This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Macro-Editing

Contents

General section

1. Summary

In most business surveys, it is reasonable to assume that a relatively small number of observations are affected by errors with a significant effect on the estimates to be published (so-called influential errors), while the other observations are either correct or contain only minor errors. For the purpose of statistical data editing, attention should be focused on treating the influential errors. *Macro-editing* (also known as *output editing* or *selection at the macro level*) is a general approach to identify the records in a data set that contain potentially influential errors. It can be used when all the data, or at least a substantial part thereof, have been collected.

Macro-editing has the same purpose as selective editing (see "Statistical Data Editing – Selective Editing"): to increase the efficiency and effectiveness of the data editing process. This is achieved by limiting the costly manual editing to those records for which interactive treatment is likely to have a significant effect on the quality of the estimates. The main difference between these two approaches is that selective editing selects units for manual follow-up on a record-by-record basis, whereas macro-editing selects units by considering all the data at once. It should be noted that in macro-editing all actual adjustments to the data take place at the *micro* level (i.e., for individual units), not the *macro* level. Methods that perform adjustments at the macro level are discussed in the topic "Macro-Integration".

2. General description

2.1 Introduction to macro-editing

Macro-editing is a general approach to identify potentially influential errors in a data set for manual follow-up. It can be used when all the data, or at least a substantial part thereof, have been collected. In addition, the method is particularly effective when it is applied to data that contain only a limited number of large errors. Given these conditions, macro-editing is typically applied towards the end of a data editing process. At that stage, the errors that one expects to find in the data are either remaining errors that 'slipped through' previous editing efforts or errors that were actually introduced during data processing (processing errors). Possible sources of processing errors include automated data handling (e.g., loading the wrong data set, running an application with the wrong set of parameters, a bug in the software) as well as wrong decisions made by editors during manual editing. Macro-editing may succeed in finding these errors by examining the data from a macro rather than a micro level perspective – in other words, looking at the whole data set instead of one record at a time.

Macro-editing proceeds by computing aggregate values from a data set and systematically checking these aggregates for suspicious values and inconsistencies. The following types of checks are typically used:

- Internal consistency checks. In most business surveys, the definitions of the survey variables imply that the aggregated data should satisfy certain logical or mathematical restrictions. For instance, in each stratum, total net turnover (say X) should equal the sum of total net turnover from domestic sales (X_1) and total net turnover from foreign sales (X_2); i.e., it should hold that $X = X_1 + X_2$. In addition, based on subject-matter knowledge the fraction of total net turnover from domestic sales may be expected to lie between certain bounds; i.e., $a < X_1 / X < b$ for certain constants a and b. These restrictions are the macro-level equivalents of edit rules that were used during micro-editing (see “Statistical Data Editing – Main Module”). Like edit rules, they may be either hard restrictions (identifying erroneous aggregates with certainty, such as the first example given above) or soft restrictions (identifying suspicious aggregates that may occasionally be correct, such as the second example).

- Comparisons with other statistics. It may be possible to compare aggregates to similar estimates from other data sources. If large differences occur, the corresponding aggregates are identified as suspicious. Such comparisons can be useful, if only to promote coherence between different statistical outputs. On the other hand, the comparability of aggregates from different sources is often affected in practice by conceptual and operational differences (e.g., different target populations, differences in variable definitions, different reference periods). It is important to be aware of these differences when they exist.
- Comparisons with previously published statistics. In repeated surveys, one can compare current aggregates to a time series of previously published values. If a sufficiently long time series is available, one may apply time series analysis to identify possible trend discontinuities and hence suspicious aggregates.
- Other quality information about the statistical process so far. For instance, a non-response analysis provides information on aggregates that have a high risk of being biased. If estimates of sampling errors are available, these may also be incorporated in the macro-editing procedure (see Section 2.2).

It should be noted that in macro-editing all actual adjustments to the data take place at the *micro* level, not the *macro* level. Therefore, after one has found suspicious aggregates by any of the above means, the next step is to identify individual units that contribute to these aggregates and may require further editing. The next two subsections describe two generic approaches to do this. The *aggregate method* (Section 2.2) proceeds by ‘drilling down’ from suspicious aggregate values to lower-level aggregates and, eventually, individual units. The *distribution method* (Section 2.3) examines the distribution of the microdata to identify outliers and other suspicious values. In practice, the two methods are often applied together.

2.2 The aggregate method

Given a data set that requires macro-editing, the aggregate method starts by calculating estimates of aggregates at the highest level of publication based on the current data (Granquist, 1994). These provisional publication figures are checked for plausibility and consistency, as discussed in Section 2.1. If an aggregate is identified as suspicious, the next step is to zoom in on the cause of the suspicious value by examining the lower-level aggregates that contribute to the suspicious aggregate. This procedure is sometimes called ‘drilling down’. In this way, macro-editing proceeds until the lowest level of aggregation is reached, i.e., the individual units. Finally, the units that have been identified as the most important contributors to a suspicious provisional publication figure are submitted to manual follow-up (see “Statistical Data Editing – Manual Editing”).

