WP.19
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, September 2006)

Topic (iii): Editing microdata for release

POST-EDITING: A POTENTIAL SOURCE OF INCONSISTENCIES

Supporting Paper

Prepared by Ton de Waal, Statistics Netherlands[1]

Abstract: Statistical offices collect data in order to fulfil many different needs of society for statistical information. In order to fulfil these needs efficiently statistical offices often use the same data for several different purposes. For instance, data may first be used to publish statistical figures directly. Later these data may also be used for additional publications, such as price indexes or the national accounts. For the direct publication of statistical figures an extensive edit and imputation process is generally carried out. For each additional publication a separate post-editing process may be required, however. These separate post-editing processes may lead to consistency problems. In this paper we focus on inconsistencies due to different ways of post-editing the data.

I. INTRODUCTION

1. Statistical offices collect data in order to fulfil many different needs of society for statistical information. For efficiency reasons and in order to limit the response burden statistical offices generally aim to satisfy several of those needs with the same data. Data are therefore often not collected for a single purpose, but rather for several purposes at the same time. For instance, at Statistics Netherlands data from businesses are not only collected in order to publish statistical figures on these businesses directly, but also to publish price indexes and as input for the national accounts. National accounts later give a comprehensive and coherent overview of the Dutch economy as a whole (see United Nations, 2003, for an introduction to the national accounts, and Bos, 2006, for a historic overview of the development of the Dutch national accounts).

2. Using the same data for multiple purposes may lead to consistency problems. Since there are several different purposes of the data, the data may need to be treated in different ways. In particular, the data may need to be post-edited after the first publication of statistical figures. In this paper we focus on inconsistencies due to post-editing the data. We restrict ourselves to identifying the problem in general. We do not attempt to pinpoint detailed causes of the problem, nor do we attempt to provide solutions to the problem.

---

[1] Ton de Waal (twal@cbs.nl)

3. In this paper the term post-editing is to be understood in a broad sense. The term refers, for instance, to the detection and correction of outliers, but also to the process carried out for the national accounts. This latter process incorporates confrontation of various data sources, adjustment of conflicting values, and balancing of components to totals.

4. The problem we aim to describe in this paper arises because the statistical process at a statistical office is generally a chain of related subprocesses rather than one all-compassing process. Each of these subprocesses has its own goals. Agreements, such as service level agreements, are made between departments carrying out different subprocesses in order to ensure that the overall result of the entire chain of subprocesses is of acceptably high quality. It is, however, impossible to prevent all inconsistencies that can possibly arise at various points in the chain.

5. The remainder of this paper is organised as follows. Section II sketches the edit and imputation process for producing business statistics directly. Section III describes some of the reasons for inconsistencies when the data are later used to produce price indexes and the national accounts. Section IV concludes the paper with a brief discussion.

II. PUBLISHING BUSINESS DATA DIRECTLY: THE EDITING PROCESS

6. Traditionally, statistical agencies have put a lot of effort and resources into data editing, as they considered it a prerequisite for publishing accurate statistics. In traditional survey processing, data editing was mainly an interactive (manual) activity intended to correct all data in every detail. Detected errors or inconsistencies were reported and explained on a computer screen and corrected after consulting the questionnaire, or re-contacting respondents: time and labour-intensive procedures.

7. It has long been recognised, however, that it is not necessary to correct all data in every detail. Several studies (see for example Granquist, 1984; Granquist, 1997; Granquist and Kovar, 1997) have shown that in general it is not necessary to remove all errors from a data set in order to obtain reliable publication figures. The main products of statistical offices are tables containing aggregated data, which are often based on samples of the population. This implies that small errors in individual records are acceptable. First, since small errors in individual records tend to cancel out when aggregated. Second, since if the data are obtained from a sample of the population there will always be a sampling error in the published figures, even when all collected data are completely correct. In this case an error in the results caused by incorrect data is acceptable as long as it is small in comparison to the sampling error. In order to obtain data of sufficiently high quality it is usually enough to remove only the most influential errors. The above-mentioned studies have been confirmed by many years of practical experience at several statistical offices.

