation of Migration Statistics. Population, Space and Place, 17, 521–533. DOI: http://www.dx.doi.org/10.1002/psp.624

O'Hagan, A. (1998). Eliciting Expert Beliefs in Substantial Practical Applications. The Statistician, 47, 21–35. DOI: http://www.dx.doi.org/10.1111/1467-9884.00114

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., and Rakow, T. (2006). Uncertain Judgements: Eliciting Experts Probabilities. New York: Wiley.

Poulain, M. (1993). Confrontation des Statistiques de Migrations Intra-Européennes: Vers Plus D'harmonisation? European Journal of Population, 9, 353–381.

Poulain, M. and Dal, L. (2008). Estimation of Flows within the Intra-EU Migration Matrix. Report for the MIMOSA project. Available at: http://mimosa.gedap.be/Documents/Poulain_2008.pdf (accessed November 2012).

Poulain, M., Perrin, N., and Singleton, A. (Eds) (2006). THESIM: Towards Harmonised European Statistics on International Migration. Louvain-la-Neuve: UCL Presses Universitaires de Louvain.

Raymer, J., Wiśniowski, A., Forster, J., Smith, P.W.F., and Bijak, J. (2013). Integrated Modeling of European Migration. Journal of the American Statistical Association, 108, 801–819. DOI: http://www.dx.doi.org/10.1080/01621459.2013.789435

Rowe, G. and Wright, G. (1999). The Delphi Technique as a Forecasting Tool: Issues and Analysis. International Journal of Forecasting, 15, 353–375.

Rowe, G. and Wright, G. (2001). Experts Opinions in Forecasting: The Role of the Delphi Technique. In Principles of Forecasting: A Handbook of Researchers and Practitioners, J.S. Armstrong (ed.). Boston: Kluwer Academic Publications, 125–144.

Szreder, M. and Osiewalski J. (1992). Subjective Probability Distributions in Bayesian Estimation of All-Excess-Demand Models. Discussion paper in Economics, 92-7. University of Leicester, Leicester.

United Nations (1998). Recommendations on Statistics of International Migration. Statistical Papers Series M, No. 58, Revision 1. New York: Department of Economic and Social Affairs, Statistics Division, United Nations.

Van der Erf, R. (2009). Typology of Data and Feasibility Study. MIMOSA Deliverable 9.1 B Report, Netherlands Interdisciplinary Demographic Institute, The Hague.

Van der Erf, R. and Van der Gaag, N. (2007). An Iterative Procedure to Revise Available Data in the Double Entry Matrix for 2002, 2003 and 2004. MIMOSA Discussion Paper, Netherlands Interdisciplinary Demographic Institute, The Hague. Available at: http://mimosa.gedap.be/Documents/Erf_2007.pdf (accessed November 2012).

Wiśniowski, A. and J. Bijak, J. (2009). Elicitation of Expert Knowledge for Migration Forecasts Using a Delphi Survey, CEFMR Working Paper, 2/2009. Warsaw: Central European Forum for Migration and Population Research.

Received July 2012
Revised May 2013
Accepted June 2013

VERSITA

Journal of Official Statistics, Vol. 29, No. 4, 2013, pp. 609–622, http://dx.doi.org/10.2478/jos-2013-0042

Internet Coverage and Coverage Bias in Europe: Developments Across Countries and Over Time

Anja Mohorko[1], *Edith de Leeuw*[2], *and Joop Hox*[2]

To estimate the coverage error for web surveys in Europe over time, we analyzed data from the Eurobarometer. The Eurobarometer collects data for the European Community across member and applicant states. Since 2005, the Eurobarometer has contained a straightforward question on Internet access. We compared respondents with and without Internet access and estimated coverage bias for demographic variables (sex, age, length of education) and sociopolitical variables (left-right position on a political scale, life satisfaction). Countries in Europe do differ in Internet penetration and resulting coverage bias. Over time, Internet penetration dramatically increases and coverage bias decreases, but the rate of change differs across countries. In addition, the countries' development significantly affects the pace of these changes.

Key words: Web survey; Internet; coverage; coverage bias; nonsampling error; Eurobarometer.

1. Introduction

Modern society relies on reliable and valid survey data, and almost every country in the world uses surveys to estimate important statistics, such as rate of unemployment, health indicators, opinions about the government and key issues in society, intention to vote in the coming elections, and people's satisfaction with services. Surveys are also one of the most common methods in the social sciences used to understand the way societies work and to test theories.

The last decennium has been marked by fast-paced technological changes that influence survey methods and survey quality. A dramatic change in survey methodology was caused by the development of Internet surveys (Bosnjak et al. 2006; Couper 2000). Internet surveys have many advantages, such as low costs, timely data, and more privacy due to self-completion. The latter is especially important when sensitive topics are being surveyed, and mode comparisons consistently show that Internet surveys give rise to less social desirability than interviews (e.g., Kreuter et al. 2008; Link and Mokdad 2005; for an

[1] Department of Social Informatics, Faculty of Social Sciences, University of Ljubljana, Kongresni trg 12, 1000, Ljubljana, Slovenia. Email: anja.mohorko@fdv.uni-lj.si
[2] Department of Methodology and Statistics, Utrecht University. Plantage Doklaan 40, 1018 CN Amsterdam, the Netherlands. Email: edithl@xs4all.nl and j.hox@uu.nl
Acknowledgments: The authors thank Bill Blyth for his stimulating ideas and kind support throughout this project. We also thank Mick Couper, two anonymous reviewers, and the editors of JOS for their helpful suggestions.

overview see De Leeuw and Hox 2011). In this sense, Internet surveys are indeed more like self-administered questionnaires and share their benefits, as Couper (2008) postulated.

