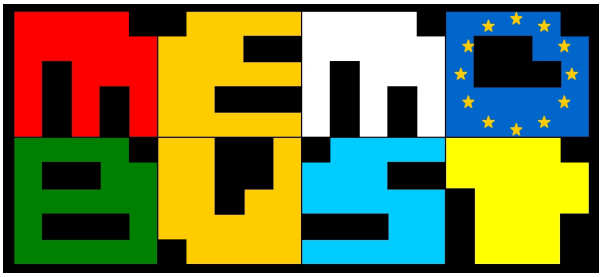

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Statistical Data Editing – Main Module

Contents

General section

1. Summary

Data that have been collected by a statistical institute inevitably contain errors. In order to produce statistical output of sufficient quality, it is important to detect and treat these errors, at least insofar as they have an appreciable influence on publication figures. For this reason, statistical institutes carry out an extensive process of checking the data and performing amendments. This process of improving the data quality for statistical purposes, by detecting and treating errors, is referred to as statistical data editing.

2. General description

2.1 Introduction to statistical data editing

Errors are virtually always present in the data files used by producers of statistics. This is true for both data obtained by means of surveys and data originating from external registers. Insofar as these errors result in inaccurate estimates of publication figures, it is important for statistical institutes to detect and treat these errors.

Errors can arise during the measurement process; if this is the case, there will be a difference between the reported value and the actual value. This can occur because the respondent does not know the actual value exactly or at all, or has difficulty finding this value and therefore makes an estimate. Another possible cause is a difference in definitions between the accounting records of businesses and the statistical institute, for example because the financial year differs from the calendar year. Furthermore, it is possible that businesses simply do not have all the information requested by the statistical institute on file. In this case, the respondent will again estimate certain values or not answer all questions. Finally, respondents may also read or understand questions incorrectly. For example, they may report in euros, while they were actually asked to report in thousands of euros (this is an example of a so-called *unit of measurement error*).

Errors may also arise during data processing. At a statistical institute, the collected data typically go through different processes, such as entering, coding, detection, imputation, weighting, and tabulation. All of these processes can introduce errors into the data. An example of this is that the manual entry of data can result in misinterpretations, for example, a ‘1’ is taken for a ‘7’ or vice versa. Similar mistakes can occur when optical character recognition is used to process survey forms automatically. Additionally, there may be errors in the processing software, and good values may incorrectly be seen as errors during the editing process.

The process of detecting and treating errors in a data file to be used for statistical purposes is called *statistical data editing*. Other commonly used terms are *data validation* and *data cleaning*. In traditional survey processing, data editing was mainly a manual activity, intended to check and correct all data items in every detail. Inconsistencies in the data were investigated and, if necessary, adjusted by subject-matter experts, who would consult the original questionnaires or recontact respondents to verify suspicious values. Overall, this was a very time-consuming and labour-intensive procedure. According to estimates in the literature, statistical institutes would spend up to 25% or 40% of their total budget on data editing (Federal Committee on Statistical Methodology, 1990; Granquist, 1995; Granquist and Kovar, 1997).

According to Granquist (1997), statistical data editing should have the following objectives, in descending order of priority:

1. To identify possible sources of errors so that the statistical process can be improved in the future;
2. To provide information about the quality of the data collected and published;
3. To detect and correct influential errors in the collected data.

In EDIMBUS (2007), a fourth objective is added:

4. If necessary, to provide complete and consistent microdata.

In line with the first objective mentioned above, the main aim of recontacts with respondents should not be to merely resolve individual observed errors, but rather to collect information on the causes of these errors. By collecting and analysing this information, a statistical institute has the opportunity to identify potential measures for improving the quality of incoming data in the future. Examples of such measures include improving the design of the questionnaire and, in particular, changing the wording of a question that many respondents found difficult to answer. In the words of Granquist (1997), "editing should highlight, not conceal, serious problems in the survey vehicle."

