WP. 26
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (v): Successful strategies for implementing new editing and imputation methods

DATA VALIDATION STRATEGY IN EUROSTAT

Supporting Paper

Prepared by Emilio Di Meglio, Eurostat

I. INTRODUCTION

1. Data validation is a fundamental issue for Eurostat. Eurostat receives a large amount of data of varying nature and therefore editing and validation are the most time consuming and complex tasks to be performed. Data validation in Eurostat is highly heterogeneous according to the domain. Most of the data production processes feature independent data validation tools developed separately for each statistical process ("stovepipes").

2. The improvements of editing and in general data validation approaches have to be seen in the framework of the effort to set up a new business architecture, that should allow to pass from a stovepipe approach to a more integrated approach. This work has already started with the CVD (cycle de vie des données, data life cycle) project, that aims ultimately to provide a coherent set of concepts, metadata structures and IT tools to be applied in all statistical domains. To improve the efficiency of the validation process and ultimately the quality of the data to be disseminated, Eurostat has adopted, in the framework of the CVD, a strategy that lays on standardisation of methods and tools. In these two fields a series of actions have been undertaken or are ongoing. In this paper we will describe the specific situation of editing and data validation practices in Eurostat, the solutions envisaged and first results of the proposed strategy.

II. CURRENT PRACTICES IN DATA VALIDATION

A. Validation and quality, some definitions

3. Eurostat's mission is to provide the European Union with a high-quality statistical information service. Quality implies that the data available to users have the ability to satisfy their needs and requirements concerning statistical information and is defined in a multidimensional way involving six criteria: Relevance, Accuracy, Timeliness and punctuality, Accessibility and clarity, Comparability and Coherence. Data validation may be defined as supporting all the other steps of the data production process in order to improve the quality of statistical information. Its three main components are the following:

- Data editing – The application of checks that identify missing, invalid or inconsistent entries or that point to data records that are potentially in error.
- Missing data and imputation – Analysis of imputation and reweighting methods used to correct for missing data caused by non-response or not resolved invalid entries.

- Advanced validation – Advanced statistical methods can be used to improve data quality. Many of them are related to outlier detection since the conclusions and inferences obtained from a contaminated (by outliers) data set may be seriously biased.

In this paper we will focus mainly on editing and outlier detection, while missing data and imputation is deemed marginal in Eurostat core production processes.

4. Before Eurostat disseminates information, data validation has to be performed at different stages depending on who is processing the data:

- The first stage is at the end of the collection phase and concerns microdata. Member States (MS) and other bodies are responsible for it, since they conduct the surveys.
- The second stage concerns country data, i.e., the country aggregates sent by Member States to Eurostat. Validation has to be performed by MS and Eurostat, according to the specific situation, at this stage.
- The third and last stage concerns European aggregates (Eurostat) data before their dissemination and it is also performed by Eurostat.

The data validation process at Eurostat takes thus place in a specific context. Eurostat data editing comes very late, sometimes more than one year after the primary data collection in MS. Moreover, Eurostat usually does not have auxiliary information at unit level. The validation process in Eurostat overlaps partly with validation by the MS. In some rare cases, surveys are conducted directly by Eurostat, in this case Eurostat is also responsible for the validation of microdata. We will mainly refer to the standard case of data received from Member States.

B. Data Editing in Eurostat

5. Eurostat checks the internal and external consistency of each data set received from Member States. The steps of statistical production processes of Eurostat often contain the following:

- Ex-post harmonization of national data to EU norms.
- Data format checking.
- Re-classification of data according to the appropriate (common) nomenclature.
- Rules on relationships between variables (consistency).
- Plausibility checks of data.
- Balance checks like differences between credits and debits and mirror flows checks.
- Aggregation of items and general consistency when breaking down information (e.g. geographical, activity breakdowns).
- Time evolution checking.

6. More precisely, different kinds of corrections are applied:

Harmonization of national data - It is necessary to ensure the comparability and consistency of national data. Statistical tables for each MS can then be compiled and published based on the common Eurostat classification. To this end, Eurostat checks that the instructions to fill in the questionnaire have been followed by the reporting countries. When deviations from the harmonised definitions are detected, Eurostat reallocates national statistics according to the common classification. E.g., this involves the following corrections:

- For the country and economic zone: ensuring that the contents of each country and economic zone have been filled in the same way.
- For the economic activity: checking if all the items (sub-items) have been aggregated in the same way by Member States.