From the onset of Internet surveys, coverage error has been a source of major concern. A main problem with Internet surveys is under-coverage resulting from the "digital divide", that is, a difference in rates of Internet access among different demographic groups (such as an unequal distribution regarding age and education for those with and without Internet access; see Couper 2000). Although Internet coverage is growing – for instance for Europe as a whole, Internet coverage increased from 15% in December 1999 to approximately 63% in June 2012 (Internet World Stats 2013) – it varies widely across countries. For example, at the beginning of the 21st century almost 15% of Europeans had Internet access, but according to the World Bank (2009) this ranged from less than 4% (e.g., Romania and Turkey) to 44% and 46% (the Netherlands and Sweden). For a more detailed overview, see Blyth (2008). This differential coverage would not be a problem if the covered part represented the general population with respect to important survey variables. However, even in countries with a high coverage a digital divide can be observed, as Internet access is unevenly distributed over the population, with highly educated and younger persons more often having an Internet connection (e.g., Bethlehem and Biffignandi 2012; Rookey et al. 2008; Couper et al. 2007). This differential coverage over countries and demographic groups may result in biased estimates of substantive variables of interest in a study. To estimate the coverage bias, one needs data on both parts of the population, that covered and that not covered.

In terms of coverage of the household population, face-to-face interviews are often viewed as the gold standard to which other modes are compared (e.g., Groves et al. 2009). Since 2005, the Eurobarometer, which is based on face-to-face interviews, contains a question about Internet access at home. This provides us with a unique data set to analyze Internet coverage and coverage bias across European countries and over time. How would substantive results change if important international studies like the Eurobarometer used Internet surveys instead of the (golden) standard face-to-face interviews? As data collection in the Eurobarometer does not depend on respondents having access to the Internet, the survey mode is held constant, and as the same battery of questions is asked over time and across countries, this data set enables us to investigate how potential *coverage* bias could influence the results if the data had been collected using Internet surveys instead of face-to-face interviews. In other words, this gives us an indication of Internet coverage and coverage bias over time and across countries.

In this study, we compare those with access to Internet at home to the whole target group of Eurobarometer face-to-face interviewees (both with and without Internet access at home). It is expected that the coverage bias between the two groups differs between countries and will decrease over time for all countries. We also expect that the rate of decrease may be different in different countries and that social and economic indicators at the country level may explain some of these differences.

In the following sections, we first describe the available data and the analysis methods used. We then present our results on trends in Internet coverage at home and the resulting coverage bias for available demographic variables and sociopolitical variables. This is followed by a multilevel analysis to model the changes over time and the influence of

socioeconomic development on these trends. We end with a critical discussion and implications for research.

2. Method

2.1. Available Data

2.1.1. Eurobarometer

The Eurobarometer collects data for the European Community across EU members and applicant countries four to eight times a year. The Eurobarometer has a separate data collection for East and West Germany, the Republic of Cyprus and the Turkish Republic of Northern Cyprus, and Great Britain and Northern Ireland. Therefore, the following 32 countries were included in the analyses: Austria, Belgium, Bulgaria, Croatia, Cyprus (Republic and TCC), Czech Republic, Denmark, Estonia, Finland, France, Germany (East and West), Great Britain, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Northern Ireland, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, and Turkey. Since 2005, the Eurobarometer contains a yearly question about Internet access at home.

Each wave of the Eurobarometer consists of face-to-face interviews and includes a core questionnaire plus an additional questionnaire with special topics. For each standard Eurobarometer survey, new and independent samples are drawn; since October 1989, the basic sampling design has been a multi-stage probability sample. To ensure the total coverage of each country, the sampling in the first stage is based on a random selection of sampling points (PSU) after stratification by the distribution of the national, resident population in terms of metropolitan, urban, and rural areas, that is, proportional to the population size. Within the PSUs addresses are then selected using random route procedures, followed by a random selection of a person at the address (for more details on sampling and coverage, see GESIS Eurobarometer Survey series 2013).

Every household survey suffers from nonresponse (Bethlehem et al. 2011; De Leeuw and De Heer 2002; Groves and Couper 1998), and the Eurobarometer is no exception. Unfortunately, there is no detailed information on response rates made available publicly and on a regular basis by the principal investigator, the European Commission's Eurobarometer unit. Still, there is some indication that response rates vary between countries. For instance, Busse and Fuchs (2012) note that for the 2002 Eurobarometer, response rates varied between rates of around 70% for East and West Germany and 40% or less for Ireland, Denmark and the UK. No systematic nonresponse studies are available. However, the Eurobarometer data do include integrated design and poststratification weights to adjust the realized samples to EUROSTAT population data (Moschner 2012). These weights will be used in estimating the coverage bias indicators.