Currently at most statistical institutes, statistical data editing is used primarily with the third and fourth of the above goals in mind: correcting errors that have a significant influence on publication totals and providing complete and consistent data. Although it is widely acknowledged in the data editing literature that the information obtained during editing could and should also be used to improve aspects of the statistical process for a repeated survey, the development of practices to achieve this goal still appears to be a rather neglected area. Some statistical institutes have had good experiences with standardised debriefings of editing staff as a device for identifying possible improvements in questionnaire design (Rowlands et al., 2002; Hartwig, 2009; Svensson, 2012). An overview of indicators for assessing the quality of the data before and after editing is given in EDIMBUS (2007).

Over the past decades, statistical institutes have recognised that it is usually not necessary to correct all data in every detail. Several studies have shown that reliable estimates of publication totals can also be obtained without removing all errors from a data set (see, e.g., Granquist, 1997, and Granquist and Kovar, 1997). The main output of most statistical processes consists of tables of aggregated data, which are often estimated from a sample of the population. Hence, small errors in individual records can be accepted, provided that (a) these errors mostly cancel out when aggregated, and (b) insofar as they do not cancel out when aggregated, the resulting measurement error in the estimate is small compared to the total error – in particular the natural variation in the estimate due to sampling.

The notion that not all errors need to be corrected in every detail has led to the development of more efficient editing approaches: in particular selective editing, automatic editing and macro-editing. Section 2.4 introduces these approaches, and also illustrates how they may be combined into an effective data editing process. Before that, we discuss different types of errors in Section 2.2 and edit rules in Section 2.3.

We refer to De Waal et al. (2011) and EDIMBUS (2007) for a more comprehensive description of statistical data editing.

2.2 *Types of errors*

Different editing methods have been developed for different types of errors. We will consider here the distinction between influential and non-influential errors and the distinction between systematic and random errors.

Influential errors include the errors that have a significant influence on the final publication total. An error can be influential because it was made by a business that naturally has a strong influence on the estimate, i.e., either by a large business or by a smaller one with a large sampling weight. In addition, sometimes an error is so large that it will strongly influence the total, regardless of the size of the business for which the error occurred. A notorious example of a type of error that is usually influential is the above-mentioned unit of measurement error.

It is clear that errors that have a large influence on a publication total can lead to significant bias. For this reason, it is crucial to treat these errors as effectively as possible. An efficient and timely data editing process will have to focus mainly on the detection and treatment of influential errors. The distinction between influential and non-influential errors is particularly useful in business surveys, because these often contain variables with a skew distribution in the population, such as *Turnover*.

Another distinction that is often made is that between *systematic* and *random* errors.[1] These terms do not have universally accepted definitions. In particular, UN/ECE (2000) defines a systematic error as "an error reported consistently over time and/or between responding units", while EDIMBUS (2007) defines it as "a type of error for which the error mechanism and the imputation procedure are known." The first definition refers in particular to errors that are caused by persistent response problems, which are 'not random' in the sense that they would likely be observed again if the data collection process were repeated. Examples include: the unit of measurement error mentioned in Section 2.1; different definitions used by the statistical institute and the respondent (e.g., gross turnover versus net turnover); persistent problems with data entry or coding at the statistical office. The second definition focuses on the fact that, in many cases, errors of this kind are relatively easy to detect, precisely because they are made in a consistent way. Thus, in many cases, these two definitions of systematic errors agree. In practice, the only systematic errors that can be treated as such are those for which the error mechanism is understood, i.e., errors that are systematic according to the definition of EDIMBUS (2007).

Although the above definitions of systematic errors do not mention bias, it does hold that systematic errors often produce a systematic bias in estimated figures. This is true because these errors are often made in the same way by several respondents. For random errors – i.e., errors that are not systematic as defined in the previous paragraph – the risk of a bias is smaller. On the other hand, random errors are more difficult to detect and correct reliably, precisely because little is known about the underlying causes.

