WP. 18
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (ii): Implementing editing strategies and links to other parts of processing

THE EMBEDDING OF A UNIFORM STATISTICAL PROCESS

Supporting Paper

Submitted by Statistics Netherlands[1]

ABSTRACT: A uniform statistical process that uses modern methodological editing techniques has been implemented at Statistics Netherlands for annual structural business statistics and the short-term (turnover) statistics. Parallel to the redesign of the process the organisation of the Business Statistics Division has been changed. The implementation of a general unif orm process caused new internal dependencies. Formerly the departments interacted solely in terms of output and resources. Now also coordination of changes in requirements and maintenance of joined applications has to be organized. Therefore a team was formed to manage the functional requirements of the uniform process. After several years the initial organisational structure has been evaluated and adjusted to these findings. Still some difficulties remain.

KEYWORDS: Process redesign, IMPECT, organisation

I. INTRODUCTION

1. Society changes and the availability of new methods led Statistics Netherlands to start a project called IMPECT (IMPlementation of EConomic Transformation process) in 1999. This project aimed to redesign and standardize the questionnaires, and the logistical and editing processes. In the first phase the project focused on the annual structural business statistics, in the second phase short time statistics were included (see also De Jong, 2002 and 2003).

2. Parallel to the change of processes the organisational structure also changed. Until 1999 Statistics Netherlands was organized according to the statistics (branch). There were a large number of "stovepipes": one for each individual statistic and branch of business. Standardisation of the statistical processes made a more efficient structure possible. This new organisation is structured according to the structure of the processes.

3. The interfaces between different processes correspond with the interfaces between departments and the applications that are being used by different groups of users. This paper is organised as follows. In Section II we describe the new standardized and improved statistical processes. In section III the new organisation is described. In this description we focus on management of functionality. Section IV discusses the problems that occurred and the alterations that have been made to optimize the organisation. Section V describes improvements that have been implemented in the last years. Finally the persisting problems and further development are addressed in section VI.

[1] Prepared by Jeroen van Velzen (JVLN@cbs.nl).

II. THE NEW PROCESSES

4. This section gives an overview of the major changes in the statistical processes made by the project IMPECT. The starting point of the redesign was that one general process for all statistics should be created.

A. Scope

5. The first phase of the project focused on the annual structural business statistic. It included almost all branches of business (e.g. manufacturing, construction, trade, transportation, services). Beforehand the detail of the output was described (in terms of NACE codes) in order to be able to optimize the statistical methods.

6. In the second phase the short-term statistics were added to the system. *This description will restrict itself mainly to the annual statistics for practical reasons. Descriptions of (parts of) processes that are solely for short-term statistics are in italic.*

B. Process

Business Register

7. Statistic Netherlands has a central business register since 1967. This register is based on the register from the Chambers of Commerce and describes the relation between the enterprises and the fiscal units (for value added tax). The (fiscal) turnover is used as auxiliary variable in the weighing process.

Sample

8. The sample is optimized using a Neymann allocation. Using this new allocation the quality of the output of all branches of business is optimized, focussing on the new explicit description of the output. The use of panels is rejected because the representativeness of such panels was shown to decline.

Questionnaires

9. The layout and structure of all questionnaires are harmonized. A central database has been built to describe which specific questions are asked to the different kinds of enterprises (ordered by size and branch). From this database both questionnaires and interfaces for data entry and editing are generated. For the definitions of the variables the system refers to a central database with metainformation.
The letter attached to the questionnaires and the letters used to remind the non-responding enterprises are also harmonized and have been altered to adapt to the new legal status of Statistic Netherlands. The law now offers Statistics Netherlands the opportunity to impose a financial fine upon non-responding enterprises.

Central logistic system

10. A central database is used for all specific information about the respondent. In this database all contacts with the respondent are registered. Furthermore, for each enterprise the preferred method of data collection is recorded (e.g., paper questionnaire, electronic questionnaire). In this way new modes of data collection can easily be implemented.

Reminders

11. For reminding both written reminders and telephone reminders are being used. The written reminder is generated by the system. The capacity for reminding by telephone is limited, therefore there is a strong need for optimisation of the selection process. This optimisation is done manually for annual statistics. *For short term statistics the enterprises for telephonic rappel are selected using an algorithm based on size of the enterprise and the response rate of the specific branch.*

Automatic pre-editing

12. In this stage the questionnaires (now called records) are corrected for obvious mistakes. First, these mistakes and corresponding corrections are described by subject matter experts. Secondly, data that are edited interactively are used to confirm these corrections.