The core questionnaire contains trend questions about sociopolitical orientation and standard demographic questions and, since 2005, also includes a question on having an Internet connection at home, allowing us to estimate Internet access at home and the resulting coverage bias. Besides Internet access at home, interview data on the following variables were available for all countries: sex, age, length of education, political left-right self-placement and life satisfaction (see Mohorko et al. 2011 for the question wording

used); also the year of data collection was recorded. All the data were downloaded in February and March 2011, at which point the Eurobarometer data were fully available for the years 2005 to 2009. Hence, our analysis will cover this five-year period. To assess coverage bias, we analyze three demographic variables: sex, age, and length of education, and two substantive variables: political left-right self-placement and life satisfaction. The demographic variables age, sex, and education are seen as important indicators for the digital divide (e.g., Couper 2000) and correlate with many substantive variables typically assessed in academic or market research surveys (Fuchs and Busse 2009). The substantive variables political left-right self-placement and life satisfaction give us an opportunity to directly investigate the influence of undercoverage on the assessment of two major socio-political indicators.

2.1.2. Additional Country-level Variables

The data from the Eurobarometer are individual level data, collected through face-to-face interviews in each country. Apart from Internet penetration, the countries involved in the Eurobarometer also differ on socioeconomic variables, which may influence Internet coverage. To model this, we collected socioeconomic country-level data from Eurostat, the World Bank, and the Human Development Report. Contextual country-level variables are: life expectancy at birth (in years), country's educational index, duration of primary and secondary education (in years), and urbanization (the percentage of urban population). Economic indices on country level are the percentage of employed (labor force), the Gini coefficient (which measures income inequality), Gross Domestic Product growth (GDP), and inflation. For a description of these variables and the data sources including the URL, see Mohorko et al. (2011, 2013). It should be noted that these variables are measured at the country level, but they are available for each year, hence they are time-varying predictors.

2.2. Analysis

2.2.1. Coverage and Indicators of Coverage Bias

Coverage is defined as the percentage of the population of interest that is included in the sampling frame; ideally the coverage should be 100%. Furthermore, there should be a one-to-one correspondence between the population of interest or target population and the (sampling) frame population. If this is not the case, *and* if those missing in the frame differ from the target population on a key variable of interest in the study, coverage error occurs (Biemer and Lyberg 2003; Groves et al. 2009). Groves (1989, p. 11) describes coverage error as follows: "Coverage error exists because some persons are not part of the list or frame (or equivalent materials) used to identify members of the population. Because of this they *never* can be measured whether a complete census of the frame is attempted or a sample studied."

Undercoverage is one of the main concerns for the validity of conclusions based on Internet surveys (Couper 2000). Although Internet access is growing, there are still many individuals who are not covered, and if those without Internet access differ on key measures from those with Internet access, the resulting estimators will be biased. For example, if wealthier households are more likely to have Internet access, then a survey

about household assets that is based exclusively on the Internet will produce income estimates that are too high (Lohr 2008).

To investigate *coverage* problems in Internet-based surveys, we compare the responses of the subgroup of Internet-at-home with those of the total group of Eurobarometer respondents. Since the Eurobarometer was conducted face-to-face in all countries and face-to-face surveys have the least coverage problems (Groves et al. 2009, p. 163; De Leeuw 2008, p. 125), the total Eurobarometer group in this study is regarded as a proxy for the target population. Differences between those with an Internet connection at home and the total Eurobarometer group give an indication of the bias due to *undercoverage* if an Internet survey had been implemented instead of a face-to-face survey.

The net coverage bias is defined by Lessler and Kalsbeek (1992, p. 59–60) as

$$\bar{y}_{\text{covered}} - \bar{y}_{\text{target}} = \frac{N_{\text{not covered}}}{N_{\text{target}}}(\bar{y}_{\text{covered}} - \bar{y}_{\text{not covered}}) \tag{1}$$

which is used by Bethlehem and Biffignandi (2012, p. 289) to define bias due to the non-Internet population. Based on Equation (1), we use two indices to assess the amount of coverage bias: the relative bias (Lessler and Kalsbeek 1992, p. 60) and the absolute relative bias (Groves and Peytcheva 2008). The relative coverage bias is used for descriptive purposes, as the sign of this estimate indicates the over- or undercoverage of specific groups (e.g., if more men than women have Internet access at home in a certain year and in a certain country). However, when modeling changes occur over time and across countries, positive and negative values for relative coverage can cancel each other out and the resulting regression coefficients may falsely give the impression that the overall coverage error is close to zero. Therefore, we use the absolute relative coverage bias in our multilevel analyses.

The relative and absolute relative coverage bias due to lack of Internet access are defined as

$$\text{relative coverage bias} = \frac{\bar{y}_{Int} - \bar{y}_{EB}}{\bar{y}_{EB}} \tag{2}$$

and

$$\text{absolute relative coverage bias} = \left|\frac{\bar{y}_{Int} - \bar{y}_{EB}}{\bar{y}_{EB}}\right| \tag{3}$$

where *EB* represents the total achieved Eurobarometer sample, which is viewed as our target population, and *Int* represents the covered Internet subsample. Analogous $\bar{y}_{EB}$ and $\bar{y}_{Int}$ represent the means of the Eurobarometer target population and the Internet subsample on the variable y.