It should be noted that systematic errors may or may not be influential. For instance: the unit of measurement error is usually influential, but an error where a small business with a moderate sampling

[1] Here, the terms 'systematic' and 'random' are supposed to refer to the mechanism that *causes* an error. This differs from the use of these terms in measurement error models, where they refer to the *effect* of an error on an estimator (an error being systematic to the extent that it introduces bias and random to the extent that it introduces noise). As explained in the main text, these two meanings of 'systematic' do overlap to some extent.

weight reports gross turnover instead of net turnover will usually be non-influential. The same holds for random errors.

2.3 Edit rules

To detect errors in observed data, *edit rules* are widely used. These are rules that indicate conditions that should be satisfied by the values of single variables or combinations of variables in a record. Edit rules are also commonly known as *edits* or *checking rules*. If a record does not satisfy the condition specified by an edit rule, the edit rule is said to be failed by that record. Inspection of data items that fail an edit rule is an important technique for finding errors in a data file.

A conceptual distinction should be made between so-called *hard* and *soft* edit rules. Hard edit rules (also known as *fatal* edit rules or *logical* edit rules) are edit rules that must hold by definition, such as

Turnover = *Profit* + *Costs*.

If a hard edit rule is failed by an observed combination of values, then it is certain that at least one of those values contains an error. Soft edit rules (also known as *query* edit rules) indicate whether a value, or value combination, is suspicious. For instance, the soft edit rule

Profit / *Turnover* ≤ 0.6

states that it is unusual for the value of *Profit* to be higher than 60% of the value of *Turnover*. In contrast to hard edit rules, soft edit rules can be failed by unlikely values that are in fact correct. Thus, soft edit failures should trigger a closer investigation of the data items involved, to assess whether the suspicious values are erroneous or merely unusual.

Typically, business surveys involve (mainly) numerical data. For this type of data, some commonly encountered classes of edit rules include the following:

- *Univariate edits / Range restrictions*. These edit rules restrict the range of admissible values for a single variable. A common example is the restriction that a numerical variable may attain only non-negative values, e.g., the edit rule "*Turnover* ≥ 0". Depending on the context, edits of this type can be either hard or soft.
- *Ratio edits*. These edit rules are bivariate restrictions taking the general form $a \leq x / y \leq b$, where x and y are numerical variables and a and b are constants. An example could be that the ratio of *Turnover* and *Number of Employees* (i.e., the average contribution of one employee to the total turnover of a business) should be between certain bounds. The above-mentioned edit rule "*Profit* / *Turnover* ≤ 0.6" is another example of a ratio edit. As the latter example illustrates, some ratio edits contain only a lower bound a or an upper bound b, but not both. Typically, ratio edits are soft edit rules.
- *Balance edits*. These edit rules are multivariate restrictions that relate a set of variables through a linear equality. The above-mentioned edit rule "*Turnover* = *Profit* + *Costs*" is an example of a balance edit. The general form of a balance edit is: $a_1 x_1 + \cdots + a_n x_n + b = 0$, where $x_1, \ldots, x_n$ are numerical variables and $a_1, \ldots, a_n, b$ are constants. Usually, but not always, balance edits are hard edit rules.

2.4 Overview of methods for statistical data editing

The data editing process that is considered here starts after the data have been collected and entered. It should be noted, however, that nowadays many business surveys use computer-assisted modes of data collection (see the topic "Data Collection") which often involve electronic questionnaires. With computer-assisted data collection, it is possible to perform part of the editing already at the data collection stage, for instance by building certain edit rules into the electronic questionnaire. We refer to the theme module "Questionnaire Design – Editing During Data Collection" for a discussion of the possibilities.

The specific way that the data editing process is structured will vary by statistic and by statistical institute. However, there is a general strategy that is followed in broad lines in many processes. This general strategy is shown in Figure 1; similar strategies are discussed in De Waal et al. (2011, pp. 17-21) and EDIMBUS (2007, pp. 6-8). It consists of five steps:

1. Deductive editing;
2. Selective editing;
3. Automatic editing;
4. Interactive editing (manual editing);
5. Macro-editing.