In practice, checking for suspicious aggregates is often implemented by means of score functions, similar to those that are used at the micro level in selective editing (see "Statistical Data Editing – Selective Editing"). In macro-editing, the score function is applied at the aggregate level (e.g., Farwell and Schubert, 2011). In practice, relatively simple score functions are often used, such as:

$$S_j = \frac{\hat{T}_{y_j} - \tilde{T}_{y_j}}{\tilde{T}_{y_j}}, \tag{1}$$

where $\hat{T}_{y_j}$ is the estimated total of variable y_j based on the unedited data, and $\tilde{T}_{y_j}$ is a corresponding anticipated (or predicted) total value. This score function measures the relative deviation from the anticipated value. Possible sources of anticipated values are: estimates from different data sources, such as a register or a different survey, or the value of the same total in a previous survey cycle – possibly corrected for development over time using a time series model (see also Section 2.1).

Comparisons based on ratios of aggregated values are also used, such as:

$$S_{jk} = \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right) \Bigg/ \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}}, \tag{2}$$

using notation similar to (1).

Since macro-editing is applied when all, or nearly all, data are available, there is no need to set a threshold value on the score function in advance. Instead, the aggregates can be put in order of suspicion by sorting on the absolute value of S_j or S_{jk}. In order to prevent the introduction of bias, it is important to treat large positive and large negative deviations from the anticipated values with equal care.

If the estimates are based on a sample of the population, as is often the case in business surveys, a natural amount of variation in the aggregates is expected due to sampling error. From a theoretical point of view, it is good to take this inaccuracy of the estimated aggregates into account in the score function. Thus, instead of (1), one could use

$$S'_j = \frac{\hat{T}_{y_j} - \tilde{T}_{y_j}}{se(\hat{T}_{y_j} - \tilde{T}_{y_j})},$$

and instead of (2), one could use

$$S'_{jk} = \left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right) \Bigg/ se\left(\frac{\hat{T}_{y_j}}{\hat{T}_{y_k}} - \frac{\tilde{T}_{y_j}}{\tilde{T}_{y_k}} \right),$$

where $se(.)$ indicates the standard error of an estimate. In these alternative score functions, deviations from the anticipated values are only seen as suspicious if they are large compared to the associated sampling error. This refinement is particularly important if there are large differences in accuracy between different aggregates.

For the final step in the aggregate method, the so-called 'drilling down' from suspicious aggregates to contributing individual units, the same score functions on the micro level can be used as in selective editing (see "Statistical Data Editing – Selective Editing"). The main difference is that, again, there is no need to set a threshold value in advance here, because the score function can be computed for all records at the same time. This means that the records can be sorted on their score function value and treated in order of priority.

As an alternative to the aggregate method, one could also consider working directly with the sorted record-level score function values, by manually following up records in descending order of their absolute scores and continuing until all aggregates are deemed sufficiently plausible. This was called the *top-down method*[1] by Granquist (1994).