Calculation of the plausibility indicator

13. At Statistics Netherlands we use a selective editing approach (see Lawrence and McKenzie, 2000; Hedlin, 2003). To discriminate between important and less important records, a plausibility indicator is calculated. The plausibility indicator consists of seven partial plausibility indicators, each representing a certain aspect of a record. Four indicators compare the level of a specific set of variables with the median level in the same branch size class in the previous year. With the fifth partial plausibility indicator comparisons are made for several indicators (such as turnover per person employed and gross wages per full-time equivalent) because these indicators tend to be very stable within a branch size class. The sixth plausibility indicator describes the filling-out quality in terms of the number of errors and the percentage of completion of the record involved. The seventh plausibility indicator is used to express the relationship with other sources such as individual data of the previous year and the VAT-turnover. These seven partial indicators are combined to one overall plausibility indicator. The (partial) plausibility indicators are scaled from 0 to 10 (0 is very implausible and 10 is a perfect record). To combine the partial plausibility indicators, most weight is given to the extremely low scores. A score of 0 or 1 for one of the partial indicators causes the overall plausibility indicator to drop below 5.

14. Records that score below 6 on the overall plausibility indicator are most likely to contain errors that influence the aggregate levels of the output and therefore have to be checked by statistical analysts. For more information on the plausibility indicators used at Statistics Netherlands we refer to Hoogland (2002 and 2005). *For short-term statistics the selection of records for interactive/automated editing is less complex. The selection is focused on the total turnover. The expected total turnover is calculated based on a previous observation and the number of working days in the observed period. When the observed value differs too much from the expected value and the observed and/or the expected value, the record can have a significant influence on the outcome and the record is edited interactively.*

Interactive editing

15. The implausible records (which score below 6) and the records of large enterprises (over 100 employees) are selected for interactive editing. The other records are corrected automatically. The statistical analysts use a Blaise application in which they have the following aids at their disposal:

- Data of previous years annual business statistics of the same enterprise;
- Uncorrected data (before pre-editing and interactive editing);
- VAT-data;
- Turnover from short term statistics of the same enterprise;
- All (partial) plausibility indicators (these can be re-calculated during interactive editing);
- The set of indicators used in the fifth partial plausibility indicator;
- Error messages (inconsistencies within the record).

16. The statistical analyst is guided by the plausibility indicator. The analyst is asked to enter a comment when the plausibility indicator is still low after editing. All inconsistencies within the record are corrected by the analyst. The analyst can also label the record as an outlier. Based on the comment made by the analyst and a possible outlier label extra attention will be paid to a record in the subsequent phases of analysis. *For short term statistics the editing is focused on the comparison with previous observations of the same enterprise.*

Automated editing

17. The records that are not selected for interactive editing are edited automatically using SLICE ® components (see De Waal, 2005). This software is based on the Fellegi-Holt paradigm of minimum change (see Fellegi and Holt, 1976).

18. A set of edit rules is described per record. These edit rules describe requirements that have to be met by a record. There are balance edit rules (Variable A + variable B = Variable C) and ratio rules (Variable A/variable B <= constant), that may be combined with a condition (if .. then …). First one or more sets of erroneous variables are established. The selection of the set is influenced by assigning a so-called reliability weight to each variable; a set of variables with minimal weight is considered the set of erroneous variables. Secondly the erroneous variables are imputed. The imputation model is a linear regression model based on a dataset of the previous year. Finally, the imputed records are subjected to the edit rules. The imputed variables are altered iteratively until the edit rules are met.

19. Records for which the automated editing system is not able to find a solution are edited manually (see interactive editing). *For short term statistics the automated editing using SLICE ® is not relevant because of the simplicity of the records. Remaining inconsistencies are corrected by few explicitly programmed corrections.*

Unit imputation

20. For annual structural business statistics the non-response of large enterprises is corrected by imputation. The records are estimated using the record of previous year combined with a factor for development, which is estimated based on short time statistics. When this development factor is not available a more simple imputation method is used assuming the development factor to be 1.
In short time statistics a panel survey is used. Therefore all non-response is imputed. These imputations are based on the previous observation(s) of the enterprise corrected for fluctuations in the NACE size class concerned.

Determination of outliers

21. To objectify the weighing the determination of outliers is based on a robust algorithm. The outcome of this algorithm can be overruled by the statistical analyst.

Correction for non-existing enterprises

22. The business register contains a certain amount of enterprises that do not exist (any more). This causes a selective response. Therefore both the survey and the population are corrected for non-existent enterprises. The correction factors (called existence factors) are based on a special survey for business demography.