In the remainder of this section, we give a brief outline of each of these steps. More detailed descriptions can be found in the accompanying modules on methods for statistical data editing.

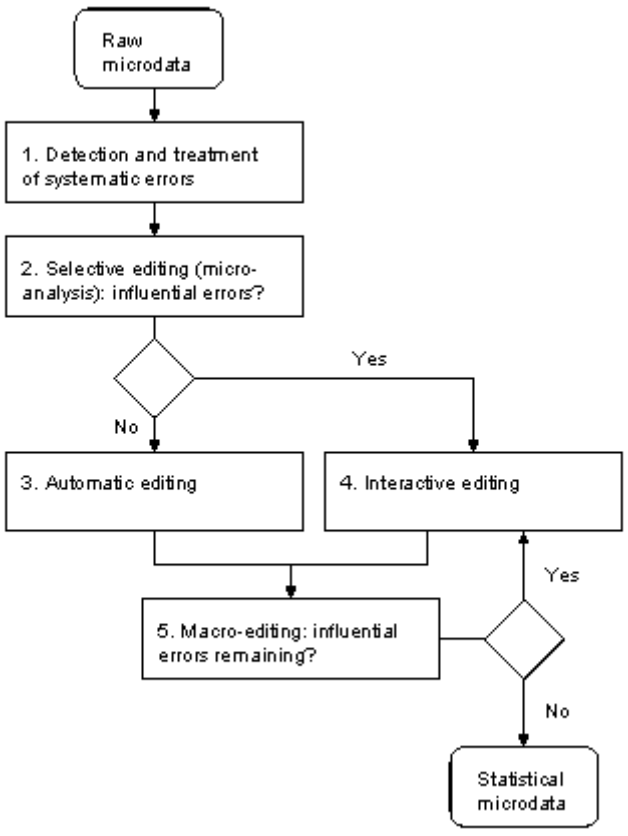

Figure 1. Example of a data editing process flow

In the first phase of the data editing process, identifiable systematic errors are detected and treated. As stated in Section 2.2, these systematic errors can lead to significant bias. Moreover, these errors can often be automatically detected and treated easily and very reliably. It is highly efficient to treat these errors at an early stage. In the remainder of the data editing process, it may then be assumed that the data contain only random errors. The detection and treatment of systematic errors is discussed in the method module "Statistical Data Editing – Deductive Editing".

After the identifiable systematic errors have been edited automatically, a decision can be taken to begin *manual editing*, i.e., manual detection and treatment of errors. This process step is performed by editors or analysts who are usually supported in this regard by software that allows, for example, edit rules to be applied to the data and values to be changed interactively. This form of editing (also known as *interactive editing*) is described in the method module "Statistical Data Editing – Manual Editing".

As mentioned above, manual editing is usually expensive and time-consuming. It is therefore better to restrict the manual work only to records that likely contain influential errors, so that the specialists' limited time can be used where it is most effective. The other records, with less important errors, can either be left unedited or, alternatively, be edited automatically (see below). Limiting interactive editing to those records that likely contain influential errors which cannot be reliably resolved automatically is known as *selective editing* or *micro-selection*. Methods that can be used in this step are discussed in the theme module "Statistical Data Editing – Selective Editing". It should be noted that the selective editing step by itself does not treat any errors; it merely assigns records to different forms of further treatment.

Most selective editing methods make use of anticipated values for the variables in a record to identify the most suspicious values in the observed data. Observed values that deviate strongly from the anticipated values may be caused by influential errors. In determining the anticipated values, information is used from sources other than the actual data file. Oftentimes, edited data from a previous period for the same statistic is used for this purpose. As such, selective editing can proceed on a record-by-record basis, and hence it is possible to start the selection process for manual editing during the data collection period, as soon as the first records are received. This is in fact the main advantage of selective editing over macro-editing, a different selection method to be discussed below.

