STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 27
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (iv): Impact of new technologies on statistical data editing

UNI-EDIT: STANDARDIZED PROCESSING OF STRUCTURAL BUSINESS STATISTICS IN THE NETHERLANDS

Invited paper

Submitted by Statistics Netherlands[1]

Abstract: Over the last two years, Statistics Netherlands has been working on redesigning its structural business statistics. The major aim of the project is to obtain a more efficient statistical process. This is achieved through selective editing and standardization of the statistical process and its systems. In 2001 the first results appeared when a new uniform system for the annual structural business statistics was implemented. This new system has been used to process the data of 2000. An important part of the new system is the editing process, which uses the latest methodology and software that Statistics Netherlands has developed. Selective editing plays a key role. Score functions are used to select the records that contribute most to the aggregates and are most likely to contain important errors. These records are checked and, if containing errors, corrected. About 50% of the records are edited in this way. The other part of the data is edited automatically. In this paper we will discuss this edit strategy. We will also present the organizational and operational issues concerning the implementation of this edit strategy, in view of the large scale on which this has been achieved.

I. INTRODUCTION

1. Society changes and so do the demands it makes on Statistics Netherlands. Statistical information has to be made available with minimal delay and in a cost-effective way. These conceptions led to the founding of a project called IMPECT (Implementation of the economical transformation process). Within this framework, different projects were started to redesign and standardize the questionnaires (UniQuest), the logistical process (LogiQuest), the editing-process (UniEdit1), the weighting (UniEdit2) and the publication strategy (MicroLab e.g.). In this paper we will focus on UniEdit: the project which deals with the editing-process.

II. PRECONDITIONS FOR UNI-EDIT

2. One of the preconditions for the standardization of the editing process, was a harmonized set of variables and questionnaires. This was dealt with in the project UniQuest. First of all, a set of global variables with an identical definition was established to be used for all questionnaires. To take the difference in branches of enterprises into consideration, specific questions and variables were added. Furthermore, brief questionnaires were developed for small enterprises to reduce the burden of paperwork. This resulted in a situation of one extended and one brief questionnaire per branch of business, each with one or more supplementary sheets with specific questions. Compared to the previous situation, a very substantial reduction of questionnaires and variables was achieved.

3. In the LogiQuest project, the input-process was redesigned and harmonized. All incoming questionnaires are registered and entered in the same way. This results in a few databases with non-edited

[1] Prepared by Arjan de Jong (gjog@cbs.nl).

records in a *Blaise*-format. The contents of these databases are merged and exported into a *SQL-server* database, which forms the point of departure for the new editing process.

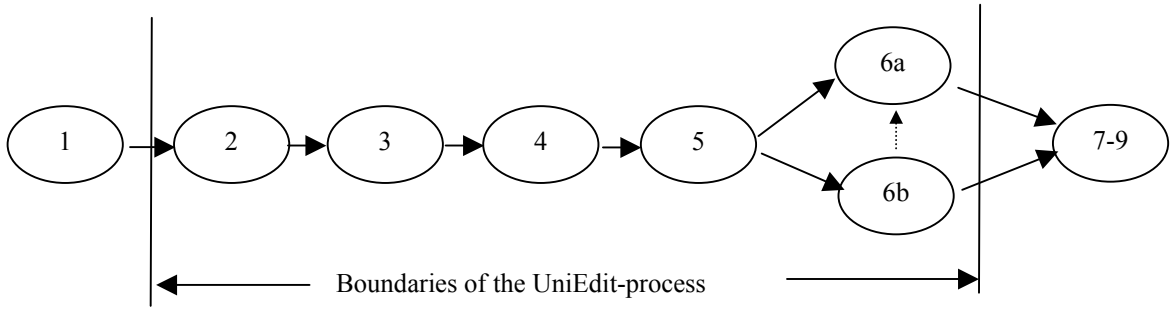
III. THE NEW EDITING PROCESS

4. The UniEdit-process has two main pillars: standardization and selective editing. Until 1999, all structural business statistics were processed in different ways: they all had their specific questionnaires, logistics, methods and systems for editing and weighting and calculation of output variables. To meet the demands of today, standardization was an absolute necessity. The same goes for selective editing. To obtain the best results with limited resources, the valuable time and energy of statistical analysts should be concentrated on records with errors that contribute most to the aggregates. The demands on less important records are limited and computerized editing is a good alternative. It is therefore essential to be able to make just the right distinction between important and less important records. For this purpose, the plausibility indicator was developed.