Weighing

23. The BASCULA® component is used to assign weights (see Nieuwenbroek, Renssen and Hofman, 2000, for more information on Bascula). This component uses VAT-data as an auxiliary variable. The VAT-data of different size classes are collapsed. VAT-data are not available for all enterprises due to timeliness and differences in definition of an enterprise by Statistics Netherlands and the definition of a fiscal unit by the tax office. Therefore post stratification on the availability of VAT-data is used.

Validation

24. In the final step of the process the statistical analyst validates the outcome in the different output cells. When this validation fails (because the analyst finds the outcome implausible) the statistical analyst can go back to the records and weights involved.

Calculating indices

25. *For short-term statistics only indices are published (average turnover in 2000 = 100). Therefore it is possible to exclude administrative, non-realistic changes in the business register. The indices estimate the growth ratio between two subsequent periods, excluding population changes due to administrative causes.*

Publication

26. A central database is filled with all records on both annual structural business statistics and short term statistics. This database is used to generate tables for Statline (the output database of Statistic Netherlands which can be visited through the Internet www.cbs.nl).

III. ORGANISATION

27. In 2000, parallel to the implementation of the new process, Statistics Netherlands was reorganized. In this re-organisation a division of Business Statistics was formed. The division is subdivided into units and subunits based on the following:

- Process: Until 1999 the organisational structure of Statistics Netherlands was divided according to the statistics (branch). There were a large number of “stovepipes”: one for each individual statistic and branch of business. Standardisation of the statistical processes made it possible to put together analogous processes of different statistics. This improves efficiency.
- Location: Statistics Netherlands has offices both in Voorburg and in Heerlen, business statistics are produced at both locations.
- Span of control: The division of business statistics contains two layers of management (departments and units). The manager of the units manages a maximum of 40 employees. When processes could not be subdivided any further and the span of control was still too large a further subdivision according to statistical output was made.

The reorganisation was accompanied by a strong deduction of staff. The number of staff working on Structural Business Statistics and Short Term Statistics was reduced from 216 to 160 fte.

A. Departments

28. The division of business statistics exists of 6 departments:

- Business registers department: responsible for the business register and the availability of the larger external registers for use in the statistical process.
- Business surveys departments (both in Voorburg and Heerlen): responsible for the data collection and editing of the survey data.
- Statistical analysis departments (both in Voorburg and Heerlen): responsible for the weighing and validation of the output, responsible for further analyses and additional publications.
- Development and support department (one department located at both locations): responsible for (management) support, research and statistical development.

B. Organisation of the business survey departments

29. The majority of the processes concerned in IMPECT are used in the departments of data collection. These departments were structured according to both process and statistics. The unit of preparation (at both locations) is responsible for the maintenance of the central logistic database and the preparation of the questionnaires. These tasks are divided between the locations. In Voorburg the maintenance of the central logistic database is located. The preparation of the questionnaires is

concentrated in Heerlen. The logistics units are responsible for survey logistics, such as sending out questionnaires, digitalisation of paper questionnaires by data entry, or scanning and the sending of reminders. Furthermore both business survey departments contain several units responsible for editing data.

C. Organisation of the analyses department

30. The statistical analysis department is not subdivided according to process because the statisticians need an overview over the last steps of the processes. The new processes are implemented in three units (1 in Voorburg, 2 in Heerlen).

D. Organisation of the administration of functionality

31. The new process uses several strongly related applications. In these applications a large set of parameters is used. To keep control over the management of this parameter set extra attention is paid to the organisation of the administrators of functionality.

32. Therefore a team of administrators has been formed. In this team the different stages of the process are assigned to different administrators: meta (questionnaires and sampling), logistics, editing, analyses, output and process planning. In the beginning the head of the Development and support department was appointed chairman of this team. He was later replaced by a project leader of the department of development and support. The team itself was formed mostly of members of the original project team (developers). They knew the ins and outs of the applications and where able to fine-tune the parameters. As the knowledge of the systems was spread in the organisation, team members were replaced by senior statistical analysts.

E. Match of new process, new organisation and administrators of functionality

33. In Table 1 an overview is given of the relations between processes, organisation and the administrators of the applications.

Process	**Which organisation(s)**	**Administrator**
Business register	Department of business register	-
Sample	Department of business register Unit preparation Heerlen	Meta
Questionnaires	Unit preparation Heerlen	Meta
Central logistic system	Unit preparation Voorburg (maintenance of central database) Unit Logistics (both locations)	Logistics
Reminding	Unit Logistics (both locations)	Logistics
Automatic pre-editing Calculation plausibility indicator Interactive editing Computerized editing	Four units responsible for editing (2 in Voorburg, 2 in Heerlen)	Editing
Unit Imputation Determination of outliers Correction for non-existent enterprises Weighing Validation	Three units of analyses (1 in Voorburg, 2 in Heerlen)	Analyses

Table 1.Relation between processes, organisation and administrators of functionality.