2.2.2. Statistical Analyses

The relative coverage bias is used for descriptive analyses over countries and time. Positive values indicate that surveys, which are exclusively conducted through the Internet, will result in estimates that are too high, whereas negative values indicate that these will result in estimates that are too low.

Multilevel analysis on the absolute relative coverage bias is used to model and explain trends over time and country for all bias indicators (sex, age, length of education, political left-right self-placement and life satisfaction). For ease of interpretation, the absolute relative coverage bias is expressed as percentage points. In the multilevel model, the lowest level represents the years, indicated by a time variable coded 2005 = 0, 2006 = 1, et cetera. To estimate change over time, we analyze a null model that always includes the linear effect of time and tests whether the variance component for the slope of time is significant. If this random component is not significant using a likelihood ratio test, it is removed from the null model. Since the plots for the effect of time in Figure 1 indicate possible nonlinearity, we test for nonlinear effects by analyzing the quadratic effect of time. If the quadratic term is not significant at the conventional 5% level, it is removed from this model; the linear term for time is always retained in the null model.

In a second step, we add country-level socioeconomic variables. Country-level variables model initial differences in bias between countries in the starting year 2005. Since the country-level variables vary across time, they may also explain change over time. Because the country-level variables are correlated with time, adding them to the model may replace (part of) the explanatory power of the time variable as estimated in the null model.

Finally, differences between countries in the rate of change over the years, as indicated by variation in the slopes of the time variable, are modeled as interactions of country-level

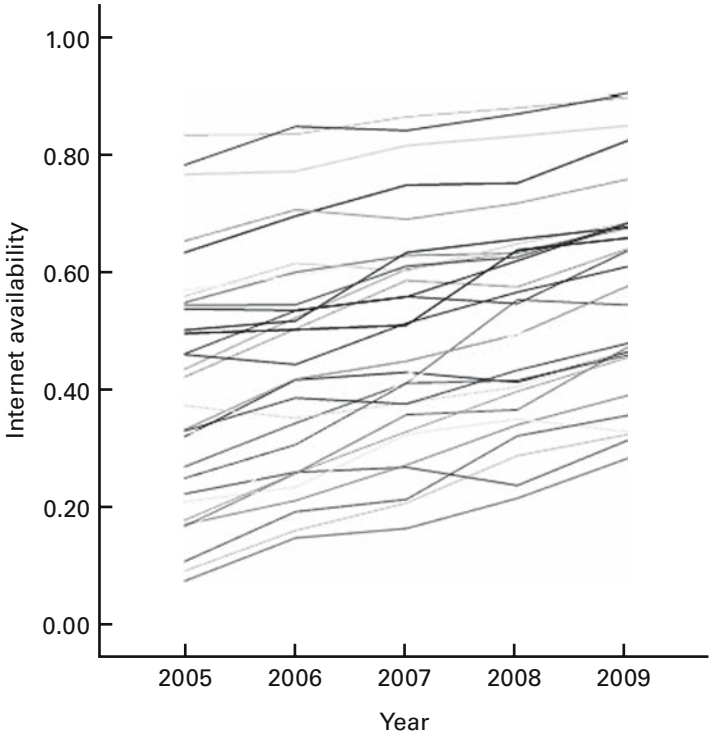

Fig. 1. Internet access at home across Europe 2005–2009, based on the Eurobarometer's weighted data. The lines represent the 32 countries/regions distinguished in the Eurobarometer

variables with the time variable. Again, effects that are not significant are removed from the model. A two-sided significance level of alpha = 0.05 is used throughout.

3. Results

3.1. Coverage Bias in European Countries

Internet access at home increases over time across Europe, but the rate of increase differs across countries (see Figure 1). The actual proportions per country and per year are presented in Appendix A. These numbers show that for countries with an initial low Internet penetration, for example Bulgaria and Romania, the proportions increase rapidly, while for countries with an initial high penetration, for example Sweden and the Netherlands, the growth is less steep.

But even with an Internet penetration above 80%, there still may be considerable differences between those with and without Internet access. This is indicated by the relative coverage bias, which is based on the standardized difference between the subgroup of those who do have Internet at home compared to the total (Internet and non-Internet at home) group. Full descriptive tables with the values of the relative coverage bias for each country in the Eurobarometer and each year are available in Mohorko et al. (2011).

For the demographic variables sex, age, and length of education, the descriptive tables indicate a digital divide. In Europe, those with Internet at home are more often male, younger, and highly educated (Mohorko et al. 2011, Appendix D, Tables D1-D3); similar patterns have been found in the USA (cf. Couper 2008). The bias for sex is relatively low and decreases strongly over time. The highest value was found for Greece with 8.5% more men than women having Internet access in 2006, which decreased to 5.5% in 2009. The lowest values (less than 1% more men) were found for countries like Sweden, Slovenia, Ireland, and the Netherlands in 2009. In general, the gender gap is closing very fast over time. Furthermore, the age difference is becoming smaller over time; younger people are still overrepresented, but for some countries (e.g., Sweden and the Netherlands) the age bias is really low (around −0.04) in 2009, while for others (e.g., Bulgaria) it is still rather high (−0.22 in 2009). The same can be seen for length of education. It should be noted that countries with the smallest digital divide regarding the demographics of age, sex, and education are also the countries with the highest Internet penetration. This gives an optimistic outlook for the future that as Internet penetration increases, the digital divide will decrease.