5. The new statistical process can be divided into different phases:

1) data-entry of non-edited questionnaires
2) error detection
3) automatic correction of obvious mistakes
4) calculation of plausibility-indicator
5) selection for either interactive or computerized editing
6) a. interactive editing
   b. computerized editing
7) imputation for unit-nonresponse and weighting
8) top-down analysis
9) storage of microdata in Microlab and generating publications

6. The UniEdit-process covers phases 2 to 6. In the diagram below, each oval represents a phase; the arrows indicate the dataflow.



A. Error detection

7. For all questionnaires a number of edit rules are defined and used in different phases of the process. In phase 2, the rules are used to detect errors. The number of violated rules per record is recorded and used in the plausibility indicator (see phase 4). In addition, the data generated in this phase can play a part in the analysis of the final results.

B. Automatic correction of obvious mistakes

8. A lot of questionnaires contain obvious mistakes. In this phase of the process, a large percentage of these errors are resolved automatically. By doing so, the workload for both interactive and computerized editing can be reduced substantially. The obvious mistakes can be classified into the following categories:

(i) **"1000-errors":** The respondent is asked to fill in the questionnaire using the unit of 1000 guilder or 1000 Euro (e.g. € 125,235 should be noted as € 125). Nevertheless, about 10% of the respondents neglect this instruction , which means an error of a factor of 1000. It is evident that such errors lead to major deviations from the estimated publication totals. Errors of this kind are detected using VAT-data, data of short-term statistics and by comparing the turnover per person-ratio with the median value of the previous year. If a 1000-error is detected, all financial variables are divided by 1000.

(ii) **erroneous negative values**: Some variables have negative values to indicate a subtraction. The minus sign however is already printed on the questionnaire. In this case, the absolute value of the variable is used.

(iii) **empty (sub)totals:** About 20 % of the questionnaires contain empty (sub)totals while one or more of the constituent variables contain values. In this phase, the (sub)totals are filled with the sum of the constituent variables. About 30% of the total amount of errors per record can be resolved by this action.

C. Calculation of the plausibility indicator

9. To differentiate important from less important records, a plausibility indicator is developed.[2] The important records are edited interactively by statistical analysts, while the remaining records are edited by computer. The pi consists of one "total indicator" and seven constituting partial plausibility indicators (PPIs), each representing a certain aspect of the questionnaire. Four PPIs (one for each part of the questionnaire) compare the level of selected variables of a record with the median level from the same NACE-size class from the previous year. In the fifth PPI, a comparison is made for several indicators (such as turnover per person employed) while PPI-six relates to the filling-out-quality (number of errors and percentage of completion) of the questionnaire involved. Last but not least, a seventh PPI is used to express the relationship with other sources, such as the individual data of the previous year and the VAT-turnover.

10. The total indicator and all PPIs are graded from 0 up to 10. A PI of 0 expresses that a record deviates much from the average record, while the figure 10 indicates a "perfect" record. The word perfect is in quotation marks: the PI indicates the plausibility and has no absolute value, a record with a PI of 10 could be erroneous while a record with a PI of 0 may prove to be correct. Records with a PI below 6 are, however, most likely to contain errors that influence the aggregate levels of publication and therefore have to be checked by statistical analysts.

D. Selection for either interactive or computerized editing

11. All records with a total indicator below 6 are selected for interactive editing by statistical analysts. The remaining records are edited by computer, with the exception of the following categories:

- records from companies with over 100 employees;
- records from small strata;
- records from branches of business which have not been surveyed before.

In all of the above-mentioned situations, records are edited interactively, regardless of the total indicator.

E. Interactive editing

12. The more important records as selected in the previous phase are edited individually. The statistical analysts use a Blaise-application in which they have the following aids at their disposal:

[2] For more detailed information about the methodology of the Plausibility-indicator, see the contributing paper of Jeffrey Hoogland

- Data of last year's annual business statistics of the same record;
- Data of this year's non-edited questionnaire of the same record (that is before the correction of the obvious mistakes);
- VAT-data;
- Turnover from the short-term statistics;
- Total plausibility indicator and all partial plausibility indicators;
- Various indicators (e.g. turnover per person employed, gross margin, gross wages per full-time equivalent);
- Error messages based on the edit rules.

13. The analyst will edit a record by resolving the errors and by editing implausible variables, guided by the plausibility indicators. A record can be finalized if the total plausibility indicator is 6 or higher and there are no errors left. In some situations however, the total plausibility-indicator will remain unsatisfactory. In these cases a record can be finalised by indicating the reason (e.g. outlier) and adding a concluding remark in a designated field.

F. Computerized editing

14. As indicated in the previous section, the statistical analyst uses a lot of information when editing a record. It is not possible to incorporate all considerations for all possible situations in a system for computerized editing. Therefore, only the less important records are edited in this way. In UniEdit, *SLICE*[3]-components, developed by the Methods and IT- Department of Statistics Netherlands, are used for computerized editing. The computerized editing consists of three phases: localization, imputation and correction.

