European Establishment Statistics Workshop, 7-9 September 2009, Stockholm, Sweden

# Towards an efficient data editing strategy for economic statistics at Statistics Netherlands

Frank Aelen & Roos Smit, Statistics Netherlands

## Abstract

*The main focus of this paper is on the new data editing strategy of Statistics Netherlands, for both STS and SBS processes. This work is done as part of the so-called HEcS programme, in which the economic statistics are being redesigned. The first major change in data editing will be an improved method for automatic editing. This method is expected to increase the number as well as the quality of the automatically edited records. Tests show that of the 50.000 records received annually for the SBS the percentage of automatically edited records will increase from 30% to approximately 70%. The second improvement is a top-down analysis strategy to selectively edit data interactively. This means that only if an aggregate is found implausible analysis will continue and only records which contribute to an implausible aggregate will be edited. For systematic or structural errors that are encountered in the analysis process, detection and correction rules are constructed that in the future can be used in the automatic data editing. In the new editing process, data from different sources can be confronted. This is an efficient means of data editing that is integrated in the estimation process and which requires only a minimal amount of micro data to be evaluated by an analyst. Currently, tools are being built to implement the new data editing strategy in the production processes of Statistics Netherlands.*

## 1. Introduction

Statistics Netherlands is currently redesigning the chain of processes for compiling its economic statistics. This is done in the so-called HEcS programme[1], which aims to: 1) compile economic statistics more efficiently; 2) improve the quality and coherence of economic statistics; 3) increase the use of administrative data, thereby reducing the response burden for businesses. At this time one of the main topics of the HEcS programme is a joint redesign of the Short-term Statistics (STS) and the Structural Business Statistics (SBS) production processes. The data sources for compiling these statistics are tax registers as well as survey data.

This paper will focus on the new data editing strategy of the HEcS programme. First, an improved method for automatic data editing for the SBS will be discussed. Subsequently, the top-down analysis strategy to selectively edit data interactively for STS as well as SBS is described.

## 2. Automatic micro data editing

### *2.1 Automatic editing in recent years*

Data editing is preferably done automatically, since interactive (manual) data editing is time consuming and costly. Not all errors in micro data, however, are suitable for automatic editing. In recent years, Statistics Netherlands edited only about 30% of the records for the SBS automatically. The automatic editing process consists of two methods: one to select records that are deemed suitable for automatic editing and a second to perform the actual editing. The selection method uses a plausibility index in which amongst other things the distance to the median for each of a number of variables or a ratio of variables is used in order to decide whether a record is suitable for automatic editing or needs to be viewed by an analyst. Influential records with strongly deviating values are not considered suitable for automatic treatment. This method was designed such that approximately 50% of the non-crucial records are deemed plausible enough for automatic editing. On the whole only 30% of the records are edited automatically since some industries do not use automatic editing and all the

[1] See also “Redesign of the chain of economic statistics in the Netherlands”, Seminar on registers in statistics – methodology and quality, 21-23 May 2007, Helsinki, Barteld Braaksma, Statistics Netherlands.

crucial records are edited interactively. Subsequently the automatic editing process, implemented in a software system called SLICE[2], uses a set of editing rules to determine the records with errors. Next the variables in these records are located that are considered to be in error. This localization of erroneous variables is done by an optimization routine based on the “minimum change” principle: it designates the least possible number of variables as erroneous such that by changing these variables all edit rules can be satisfied. As a last step in the automatic editing procedure the values of the erroneous variables are replaced by imputed values that are consistent with the edit rules.. Note that SLICE currently has no knowledge about the cause of the errors, so knowledge about this cause is not used in the correction methods.

Both the selection and detection modules are meant to select or edit incidental errors in the data using medians per industry (according to the NACE-code) and size class. They serve their purpose reasonably well, but there are some problems. First, the variance per industry times size class is often large. This makes finding errors by using the median very difficult and often correct records are unduly selected for interactive editing. Second, many of the errors in the micro data are structural errors of which the cause and proper correction method are known but not used. For example restaurants, which should put their revenue value in the service box often incorrectly, use the trade box because they sell food. This happens in about 10% of the questionnaires and would incorrectly cause a large secondary activity if left untreated. The current methods have great difficulty in correcting these kinds of errors and it is also not efficient to select all these records for interactive editing. Third, the above-described methods behave in a way like a black box and can be very incomprehensible to analysts. They have difficulties in understanding why a certain record is selected for interactive editing or why in the automatic editing process a certain correction was made. This makes controlling and predicting the system difficult and it often results in undesirable solutions.