When we take a closer look at the descriptive tables for the substantive variables political left-right self-placement and life satisfaction (for the detailed tables per country over the years, see Mohorko et al. 2011, Table D4 and D5), we again note that the differences are becoming smaller over time. On average, the coverage bias is very low for political left-right self-placement, where its bias decreases towards zero over time with the largest differences found in Bulgaria (from 0.23 in 2005 to 0.075 in 2009). It should be noted that the coverage bias for this variable does not take the same direction in all countries. For some countries, those with an Internet connection at home place themselves more on the left (e.g., Austria, West-Germany, Great Britain), for other countries they place themselves more on the right (e.g., Bulgaria). For the second substantive variable life

satisfaction, we see that in every country and every year there is a positive bias, indicating that those with Internet at home are more satisfied with life than the Eurobarometer population in general. This bias decreases slightly over time.

3.2. Changes in Coverage Bias Over Time

The change in coverage bias over time is analyzed using multilevel analysis, with years (coded 2005 = 0, . . . , 2009 = 4) nested within countries. This allows us to test whether the change over time is significant and to test if country-level variables can predict changes over time. The analysis showed that the effect of time squared was never significant, and therefore only the linear trend of time is included in the model. Table 1 presents the parameter estimates for each dependent variable for two models: a model with only the linear time indicator and a model with the time indicator and the significant country variables.

When we examine the effect of time in the first model, the results show a steady decrease in absolute relative coverage bias across time, as indicated by a negative value for the regression coefficient of time, except for political left-right self-placement where the overall effect of time is not significant. For all five bias indicators, Table 1 shows a significant and sometimes large country-level variance, which means that there were clear differences in overall bias between countries in 2005. For three out of five bias indicators, the time variable has a significant slope variation (indicated in Table 1 under "time slope variance"), which means that the biases for "age", "political left-right self-placement" and "life satisfaction" decrease at different rates across countries. Compared to the size of the regression coefficient for the time variable itself, these variances are relatively large. This indicates large differences in the rate of decline between countries for these bias indicators.

3.3. Coverage Bias and Country Differences

There are differences between countries in the size of the coverage bias and, for some variables, in the rate of the decrease of this bias over time. These differences are modeled by the direct effects of the available country-level variables: life expectancy, educational index, duration of primary and secondary education, urbanization, employment, Gini index, GDP growth rate, and inflation. The differences in rate of decrease are modeled by the interactions of these variables with the time indicator.

The explanatory variables secondary education, GDP growth rate, and inflation were never significant and are omitted from the model. Table 1 shows the estimated multilevel model and the significant regression coefficients for each of the five coverage bias indicators. The bias for political left-right placement could not be predicted by any of the available country variables. The other four coverage bias indicators can be predicted by different subsets of country-level variables. Thus differences between persons with and without Internet across countries can be predicted using different country-level variables.

Table 1 shows that coverage bias for age is higher in countries with a high income inequality as indicated by the Gini-coefficient, while coverage bias for age is lower in countries with a higher educational index, a higher life expectancy, longer duration of primary school education, and high urbanicity. In contrast, coverage bias for sex is only

Table 1. Absolute relative coverage bias for selected variables predicted by year and country-level variables. Multilevel model with regression coefficients (b), variance components (s^2), and corresponding standard errors (se)

	Bias Sex Composition		Bias Age Composition		Bias Length of Education		Bias Political Left-Right Self-Placement		Bias Life Satisfaction	
Model	Year (2005–09)	Country predictors	Year (2005–09)	Country predictors	Year (2005–09)	Country predictors	Year (2005–09)	Country predictors	Year (2005–09)	Country predictors
Fixed part	b (se)	b (se)	b (se)	b (se)	b (se)	b (se)	b (se)	b (se)	b (se)	b (se)
Intercept	2.87 (.26)	0.10 (1.34)	16.10 (.83)	82.69 (16.07)	10.11 (.97)	62.91 (12.03)	2.83 (.74)	3.02 (.41)	5.55 (.92)	84.60 (12.46)
Time	−0.27 (.08)	−0.27 (.08)	−0.60 (.11)	−0.34 (.11)	−0.78 (.11)	−0.69 (.17)	0.05 (.16)^{ns}	–	−0.35 (.13)	–
Country variables										
Gini coefficient		0.09 (.04)		0.23 (.08)		–		–		–
Educational index		–		−0.21 (.08)		−0.30 (.12)		–		–
Employment		–		–		−3.82 (1.58)		–		−4.26 (1.23)
Life expectancy		–		−0.59 (.19)		–		–		−0.66 (.17)
Primary school dur.		–		−0.92 (.44)		–		–		–
Urbanicity		–		−0.10 (.05)		−0.15 (.06)		–		−0.1 (.04)
Random part	s^2 (se)	s^2 (se)	s^2 (se)	s^2 (se)	s^2 (se)	s^2 (se)	s^2 (se)	s^2 (se)	s^2 (se)	s^2 (se)
Residual variance	1.96 (.25)	1.96 (.25)	2.13 (.31)	2.45 (.32)	3.98 (.50)	1.84 (.33)	2.80 (.40)	2.80 (.40)	2.77 (.40)	3.57 (.53)
Country variance	0.98 (.35)	0.80 (.32)	20.98 (5.66)	7.64 (2.26)	27.91 (7.29)	22.69 (6.68)	15.64 (4.40)	15.30 (4.26)	25.39 (6.89)	6.67 (2.15)
Time slope variance	–	–	0.15 (.09)	–	–	0.45 (.23)	0.51 (.21)	0.49 (.20)	0.26 (.15)	–