IV. DIFFICULTIES

34. Implementation of a new process and new applications is always a difficult task, especially when the implementation is accompanied by a cut back of staff (25 %). In this section the major difficulties that occurred are described.

A. Delegation of change management

35. A major redesign of a process needs a strong, central control and coordination during implementation. Subsequently control and coordination have to be delegated. In this latter phase the administrators of functionality were assigned. In the first year that the new processes were used still numerous (detailed) change requests in the functionality of the applications were submitted by users. Not all requests could be awarded which lead to dissatisfaction among the statisticians. Therefore unit managers claimed influence on ranking of change requests. There was tension between administrators and unit managers. As the number of statistics produced with the general application increased, the number of units involved also increased. To determine the priority of the change requests proved to be hard.

B. Expansion toward other statistics

36. As time and capacity were limited during the project, the flexibility of the applications is limited out of sheer necessity. More detailed analyses of processes of different business statistics showed the similarity between the processes to decline. In general the data collection processes are more easily extended to other statistics. The processes of analyses show more statistic specific elements. These two arguments cause the expansion towards other statistics to slow down.

C. Competence of statisticians

37. The new methods are state of the art and therefore the statisticians are challenged to learn about these new methods. Not all statisticians are willing and able to become sufficiently familiar with the new techniques. Only a small number of (senior) statistical analysts can oversee and discuss further improvement of the implemented methods. This caused both disproportional workload and risks of continuity.

D. Overview

38. Finally, centralized coordination of methods, questionnaires and output causes the statisticians to lose the overall view over all relevant processes that lead to a publication. Therefore they do not have (and feel) an integral responsibility. The commitment that follows from integral responsibility is missing, causing the incentive for maximal contribution to diminish.

V. IMPROVEMENTS

39. An internal evaluation of the organisational structure led to a reassignment of tasks to units in 2004. The most important changes are the fusion of the editing and analysis processes within units (per branch of business) and the establishment of a contact centre for telephonic reminders and a help desk for the respondents.

40. The combination of editing and analysis enables the statistician to oversee a larger part of the (statistically relevant) processes. Furthermore workload is more equally divided over the year because the same persons work on subsequent processes.

41. The establishment of a contact centre makes it possible to further optimize the efficiency of reminding. In order to guarantee sufficient response a service level agreement between departments depending on each other's effort has been drawn up. In the coming year an analogous service centre will be established for electronic data collection. All reminding activity will be concentrated in these centres. By organising processes in separate services it is possible to include additional statistics.

42. This service-structured organisation is less suitable for editing and analysis. For these processes focus is on the re-use of methods and the centralised use of metadata. Each unit is focussed on a specific (group of) statistics. Further generalisation of questionnaires and logistic processes is supported by a new project (Prodonna). In this project the applications are further improved.

VI. PERSISTING DIFFICULTIES AN FURTHER DEVELOPMENTS

43. In the coming years the statistical processes at Statistic Netherlands will be characterized by the increasing use of complex methods in large integrated applications. The developments in ICT enable us to further intensify re-use of data that forces us to focus more and more on integration. Processes are too complex to be overseen by one person or even a small group of statisticians. Especially processes, in which subject-matter knowledge is essential, require an organisational structure that combines overview and coordination with commitment and clear responsibilities for the statisticians.

44. New projects have been started for further development of the statistical process at Statistics Netherlands. The focus is on two aspects: integration of output and intensified use of register data. The integration of the output will cause new and more intensive interaction between units that publish different but related statistics. Furthermore joined use of registers will create the need for a new service: coordinated administration of registers and publications based upon these registers.

REFERENCES

De Jong, A. (2002), *Uni-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. (2005), *SLICE 1.5: A Software Framework for Automatic Edit and Imputation* UN/ECE Work Session on Statistical Data Editing, Ottawa.

Fellegi, I.P. and D. Holt (1976). A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association 71*, pp. 17-35.

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.

Hoogland, J. (2005), *Evaluation of Score Functions for Selective Editing of Annual Structural Business Statistics*. UN/ECE Work Session on Statistical Data Editing, Ottawa.

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

Nieuwenbroek, N., R. Renssen and L. Hofman (2000), Towards a Generalized Weighting System. *Proceedings of the Second International Conference on Establishment Surveys,* Buffalo, pp. 667-676.