Records that are not selected for manual editing can be processed by *automatic editing* instead. The automatic treatment of random errors and other errors for which the cause cannot be established usually takes place in two steps. First, the best possible determination is made of what values in a record are incorrect. This is trivial if a value does not fall in the permissible range according to a univariate edit, such as a negative number of employees or an improperly missing value. As such, the value is then certainly incorrect. In many cases, however, inconsistencies can occur for which it is not immediately clear which value or values are responsible. If, for example, the hard balance edit

$$Total\ Costs = Personnel\ Costs + Capital\ Costs + Transport\ Costs + Other\ Costs$$

is not satisfied, then it is clear that (at least) one of the reported values must be erroneous, but it is usually not obvious which one. The problem of identifying the erroneous values in an inconsistent record is known as the *error localisation problem.*

In automatic editing of business survey data, the error localisation problem for random errors is usually solved by applying the *Fellegi-Holt paradigm*, which states: a record should be made

consistent by changing the fewest possible items of data (Fellegi and Holt, 1976). Methods for automatic error localisation based on the Fellegi-Holt paradigm are discussed in the method module "Statistical Data Editing – Automatic Editing".

Once the erroneous values have been detected, they are replaced with better values by means of *imputation*. Automatic imputation relies on (explicit or implicit) mathematical models that use information from the correctly observed values to predict the values that were incorrectly observed or missing. We refer to the topic "Imputation" for a discussion of this subject.

Instead of applying automatic editing, one may also choose not to edit the records that are not selected for interactive treatment by the selective editing procedure. In fact, one may argue that it is not necessary to edit these records, because they will not contain any influential errors, assuming that the selective editing procedure works as intended. Nevertheless, there are reasons why automatic editing may be of use in practice (see also De Waal and Scholtus, 2011). Firstly, it is often desirable to resolve at least all obvious inconsistencies (values that fail hard edit rules), even when these are not influential as such. This is especially true if the microdata are to be released to external users. Secondly, automatic editing provides a relatively inexpensive way to test the quality of a selective editing procedure. If the selection procedure is working correctly, then the records that are not selected for interactive treatment should require only minor adjustments with little influence on a publication figure. Thus, if many influential adjustments are made during automatic editing, this may indicate that the design of the selective editing procedure needs to be improved.

In the final phase of the process in Figure 1, provisional publication figures are calculated and analysed using historical data or external sources. This analysis is called *macro-editing* or *output editing*. If the aggregate figures are implausible, the underlying individual records are examined by, for example, further analysing outliers or influential records and adjusting these as necessary. In Figure 1, this is indicated by the arrow leading back from macro-editing to interactive editing. The errors detected at this stage may be errors that were not found in earlier phases of the data editing process or errors that were actually introduced by the process. In macro-editing, the detection of errors begins at an aggregated level, but the adjustment always takes place in the underlying microdata, i.e., the records of individual respondents. As soon as the provisional figures are considered plausible, the statistical data editing process is completed. For more information on this step, see the module "Statistical Data Editing – Macro-Editing".

In the macro-editing step, as well as during selective editing and manual editing, mathematical techniques for outlier detection are often applied. An extensive discussion of outlier detection in the context of statistical data editing can be found in EDIMBUS (2007).

The process in Figure 1 should be viewed as a prototype. In practice, not all of the steps will be undertaken for all statistics, or a different order of process steps may be used. For instance, it was already mentioned that automatic editing is not always included in the process. Another example is that the selection of records for manual editing is often partly based on other criteria than only whether a record contains influential errors. As such, important or complex businesses are frequently identified as crucial, meaning that their data are always inspected manually. Examples of such businesses could be those that are individually responsible for a significant portion of turnover in their sector. See, e.g., Pannekoek et al. (2013) for a further discussion of the design of an editing process.