## *2.2 New and improved*

Recently, several modules were added to the system of detection and optimisation. These new modules correct common structural errors in the data, which results in a better quality of publication data as well as a lower disapproval rate of the optimisation module[3]. Taking a closer look at the data, however, revealed that not all records selected for automatic editing were actually suitable for this. Correcting the flaws made in automatic editing took up a lot of analysis time. Therefore, in summer 2008 a test was started to develop a new selection module to better facilitate the new automatic editing and at the same time examine whether more additions to SLICE could further improve data quality. The following industries were incorporated in the test: bakeries, computer service industries, supermarkets and plumbers. The main principle behind the new modules is to use expert knowledge in deciding which parameters to use and if desirable even set limits according to an expert guess. This proves especially beneficial in deciding how to select and correct structural errors as well as setting limits in industries and size classes with a small population. The percentage of records selected for automatic editing increased to 70%, for all industries except the always very difficult plumbers. For the plumbing industry the percentages remained about the same as before, but very different records were selected for automatic editing. For all industries the quality of the automatically edited records was conceived by experts as better than before. Based on these results the modules are now being developed further and they should be production-ready by October of this year. Below, the new modules are explained further.

### *2.2.1 A new selection module*

One of the main features of the new selection module is its flexibility. This flexibility is needed because the module has to facilitate a large number of very different industries. So in the new module it is possible to define selection rules containing functions of a number of variables, for example ratios and relations between variables such as inequalities (A>B). For instance a record should not be selected for automatic editing if A>B and A>500. The value 500 in the example can either be an expert guess, the value from last year, a median or any other calculated value. It is also possible to use reference material from other data sources. This makes the new module very flexible. Tests show that the module is able to make a precise distinction between records that are suitable for automatic editing

[2] See also “From CherryPi to SLICE”, 1999, De Waal, T. & H. Wings, Statistics Netherlands.

[3] See also “Onderzoeksrapport PI” (in Dutch only), 2008, Smit R., Statistics Netherlands.

and records that need interactive attention. The selection module is used twice. Once before automatic editing to select records with crucial mistakes such as no revenue or the record does not describe the right financial year. The selection module is used a second time after al the automatic editing is done to determine if the record is correct. This time the module detects deviant ratios and classification errors.

*2.2.2 Automatic editing*

In order to correct structural errors as well as some other shortcomings of SLICE[4] the optimisation was expanded with an editing module that can be programmed to make any desirable edit. These edits use an expert set boundary or imputation or data from previous years. For example the social security sums of the wages are often reported poorly. For the computer service industry this problem was solved by stating that if the wages are over 10.000 and the social security sum is zero, the median of the (social security sum/wages) for the industry times size class is imputed. Here the 10.000 is an expert set boundary while the imputation uses data from last year, but other options are also possible.

## 3. Macro-editing: an efficient top-down strategy for interactive micro data editing[5]

### *3.1 Introduction to macro-editing*

Micro data may still contain substantial errors after automatic editing. Additional interactive (manual) data editing is therefore needed. However, interactive editing of large amounts of micro data by an analyst is time consuming and costly and should therefore be restricted to a minimum. Ideally only micro data that have a significant influence on the final output and that are not deemed plausible, should be selected for interactive editing. Macro-editing is a means of making such an efficient selection. It follows a top-down approach where first aggregate data are considered. Only if an aggregate is found implausible, analysis will continue on more detailed underlying aggregates and finally micro data. Furthermore, only records which contribute significantly to an implausible aggregate will be edited.

Macro-editing consists of two main elements: data selection and data correction. In the remainder of this paper we concentrate on the selection process. Interactive correction of data is usually performed by an analyst with extensive knowledge of the relevant industries. After correction of incorrect micro data the aggregates are evaluated again in an iterative process until the aggregates on the output level are deemed correct. The plausibility of the aggregates can be evaluated according to the different viewpoints detailed in the next section.

### *3.2 Different viewpoints of macro-editing*

The proposed macro-editing strategy for economic data consists of the following viewpoints.