Note: the explanatory variables secondary education duration, GDP-growth and inflation had no significant effects, and are omitted.
"–" indicates parameter tested but removed because parameter was not significant at a two-sided alpha of 0.05.
^{ns} indicates non-significant coefficient for time.

associated with the Gini coefficient; coverage bias for sex is higher in countries with high income inequality (high Gini). Coverage bias in length of education is lower in countries with a higher educational index, a higher employment level, and a higher urbanicity. Coverage bias in life satisfaction is lower in countries with a higher employment rate, higher life expectancy and high urbanicity.

There were no significant interactions with time, meaning that the available country-level variables do not predict the differences in the rate of bias decrease. When we compare the model with country variables added to the model with only time as predictor, an interesting pattern emerges. For all four bias indicators with a significant effect of time, Table 1 shows that adding country-level variables to the model decreases the size of both the regression coefficient for time and the variance across countries. Thus part of the effect of time is the result of changes over time in country-level variables. The signs of the regression coefficients for the country variables suggest that, in general, coverage bias decreases when education, employment, life expectancy, and urbanicity increase. In other words, differences between persons with and without Internet access decrease when the value of these variables increase. In contrast, the differences between persons with and without Internet access increase when the income distribution is more unequal.

4. Conclusion and Discussion

As expected, Internet penetration has increased over time in all countries included in this study. As a result, the absolute relative bias in the estimates of four out of five variables has also decreased; only political left-right self-placement does not show this trend. In other words, differences in age, sex, education, and life satisfaction between those with and without Internet access are diminishing. Multilevel analyses show that for those four bias indicators, the decrease in coverage bias over time differs across countries and that the countries' development affects the pace of this decrease. For age and life satisfaction, the variation in decrease is fully explained by the country-level variables in the model, albeit only partially for sex and education.

The general trend is that higher levels of economic development, education, and health are associated with lower coverage bias, whereas higher income inequality is associated with higher levels of bias. Given the general economic and demographic trends, one conclusion of our study is that coverage bias due to low Internet penetration is disappearing across countries in Europe. The multilevel analyses also show variation across countries in both the initial level and rate of decrease of coverage biases for demographic variables. This shows that the "digital divide" (Couper 2000) not only differs between countries, but also is diminishing at different rates over time in these countries.

Our measure of Internet penetration and coverage bias is based on a question in the Eurobarometer that inquires specifically about Internet access at home. However, there are alternative ways to access the Internet, for instance at work, in libraries, or on mobile devices. For this reason, our analyses are based on the assumption that for surveys that consist of more than a couple of pop-up questions, respondents will prefer to answer in an environment where they have time, feel comfortable, and have privacy. Although mobile Internet is promising, only one third of the population was covered by mobile Internet in Europe in 2007. Furthermore, coverage biases for demographic variables for the mobile

web were larger than for landline Internet (Fuchs and Busse 2009). The use of mobile Internet on telephones and tablet devices is likely to increase further in the near future, which will necessitate a change in the measurement of Internet access. Provided that survey methodologists adapt their surveys to these new devices (e.g., Callegaro 2010), this will not change our conclusion that coverage bias for Internet surveys is decreasing over time.

This study focuses on coverage bias. Good coverage is a necessary but not a sufficient condition for high quality survey data. Other error sources exist, such as nonresponse error or mode effects. Meta-analyses (Cook et al. 2000; Lozar Manfreda et al. 2008) show that Internet surveys yield on average 11% lower response rate than other modes. Clearly, measures should be taken to increase this response rate. For a discussion of such measures we refer to Dillman et al. (2009). Compared to face-to-face interviews, responses to Internet surveys may differ due to mode effects, especially when sensitive topics are addressed. For a discussion, we refer to De Leeuw and Hox (2011), Dillman et al. (2009), and Kreuter et al. (2008).

In our study we treat the data from the face-to-face Eurobarometer samples as a representative sample of the total target population, and our results are conditional on the selection and nonresponse processes in the Eurobarometer. Therefore, in estimating the bias indicators, we used the design and post-stratification weights included in the Eurobarometer data. Nevertheless, nonresponse in the Eurobarometer samples can still affect our results. The use of adjustment weights amounts to treating nonresponse as missing at random (MAR, cf. Little and Rubin 2002). However, if the nonresponse in the Eurobarometer were related to Internet access itself (and were therefore missing not at random or MNAR), there is a potential for nonresponse bias. Hence we view our findings as an indication of a generally decreasing coverage bias in the countries studied, but not as precise estimates of this bias.