Corrections (imputation) using data from the same or another Member State

- Some variables can be corrected using other variables or functions of them (estimator imputation). Correction of data for a given zone, a given item and a given year can be performed using an average proportion involving another zone, another item and other years, with data coming from the same or another MS.

C. Data Editing in Eurostat: some examples

7. Here a few example of editing checks, collected in an internal survey of 2007 are given. For Foreign Trade data, transport statistics and Eurofarm survey:

- Checking for invalid nomenclature codes, i.e., some variables have to assume values of a given codelist (nomenclature).
- Checking for invalid combinations of values among variables.
- Detection of non-admissible values, i.e., checking if a variable is within a certain interval range.

8. For Labour Force Survey:

- Checking the variables' attributes such as data format, length, type or nomenclature codes.
- Comparison of variables to detect eventual inconsistencies.

9. The checking patterns are common to several surveys, a standardization of methods and tools is therefore possible.

D. Advanced validation: outlier detection

10. As Eurostat often has to further harmonise the data coming from Member States, outlier detection is an important step. Furthermore Eurostat is in a privileged position to detect outliers as it can compare the data across countries and discover more easily anomalies. Empirical (numerical) error detection procedures are used for this purpose, namely the construction of admissibility intervals (for quantitative variables). When a given value falls outside such intervals, a warning is issued leading to an appropriate treatment. Visual inspection of points of clouds remains a widespread basis but is cumbersome and no multidimensional and robust standard software tool is available yet. Labour Force Statistics or Foreign Trade Statistics are examples of the application of these procedures in a similar fashion to what is done in Member States.

E. Drawbacks of the current practices

11. Data validation processes, drawing on basic but sound methodology, are highly heterogeneous according to domain. The stovepipe approach of the domain specific statistics features independent data validation tools (including confidentiality treatment, outlier detection, auto-correction, calculation of indicators) developed separately for each statistical process. Customised solutions are the rule and no common standardized architecture is yet in place. Editing rules are usually frozen in the dedicated system and their maintenance is not straightforward given also staff mobility inside the organization.
More and more data flows are controlled via eDAMIS, a web-based application that allows Member States to send directly data to appropriate Eurostat databases. eDAMIS has some checks embedded and this is a first step towards the standardization of the validation process. However, the validation process in Eurostat generates a lot of information flows between MS and Eurostat. In many cases it is required in fact, that Eurostat data should be coherent with the national data. For this reason Eurostat will have to discuss errors and outliers with the Member States. The first challenge is to keep track of this process through metadata. The second challenge is to measure the real impact of this validation step on the quality of the disseminated data. The current situation pleas for a more integrated approach of data validation in Eurostat and MS by sharing editing rules and associated tools.

III. A NEW STRATEGY FOR DATA VALIDATION

12. Efficiency of ESS data validation process can be improved according to the following ideas. The first step is the standardisation and the rationalisation of the validation process inside Eurostat. Eurostat has taken a series of measures that have to be seen in the broader framework of the CVD that contribute to the creation of a more integrated architecture. The second step is to foster the integration of validation processes in MS and Eurostat, seeking the use of EU wide editing rules in MS processes and promoting the development of a common distributed architecture providing different type of services supporting the steps of the ESS statistics production process. The second step can be seen as an extension of the CVD

concept outside Eurostat. The different actions, the tools, their impact and expected impact are described in this section.

A. CVD, the Data Life Cycle project in Eurostat

13. Over the years in Eurostat, the initially homogeneous statistical production environment on central computers has evolved into a heterogeneous range of production systems in which practically every statistical domain in Eurostat has its own specific IT system to collect, validate and analyse statistical data. The CVD project aims to provide a coherent set of concepts, metadata structures and IT tools to be applied in all statistical domains. It aims at creating significant benefits, such as economies of scale by delivering generic IT tools and functionality to statisticians to create the basis for key corporate objectives, such as quality management and improved mobility of statisticians and domain managers. The CVD framework aims at standardization and improvement of editing and outlier detection through use of common tools and methodology.