*3.2.1 Population overviews including dynamics*

A thorough understanding of the target population including dynamics is essential in the evaluation of publication data. Information is needed about the size of the population and how the population is divided into different industries, size classes, etc. What has changed since the previous period regarding births, deaths, fusions, take-overs, etc? This information can be evaluated at a micro level for influential enterprises, but also on an aggregate level for large amounts of enterprises simultaneously.

*3.2.2 Sorted tables with relevant variables or indicators*

Micro data as well as aggregate data on different levels can be shown in tables. The tables can be sorted using relevant variables or indicators such as turnover per employee in order to identify extreme and possibly implausible values. These variables and indicators are industry specific. Data are preferably investigated on an aggregate level first, after which the analyst can zoom in on more detailed aggregates or even micro data.

[4] See also “TMO-visie op het IMPECT-1 microgaafmaakproces” (in Dutch only), 2004, Hoogland, J., J. Pannekoek, T. de Waal (2004), Statistics Netherlands.

[5] See also “Analyse en interactief gaafmaken in de HEcS-keten” (in Dutch only), 2009, Frank Aelen et al., Statistics Netherlands.

#### *3.2.3 Visualisations*

The same variables and indicators can also be visualized in different graphs such as scatter plots, box plots, spider plots, bar plots, histograms, line graphs, etc. These graphs can be shown on different levels of aggregation, with the possibility of zooming in on relevant sub-aggregates or micro data.

#### *3.2.4 Statistical measures, plausibility measures and risk indicators*

Statistical measures that give insight into the distribution of the micro data on different levels of aggregation are for instance the mean value, median value, variance, inter quartile distance, minimum, maximum, skewness, kurtosis, etc. Robust measures like the median can be used for instance to give robust estimates of growth rates, which can be compared with the regular estimates. Plausibility measures give an indication of the quality of a record, for instance by comparing measured values with expected values. In our experience, plausibility measures that try to incorporate the quality of a large number of variables in a record, are problematic. The number of variables involved should be limited. A plausibility measure of a record multiplied by the influence of the record on the output level, gives a risk indicator. Risk indicators can be used in tables to order records with respect to their priority for interactive review, thus supporting an efficient interactive editing process.

#### *3.2.5 Problems in the aggregation of micro data*

Besides errors in micro data, also errors in the process of compiling aggregate values on the output level can occur. This is especially true when weighting sample data. For instance, sample weights can become extremely large or even negative when the weighting model contains too many variables, the weighting aggregates are too detailed or when large differences occur between sample response and the reference material that is used in the weighting. When extreme sampling weights occur, it is important to investigate its cause.

#### *3.2.6 Process measures*

A number of process measures give further insight into the quality of output data and the underlying records. Examples are numbers and percentages of response records, automatically and interactively corrected records, records that did not comply with certain editing rules, imputed records, outliers, etc. Comparing these measures with those of previous periods can give additional information.

### ***3.3 Evaluation of plausibility of aggregate data***

Expressing the plausibility of publication data in one comprehensive number is not feasible due to the complexity of the statistical processes of compiling the publication data. Too many factors are involved. However, it is feasible to express the plausibility of publication data by means of the above outlined viewpoints (see paragraph 3.2). In the assessment of the plausibility, besides general economical knowledge and knowledge about industry specific developments, all of these viewpoints should be included. Each viewpoint by itself is not broad enough, but together they are.

### ***3.4 Feedback to automatic data editing***

The processes of automatically editing micro data and interactively editing micro data using top-down macro-editing, should be seen as one integral editing approach. When analysts often encounter certain types of data errors in their analyses – and when it is obvious how these errors can be corrected – this information can be used to improve the automatic detection and correction rules. This feedback makes the editing process more efficient.

### ***3.5 Macro-editing tool***

The above-described viewpoints of macro-editing are currently being implemented in a tool[6]. This tool will be generic to a large extent, so that it can also be used for other statistics in the future. An important feature of this tool is that analysts (or at least the team leaders) should be able to set up their own analyses, that are often industry specific. In order to monitor the extent in which the top-down approach is followed by the analysts, relevant information will be stored about which micro data have been edited. This information can subsequently be used to further instruct the analysts.

[6] See also “Macro selection and micro editing: a prototype”, 2009, Wim Hacking, Statistics Netherlands.