A potential alternative data source for a future follow-up study would be the European Social Survey (ESS), which recently added a question on Internet access to the core module. Like all surveys, the ESS also has differential nonresponse across countries, but the ESS response rates and sources of nonresponse are well documented and available for more in-depth analyses (Stoop et al. 2010). Ideally, in some countries it may be possible to validate survey-based information on Internet access with registry data.

In conclusion, even if Internet coverage is not complete, Internet surveys may still compete with other survey modes. For instance, in 2008 the Netherlands had an 86% Internet coverage, while the landline telephone coverage was around 60–70% (Bethlehem et al. 2011, p. 100 and p. 102). The same trend can be seen in other countries; for instance, Smyth and Pearson (2011, p. 16 and p. 17) report that in 2008 the US had an Internet coverage of just over 70%, and random digit dialing landline telephones had a coverage of about 78%. However, landline telephone coverage is decreasing (cf. Busse and Fuchs 2012; Mohorko et al. 2013), while Internet coverage is rapidly increasing over time – as this study shows.

Appendix A:
Growth of internet access at home across Europe: 2005–2009 based on the Eurobarometer weighted data for that time period

Country\Year	2005	2006	2007	2008	2009	Grand Total
Austria	0.46	0.53	0.56	0.55	0.56	0.53
Belgium	0.55	0.60	0.63	0.63	0.62	0.61
Bulgaria	0.09	0.16	0.21	0.29	0.32	0.21
Croatia	0.32	0.42	0.43	0.41	0.41	0.40
Cyprus Rep.	0.33	0.37	0.43	0.48	0.48	0.42
Cyprus (TCC)	0.27	0.34	0.41	0.55	0.53	0.42
Czech Rep.	0.33	0.42	0.45	0.49	0.53	0.44
Denmark	0.77	0.77	0.81	0.83	0.84	0.80
Estonia	0.44	0.52	0.61	0.63	0.62	0.56
Finland	0.63	0.69	0.75	0.75	0.79	0.72
France	0.42	0.50	0.59	0.57	0.62	0.54
Germany East	0.46	0.44	0.51	0.57	0.61	0.52
Germany West	0.56	0.62	0.60	0.65	0.69	0.62
Great Britain	0.57	0.59	0.63	0.63	0.65	0.61
Greece	0.22	0.26	0.27	0.24	0.30	0.26
Hungary	0.17	0.21	0.27	0.34	0.37	0.27
Ireland	0.50	0.50	0.51	0.64	0.65	0.56
Italy	0.37	0.35	0.38	0.41	0.39	0.38
Latvia	0.16	0.30	0.42	0.50	0.50	0.38
Lithuania	0.17	0.26	0.36	0.37	0.38	0.31
Luxembourg	0.65	0.71	0.69	0.72	0.76	0.71
Malta	0.50	0.52	0.63	0.66	0.66	0.59
Northern Ireland	0.54	0.54	0.56	0.62	0.65	0.58
Poland	0.25	0.31	0.41	0.42	0.48	0.37
Portugal	0.21	0.24	0.33	0.35	0.33	0.29
Romania	0.11	0.19	0.22	0.32	0.33	0.23
Slovakia	0.18	0.26	0.33	0.40	0.43	0.32
Slovenia	0.54	0.54	0.61	0.63	0.64	0.59
Spain	0.33	0.39	0.38	0.43	0.41	0.39
Sweden	0.78	0.85	0.84	0.87	0.88	0.84
The Netherlands	0.83	0.83	0.86	0.88	0.90	0.86
Turkey	0.08	0.15	0.17	0.22	0.25	0.17
Grand Total	0.40	0.45	0.50	0.53	0.55	0.49

5. References

Bethlehem, J. and Biffignandi, S. (2012). Handbook of Web Surveys. Hoboken, NJ: Wiley.

Bethlehem, J., Cobben, F., and Schouten, B. (2011). Handbook of Nonresponse in Household Surveys. New York: Wiley, Wiley Series in Survey Methodology.

Biemer, P.P. and Lyberg, L.E. (2003). Introduction to Survey Quality. New York: Wiley, Wiley Series in Survey Methodology.

Blyth, B. (2008). Mixed-Mode: The Only "Fitness" Regime? International Journal of Market Research, 50, 241–266.

Bosnjak, M., Forsman, G., Isaksson, A., Lozar Manfreda, K., Schonlau, M., and Tuten, T. (2006). Preface to JOS Special Issue on Web Surveys. Journal of Official Statistics, 22, iii.

Busse, B. and Fuchs, M. (2012). The Components of Landline Telephone Survey Coverage Bias. The Relative Importance of No-Phone and Mobile-Only Populations. Quality and Quantity, 46, 1209–1225. DOI: http://www.dx.doi.org/10.1007/s11135-011-9431-3

Callegaro, M. (2010). Do You Know Which Device Your Respondent Has Used to Take Your Online Survey? Survey Practice, December: www.surveypractice.org. Available at http://surveypractice.org/2010/12/08/device-respondent-has-used/ (accessed August 2012).