Many business surveys have a longitudinal aspect. Sometimes, a panel of units is followed over time during multiple rounds of the same survey. Even for cross-sectional business surveys, the largest units in the population are usually observed in each survey round. This implies that during a particular survey round, at least for part of the responding units, historical data are available. These historical data may be used in various ways during several steps of the editing process; for example, they are often used to determine anticipated values for selective editing. We refer to the theme module "Statistical Data Editing – Editing for Longitudinal Data" for more details on this aspect of statistical data editing.

Finally, it should be noted that, traditionally, applications of statistical data editing have been aimed mainly at survey data. More recently, the use of administrative data for statistical purposes has become increasingly important. These data require an editing process that is in some respects different from the typical editing process for survey data. For instance, for statistics based on administrative data, often all the data (or a large proportion thereof) become available at the same time. In that case, it is not necessary to use micro-selection methods, and we can start immediately with output editing. We refer to the theme module "Statistical Data Editing – Editing Administrative Data" for a discussion of editing in the context of statistics based on administrative data.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

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

De Waal, T. and Scholtus, S. (2011), Methods for Automatic Statistical Data Editing. Paper presented at the 2011 KSS International Conference on Statistics and Probability, Busan.

EDIMBUS (2007), *Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys*. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/RPM_EDIMBUS.pdf.

Federal Committee on Statistical Methodology (1990), *Data Editing in Federal Statistical Agencies*. Statistical Policy Working Paper 18, U.S. Office of Management and Budget, Washington, D.C.

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

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

Granquist, L. and Kovar, J. (1997), Editing of Survey Data: How Much Is Enough? In: L. E. Lyberg, P. Biemer, M. Collins, E. D. De Leeuw, C. Dippo, N. Schwartz, and D. Trewin (eds.), *Survey Measurement and Process Quality*, John Wiley & Sons, New York, 415–435.

Hartwig, P. (2009), How to Use Edit Staff Debriefings in Questionnaire Design. Paper presented at the 2009 European Establishment Statistics Workshop, Stockholm.

Pannekoek, J., Scholtus, S., and van der Loo, M. (2013), Automated and Manual Data Editing: A View on Process Design and Methodology. *Journal of Official Statistics* **29**, 511–537.

Rowlands, O., Eldridge, J., and Williams, S. (2002), Expert Review Followed by Interviews with Editing Staff – Effective First Steps in the Testing Process for Business Surveys. Paper presented at the 2002 International Conference on Questionnaire Development, Evaluation, and Testing Methods, Charleston, South Carolina.

Svensson, J. (2012), Editing Staff Debriefings at Statistics Sweden. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.

UN/ECE (2000), *Glossary of Terms on Statistical Data Editing*. United Nations, Geneva.

Interconnections with other modules

8. Related themes described in other modules

1. Questionnaire Design – Editing During Data Collection
2. Data Collection – Main Module
3. Statistical Data Editing – Selective Editing
4. Statistical Data Editing – Macro-Editing
5. Statistical Data Editing – Editing Administrative Data
6. Statistical Data Editing – Editing for Longitudinal Data
7. Imputation – Main Module

9. Methods explicitly referred to in this module

1. Statistical Data Editing – Deductive Editing
2. Statistical Data Editing – Automatic Editing
3. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. GSBPM Sub-process 5.3: Review, validate and edit

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1. Statistical data editing

Administrative section

14. Module code

Statistical Data Editing-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	09-12-2011	first version	Sander Scholtus	CBS (Netherlands)
0.2	20-04-2012	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	19-06-2012	minor improvements	Sander Scholtus	CBS (Netherlands)
0.3.1	16-07-2013	adjusted to new template; minor improvements	Sander Scholtus	CBS (Netherlands)
0.4	31-10-2013	minor improvements based on comments by Italian reviewer and Editorial Board	Sander Scholtus	CBS (Netherlands)
0.4.1	31-10-2013	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:10