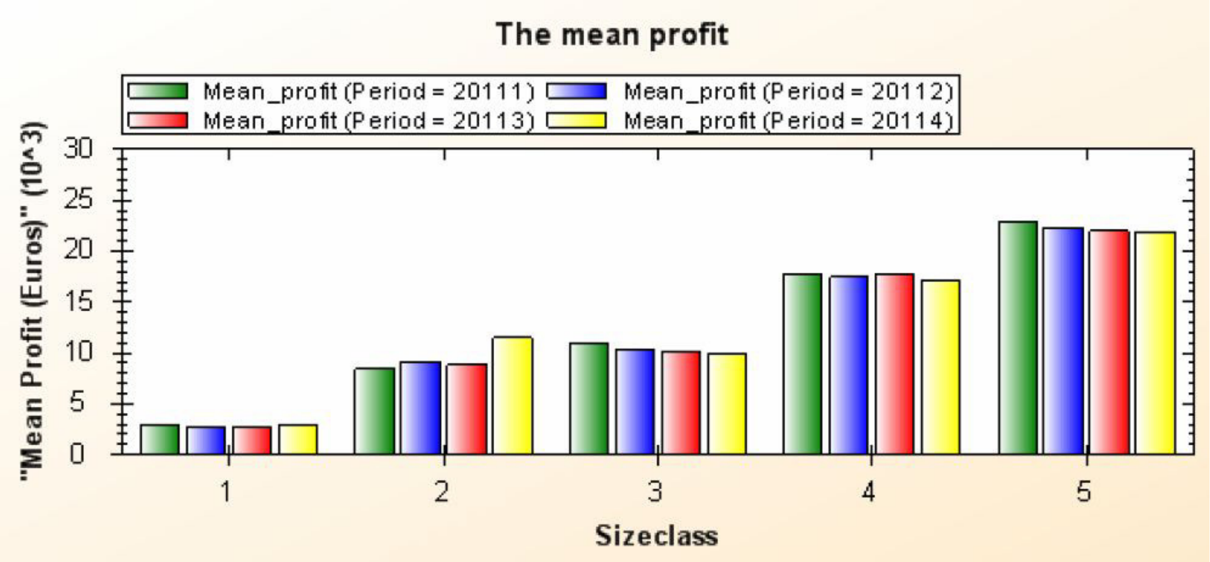


Figure 1. Example of a histogram for macro-editing (taken from Hacking and Ossen, 2012).

In addition to score functions, graphical aids can also be useful for identifying suspicious aggregates. As an example, Figure 1 shows a histogram that compares the mean value of profit across several reference periods and several size classes. It is seen that the mean profit in the last period for size class 2 is unusually high in comparison with previous periods and other size classes. This could be a reason to identify this aggregate as suspicious and drill down to the contributing units.

2.3 The distribution method

Another method for selecting individual units for manual editing, given all or most of the data, is known as the distribution method. This method tries to identify observations that require further treatment by applying techniques for detecting *outliers*, i.e., observations that deviate from the distribution of the bulk of the data. For the purpose of macro-editing, records are then prioritised for manual follow-up by ordering them on some measure of 'outlyingness'. A discussion of outlier detection techniques in the context of statistical data editing can be found in EDIMBUS (2007).

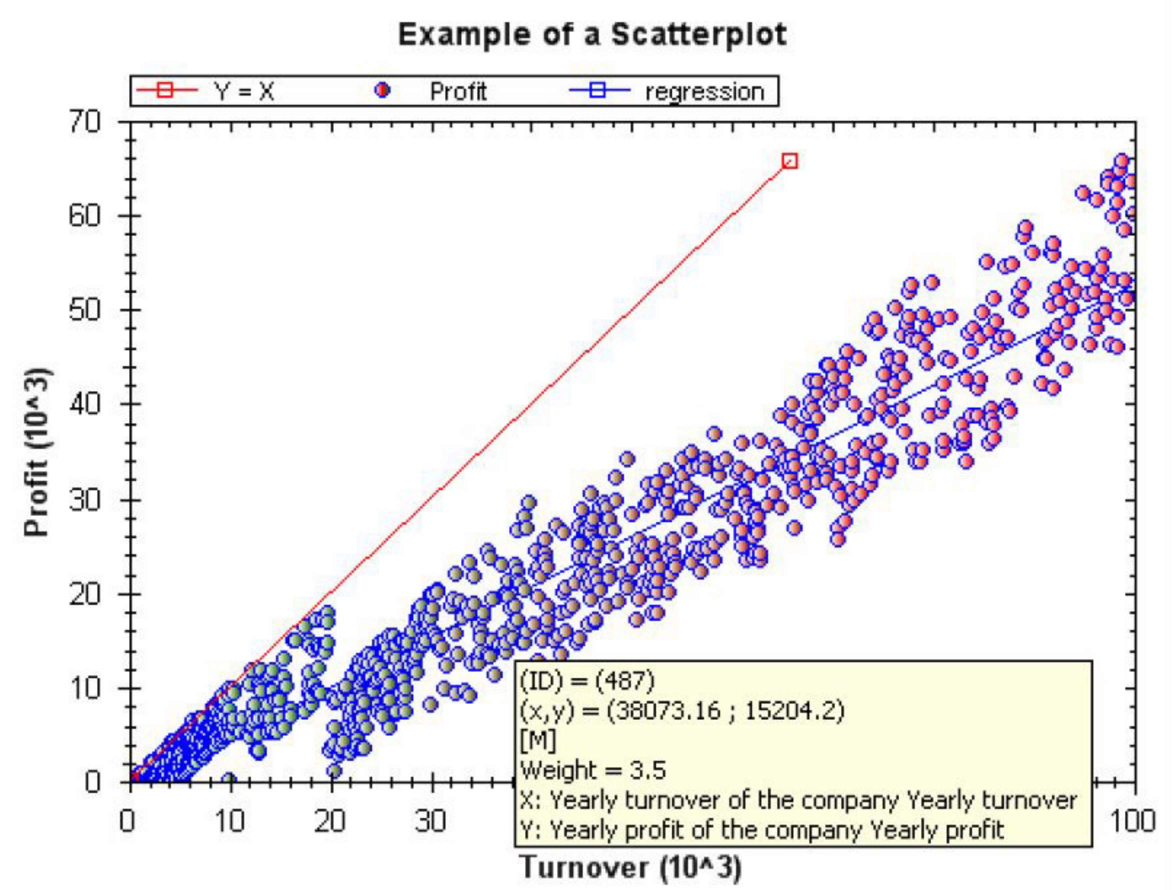
---

[1] The name 'top-down method' is a potential source of confusion, because it is sometimes used as a synonym for the aggregate method (e.g., De Waal et al., 2011, p. 208). This probably derives from the fact that the aggregate method starts at 'top level' aggregates and 'drills down' to lower-level aggregates.

