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