Localization

15. In the first phase, a record is subjected to a set of edit rules. These edit rules describe requirements that have to be met by a record. There are line-edit rules (variable A + variable B – variable D = 0) and ratio-rules (variable A / variable B <= Constant), whether or not combined with a condition (IF condition THEN rule). There are however limitations: a ratio-rule consists of just one variable as numerator and one as denominator; in the conditional statement only a line-edit rule can be used. The edit rules are defined per questionnaire. The localization module establishes one or more sets of erroneous variables. The localization can be influenced by attribution of confidence weights: variables which are more likely to be correct are attributed a larger weight than other variables. The solution with the smallest sum of confidence weights is considered to be the best.

Imputation

16. For the variables that are indicated as erroneous by the localization-module, a more accurate value has to be calculated. The applied version of the imputation engine has four options to choose from: mean value imputation, ratio-regression, mono linear regression or user defined linear regression. The imputation method and predictor are defined and recorded per questionnaire per variable, the coefficients are calculated by the imputation module (if user defined linear regression is chosen, the coefficients have to be defined as well). The coefficients are calculated based on a "reference file" which contains correct records of the previous year for the same branch. The imputation however takes no account of edit rules, which means that an imputed record is not necessarily correct. Therefore, a third phase is required.

Correction

17. In this final phase imputed records are subjected once again to the edit rules. If a record is not in compliance with the edit rules, imputed variables are adapted until the record is correct. Records, for which SLICE is not able to find a solution, are passed on unedited to interactive editing (see the

[3] SLICE: Statistical Location, Imputation and Correction of Errors

discontinuous arrow from 6b to 6a in the diagram on page 2). In the present situation, this concerns about 5 % of all records.

IV. ORGANIZATIONAL ASPECTS

A. The UniEdit project

18. As stated in the introduction, UniEdit was one of several projects within the IMPECT-framework. Every part of the logistical and statistical process was redesigned, from questionnaire and variables to editing strategy and weighting. In addition, there was no system to fall back on: the structural business statistics of the year 2000 had to be processed with the IMPECT-systems. This put the project and the people involved under a lot of pressure. As a result of this, next best solutions were often chosen to remain within planning. Furthermore, it was not always possible to test as thoroughly as desired.

B. Organizational structure of Statistics Netherlands

19. Until 1999, as mentioned in chapter III, all structural business statistics were processed in different ways: they all had their specific questionnaires, logistics, methods and systems for editing and weighting and calculation of output variables. The organizational structure of Statistics Netherlands was based upon this approach. There was a number of "stovepipes": one for each individual statistic and branch of business. The knowledge *of a branch of business* was the basis for the distinction between departments. The standardization of both questionnaires and logistical and editing processes, made it possible to adapt the organisational structure to improve efficiency. In this new organizational structure, the knowledge *of a process* is of primary importance. For example, the input-process for all relevant statistics is now totally independent of the branch of business; all knowledge on logistics is consequently concentrated in one department.

20. Although the merits of the current organizational structure are paramount, a notable demerit needs to be mentioned as well. After finalizing the editing process, the data is transferred to the Analysis Department. Hence, the statistical analyst cannot see the significance of the record he is editing, which makes it more difficult to take the right edit decisions. Besides, this situation might lead to a certain degree of detachment. To prevent this from happening, excellent communication between the departments involved is essential.

V. CONCLUSIONS

21. A thorough evaluation of all UniEdit-processes and results is being carried out this year to reveal any weaknesses and highlight where improvements need to be made. At the moment, it is not yet possible to draw final conclusions. The following observations however, can be made:

- Implementing a new edit strategy on such a large scale as Statistics Netherlands has done, puts high demands on all people involved. Risks are implied, when in addition almost all other relevant circumstances (logistics, questionnaires, variables) are being changed. It would have been more manageable if either the development had taken place successively or the pressure of time had been less. Nonetheless, given the circumstances, the new UniEdit-system is a major accomplishment.

- The correction of obvious mistakes, especially the correction of "1000-errors", is of major importance. In fact, this phase produces the largest adjustment to the publication totals. The filling of empty (sub)totals proves to be useful as well: it reduces the number of errors substantially and speeds up interactive editing.

- The plausibility indicator is a powerful tool for selection of records and provides auxiliary information for interactive editing. Finding the right parameters, given the complexity of the questionnaires, has proven difficult. Further research will be necessary to optimize parameters.

- Computerized editing reduces the number of employees needed for the editing process. Using the right parameters, a lot of records can be processed in a very short time and with reasonable quality. It is nevertheless not a solution for all situations. Given the fact that editing a complex questionnaire requires professionalism, computerized editing is not able to replace all human intelligence.