Cook, C., Heath, F., and Thompson, R.L. (2000). A Meta-Analysis of Response Rates in Web- and Internet-Based Surveys. Educational and Psychological Measurement, 60, 821–836. DOI: http://www.dx.doi.org/10.1177/00131640021970934

Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches. Public Opinion Quarterly, 64, 464–494. DOI: http://www.dx.doi.org/10.1086/318641

Couper, M.P. (2008). Designing Effective Web Surveys. New York: Cambridge University Press.

Couper, M.P., Kapteyn, A., Schonlau, M., and Winter, J. (2007). Noncoverage and Nonresponse in an Internet Survey. Social Science Research, 36, 131–148. DOI: http://www.dx.doi.org/10.1016/j.ssresearch.2005.10.002

De Leeuw, E.D. (2008). Choosing the Method of Data Collection. International Handbook of Survey Methodology, E.D. de Leeuw, J.J. Hox, and D.A. Dillman (eds). New York: Routledge, Taylor & Francis, European Association of Methodology (EAM) Methodology Series.

De Leeuw, E.D. and De Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In Survey Nonresponse, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: Wiley, Wiley Series in Survey Methodology.

De Leeuw, E.D. and Hox, J.J. (2011). Internet Surveys as Part of a Mixed Mode Design. Social Research and the Internet. In Advances in Applied Methods and New Research Strategies, M. Das, P. Ester, and L. Kaczmirek (eds). New York: Routledge, Taylor & Francis, European Association of Methodology (EAM) Methodology Series.

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2009). Internet, Mail, and Mixed-Mode Surveys; The Tailored Design Method. New York: Wiley, Wiley Series in Survey Methodology.

Fuchs, M. and Busse, B. (2009). The Coverage Bias of Mobile Web Surveys Across European Countries. International Journal of Internet Science, 4, 21–33. Available at: http://www.ijis.net/ijis4_1/ijis4_1_fuchs_pre.html (accessed July 2012).

GESIS, Eurobarometer Survey Series (2013). Available at: http://www.gesis.org/eurobarometer-data-service/survey-series (accessed July 2013).

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: Wiley, Wiley Series in Survey Methodology.

Groves, R.M. and Couper, M.P. (1998). Nonresponse in Household Interview Surveys. New York: Wiley, Wiley Series in Survey Methodology.

Groves, R.M. and Peytcheva, E. (2008). The Impact of Nonresponse Rates on Nonresponse Bias – A Meta-Analysis. Public Opinion Quarterly, 72, 167–189. DOI: http://www.dx.doi.org/10.1093/poq/nfn011

Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2009). Survey Methodology. New York: Wiley, Wiley Series in Survey Methodology.

Internet World Stats. (2013). Available at: http://www.internetworldstats.com/stats.htm (accessed June 2013).

Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys. The Effect of Mode and Question Sensitivity. Public Opinion Quarterly, 72(5), 847–865. DOI: http://www.dx.doi.org/10.1093/poq/nfn063

Lessler, J.T. and Kalsbeek, W.D. (1992). Nonsampling Error in Surveys. New York: Wiley.

Link, M.W., and Mokdad, A.H. (2005). Effects of Survey Mode on Self-Reports of Adult Alcohol Consumption: A Comparison of Mail, Web, and Telephone Approaches. Journal of Studies on Alcohol, March 2005, 239–245.

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: Wiley.

Lohr, S.L. (2008). Coverage and Sampling. In International Handbook of Survey Methodology, E.D. de Leeuw, J.J. Hox, and D.A. Dillman (eds). New York: Routledge, Taylor & Francis, European Association of Methodology (EAM) Methodology Series.

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., and Vehovar, V. (2008). Web Surveys versus Other Survey Modes – A Meta-Analysis Comparing Response Rates. International Journal of Marketing Research, 50, 79–104.

Mohorko, A., De Leeuw, E., and Hox, J. (2011). Internet Coverage and Coverage Bias in Countries Across Europe and over Time: Background, Methods, Question Wording and Bias Tables. Available at: www.joophox.net (accessed June 2013).

Mohorko, A., De Leeuw, E., and Hox, J. (2013). Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage across Countries and over Time. Survey Methods: Insights from the Field. Available at: http://surveyinsights.org/?p = 828 (accessed February 2013).

Moschner, M. (2012). GESIS, Weighting overview. Available at: http://www.gesis.org/eurobarometer-data-service/survey-series/candidate-countries-eb/weighting-overview/ (accessed February 2013).

Rookey, B.D., Hanway, S., and Dillman, D.A. (2008) Does a Probability-based Household Panel Benefit from Assignment to Postal Response as an Alternative to Internet-only? Public Opinion Quarterly, 72, 962–984. DOI: http://www.dx.doi.org/10.1093/poq/nfn061

Smyth, J.D. and Pearson, J.E. (2011). Internet Survey Methods: A Review of Strengths, Weaknesses, and Innovations. In Advances in Applied Methods and New Research Strategies, M. Das, P. Ester, and L. Kaczmirek (eds). New York: Routledge, Taylor & Francis, European Association of Methodology (EAM) Methodology Series.

Stoop, I., Billiet, J., Koch, A., and Fitzgerald, R. (2010). Improving Survey Response. Lessons Learned from the European Social Survey. New York: Wiley.

World Bank (2009). Available at: http://data.worldbank.org/ (accessed May 2011).

Received December 2011
Revised March 2013
Accepted August 2013