8. Instead of the traditional interactive approach modern, techniques such as selective editing, automatic editing and macro-editing can be applied. Selective editing (cf. Lawrence and McKenzie, 2000; Hoogland, 2002; Hedlin, 2003) is applied to split the data into two streams: the critical stream and the non-critical stream. The critical stream consists of those records that are the most likely to contain influential errors; the non-critical stream consists of records that are unlikely to contain influential errors. The records in the critical stream are edited in the traditional, manual manner. The records in the non-critical stream are either not edited or are edited automatically. When automatic editing is applied a record, it is entirely edited by a computer instead of by a clerk. Automatic editing can lead to a substantial reduction in costs and time required to edit the data. For an overview of algorithms for automatic editing, see De Waal and Coutinho (2005).Macro-editing (cf. Granquist, 1990; De Waal, Renssen and Van de Pol, 2000), i.e. verifying whether figures to be published seem plausible, is often an important final step in the editing process. Macro-editing can reveal errors that would go unnoticed with selective editing or automatic editing.

9. Selective editing, automatic editing and macro-editing can be used in combination. For instance, in the uniform system for short-term statistics at Statistics Netherlands we apply a combination of selective editing and macro-editing (see De Jong, 2003). In the uniform system for annual structural business surveys at Statistics Netherlands we apply an edit and imputation approach that consists of the following main steps (see De Jong, 2002):

- application of selective editing to split the records in a critical stream and a non-critical stream;
- editing of the data: the records in the critical stream are edited interactively, the record in the non-critical stream are edited and imputed automatically;
- validation of the publication figures by means of (graphical) macro-editing.

10. Partly for efficiency reasons (combinations of) selective editing, automatic editing and macro-editing have become very popular in recent years. The main idea behind these approaches is that only a subset of the records are edited manually, namely the most influential or most risky records for a certain domain. To determine which records should be edited manually one usually selects in advance a number of domains for which the quality of the data should be of an acceptably high level. These domains can, for instance, be based on combinations of the industry code and the size class of an enterprise. Quality of the microdata is only guaranteed for the selected domains, not for subsets of the domains and certainly not for individual records.

11. We end this section with a remark on outliers for short-term statistics. An important aim of short-term statistics, such as monthly business statistics, is to estimate the growth of the economic activity in a certain domain. In order to estimate the month-to-month growth on a detailed aggregation level, i.e. small domains, accurately it may be necessary to treat relatively many observations as outliers (see, e.g., Chambers, 1987, and Maronna, Martin and Yohai, 2006, for more on outliers). These outliers would otherwise cause too much fluctuation in the data, and hence would lead to (too) inaccurate estimates for the month-to-month growth. We will return to this remark later.

III. CAUSES FOR INCONSISTENCIES IN LATER STAGES OF THE PROCESS

12. After business data have been used to publish statistical figures directly, they may later be used to publish information on price indexes and as input for the national accounts. This may lead to inconsistencies. In this paper we distinguish between four causes for inconsistencies: different domains, different moments in time, different background information, and different aggregation levels. Below we briefly examine each of these causes.

13. Different domains. In order to publish reliable price indexes and national accounts, other domains may be important than the domains selected for direct publication. If this is the case, the microdata may need to be post-edited because the quality of the data is only guaranteed for the domains for direct publication, not for other domains or combinations thereof. This problem does occur at Statistics Netherlands. When price indexes are determined, the important domains for which high quality statistical microdata is required, differ from the domains used for direct publication for which the quality is guaranteed. Therefore, for these important domains for determining price indexes outliers are (again) detected and treated. The detected outliers may differ from the outliers detected earlier in the chain of subprocesses. This is a side-effect of using (combinations of) selective editing, automatic editing and macro-editing where the quality of the microdata is only guaranteed for the selected domains. The problem can be resolved by making better agreements on for which domains the quality should be guaranteed. By making such agreements more data may have to be edited earlier in the process than is done at the moment.

14. Different moments. Consistency problems between publication figures based on the same microdata can also occur if these figures are calculated at different moments in time. For instance, when the national accounts are being compiled more may be known about the population. That is, incorrect

industry codes or size class codes for certain businesses may have been corrected. Also, late response may not yet have been incorporated into the microdata when statistical figures are published directly but may have been incorporated into the microdata when the results for the national account are being determined.

15. Different background information. When different background information is used to determine statistical figures, inconsistencies may obviously arise. Different background information may, for instance, be used while imputing for missing values or when calculating raising weights.

16. This cause for inconsistency is related to the previous one regarding different moments in time. As mentioned before, the process at a statistical office can generally be subdivided into a chain of related subprocesses. The later a subprocess occurs, the more background information is available. For instance, at Statistics Netherlands the national accounts are produced at the end of the chain of subprocesses. To produce these national accounts figures, i.e. aggregated data, from various available sources, e.g. short-term statistics and annual structural business statistics, are confronted with each other. When differences between sources occur, these differences are resolved by adjusting figures, usually on a macro-level. If figures that are adjusted have already been published, the results of the national accounts differ from already published figures.