Theoretically speaking, there exists some overlap between this approach and the above approach based on score functions, because many common criteria for detecting outliers can be expressed as score functions; see, e.g., De Waal et al. (2011).

Graphical displays can also be useful for detecting observations that deviate from the distribution of the bulk of the data. Common examples include box plots, scatterplots, and other techniques from Exploratory Data Analysis (Tukey, 1977). Figure 2 gives an example of a scatterplot that could be used in this context. A graphical analysis can be particularly effective if the software allows an editor to interact with a display. In the plot of Figure 2, whenever a user moves his mouse to one of the points, information about the relevant unit is automatically displayed. This can be taken one step further by letting a user access a record for further editing by simply clicking on the point that represents the record in the graphical display. See, e.g., Bienias et al. (1997) and Weir et al. (1997) for examples of applications of graphical macro-editing. For some more recent innovations, see Tennekes et al. (2012).

In practice, the distribution method is often applied in conjunction with the aggregate method. Thus, the macro-analysis starts by identifying suspicious aggregates at the highest level and ‘drills down’ to suspicious aggregates at a lower level. Subsequently, the distribution method is applied to identify the records that are likely to contribute most to the total error in the identified low-level aggregates.



(ID) = (487)
(x,y) = (38073.16 ; 15204.2)
[M]
Weight = 3.5
X: Yearly turnover of the company Yearly turnover
Y: Yearly profit of the company Yearly profit

Figure 2. Example of a scatterplot for macro-editing (taken from Hacking and Ossen, 2012).

3. Design issues

4. Available software tools

Many statistical offices have developed macro-editing tools. Quite often, several such tools exist within one office, each one dedicated to a particular survey.

Statistics Netherlands has developed a generic macro-editing tool called *MacroView*; see Ossen et al. (2011) and Hacking and Ossen (2012). It is currently used for macro-editing in the production processes of the Dutch structural business statistics and the Dutch short-term statistics, as well as several smaller statistical processes. It is currently not made available to other statistical offices.

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

Bienias, J. L., Lassman, D. M., Scheleur, S.A., and Hogan, H. (1997), Improving Outlier Detection in Two Establishment Surveys. In: *Statistical Data Editing, Volume 2: Methods and Techniques*, United Nations, Geneva, 76–83.

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

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.

Farwell, K. and Schubert, P. (2011), A Macro Significance Editing Framework to Detect and Prioritise Anomalous Estimates. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.

Granquist, L. (1994), Macro-Editing – a Review of Some Methods for Rationalizing the Editing of Survey Data. In: *Statistical Data Editing, Volume 1: Methods and Techniques*, United Nations, Geneva, 111–126.

Hacking, W. and Ossen, S. (2012), User Manual MacroView. Report PMH-20121125-WHCG, Statistics Netherlands, Heerlen.

Ossen, S., Hacking, W., Meijers, R., and Kruiskamp, P. (2011), MacroView: a generic software package for developing macro-editing tools. Working Paper, UN/ECE Work Session on Statistical Data Editing, Ljubljana.

Tennekes, M., de Jonge, E., and Daas, P. (2012), Innovative Visual Tools for Data Editing. Working Paper, UN/ECE Work Session on Statistical Data Editing, Oslo.

Tukey, J. W. (1977), *Exploratory Data Analysis*. Addison-Wesley, London.

Weir, P., Emery, R., and Walker, J. (1997), The Graphical Editing Analysis Query System. In: *Statistical Data Editing, Volume 2: Methods and Techniques*, United Nations, Geneva, 96–104.

Interconnections with other modules

8. Related themes described in other modules

1. Statistical Data Editing – Main Module
2. Statistical Data Editing – Selective Editing
3. Macro-Integration – Main Module

9. Methods explicitly referred to in this module

1. Statistical Data Editing – Manual Editing

10. Mathematical techniques explicitly referred to in this module

1. n/a

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. MacroView

13. Process steps explicitly referred to in this module

1. Statistical Data Editing

Administrative section

14. Module code

Statistical Data Editing-T-Macro-Editing

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	04-03-2013	first version	Sander Scholtus	CBS (Netherlands)
0.2	18-04-2013	improvements based on Swedish review	Sander Scholtus	CBS (Netherlands)
0.3	19-07-2013	minor improvement based on second Swedish review	Sander Scholtus	CBS (Netherlands)
0.3.1	09-09-2013	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 18:12