17. Different aggregation levels. The level of detail can also lead to consistency problems. As noted earlier in paragraph 11, in order to publish the growth from one month to the next for short-term statistics one may have to consider relatively many observations as outlier, because else these monthly figures would fluctuate too much due to their variance. By assigning observations the status of outlier – and adjusting their raising weights accordingly – the variance can be lowered at the expense of introducing bias. One hereby aims to minimise the mean squared error. The national accounts may use less detailed aggregation levels, because there the primary aim is to give a coherent overview of the entire economy rather than detailed data for a specific branch of industry. For these less detailed aggregation levels, fluctuations due to variance are a less important issue. Hence, one can focus on reducing bias by assigning fewer observations the status of outlier. In principle, the converse where the national accounts want to determine figures on a more detailed level can also happen. This is especially the case for domains that are not very well observed. In such a case the national accounts may assign the status of outlier to more observations than has been done for direct publication. Whenever the outliers for direct publication and for the national accounts differ the national accounts can either make changes in the microdata, or – more usually – make adjustments on a macro-level.

IV. DISCUSSION

18. In this paper we have seen that when the same data are used for multiple purposes, inconsistencies can arise. Partly these inconsistencies are inevitable. This is in particular the case if data are used for different purposes at different moments in time. Later in time more and better (background) information is available. For instance, editing and imputation can hence be carried out more accurately at a later moment in time. Inconsistencies arising due to increased information indicate improvement of the quality of the data. Such inconsistencies can be explained to and understood by users of the data.

19. Other inconsistencies are not inevitable. For example, inconsistencies due to the fact that other domains are considered important for the calculation of price indexes than for the direct publication of figures on business statistics can be avoided by reached a consensus on the domains for which the quality should be guaranteed. Such inconsistencies are hard to explain to users of the data, and should be avoided.

20. Finally, some other inconsistencies seem to lie in the middle between inevitable and avoidable. For example, inconsistencies due to different aggregation levels can partly be resolved by making agreements and by developing methodology aimed at reducing inconsistency. Making agreements and developing methodology are probably not sufficient to resolve all inconsistencies due to different aggregation levels. Explaining such complex inconsistencies to users of the data seems to be quite hard.

20. The most important lesson that can be drawn from this paper is that when one publishes statistical figures one is not alone. Basically the same, or strongly related, information is often published at an earlier or a later stage in the entire chain of subprocesses. Coherence and consistency across the entire chain of subprocesses and corresponding publications is what we should aim for. To achieve such coherence and consistency the methods of the various subprocesses should geared to one another and agreements should be made between the various departments involved.

References

Bos, F. (2006), The Development of the Dutch National Accounts as a Tool for Analysis and Policy. *Statistica Neerlandica 60*, pp. 206-224.

Chambers, R.L. (1987), Outlier Robust Finite Population Estimation. *Journal of the American Statistical Association 81*, pp. 1063-1069.

De Jong, A. (2002), *Unit-Edit: Standardized Processing of Structural Business Statistics in the Netherlands.* UN/ECE Work Session on Statistical Data Editing, Helsinki.

De Jong, A. (2003), *IMPECT: Recent Developments in Harmonized Processing and Selective Editing*. UN/ECE Work Session on Statistical Data Editing, Madrid.

De Waal, T. and W. Coutinho (2005), Automatic Editing for Business Surveys: An Assessment of Selected Algorithms. *International Statistical Review 73*, pp. 73-102.

De Waal, T., R. Renssen and F. Van de Pol (2000), Graphical Macro-Editing: Possibilities and Pitfalls. *Proceedings of the Second International Conference on Establishment Surveys*, Buffalo, pp. 579-588.

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

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

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

Granquist, L. and J. Kovar (1997), Editing of Survey Data: How Much is Enough?. In: *Survey Measurement and Process Quality* (ed. Lyberg, Biemer, Collins, De Leeuw, Dippo, Schwartz & Trewin), John Wiley & Sons, Inc., New York, pp. 415-435.

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

Hoogland, J. (2002*).* Selective Editing by Means of Plausibility Indicators. UN/ECE Work Session on Statistical Data Editing, Helsinki.

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

Maronna, R.A., R.D. Martin and V.J. Yohai (2006), *Robust Statistics*. John Wiley & Sons, Ltd., Chichester.

United Nations (2003), *National Accounts: A Practical Introduction*, ST/ESA/STAT/SER.F/85, New York.