14. The CVD programme encompasses many different projects that will impact statistical production processes in Eurostat (see Figure 1). The key projects are:

- Single entry point (SEP), already operational. It consists of the eDAMIS web application that supports the transmission of statistical data from Member States to Eurostat. It allows performing some simple editing checks.
- Production tools and Building Blocks, standard applications that allow performing different steps of the CVD. The building blocks concerning editing and outlier detection are respectively called EBB and ODBB.
- Reference environment, loaded with data from production databases ready to be transferred to the Internet portal.
- Dissemination tools (Internet portal)
- Metadata Handler (MH)

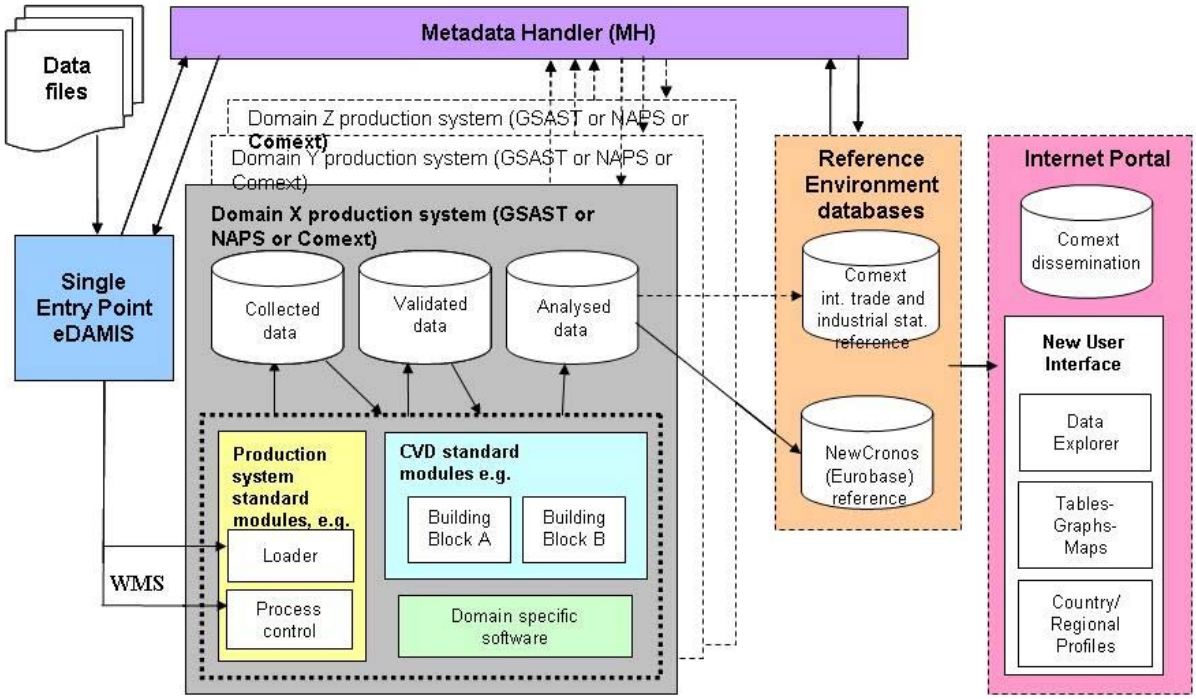

FIGURE 1: CVD architecture

B. Building blocks: ODBB and EBB

15. Building Blocks (BB) are to-large-extent customisable IT tools created for usage by the statistical applications according to the specific domain needs. The building blocks provide standard functions in a specific area. The building blocks having data validation as the primary goal are the following:

- Data editing (EBB)
- Outlier detection (ODBB)

16. The Editing Building Block (EBB) is a generic software mainly aiming at editing of statistical data. It has been designed to prepare data for loading into a reference database. EBB is based upon a flow process where data files of different formats (CSV, FLR, GESMES, SDMX-ML) enter into the system and flow through a series of steps, each performing specific operations producing a resulting file(s). EBB executes intra-cell, intra-record (horizontal) and inter-record (vertical) editing rules, optionally with reference data (lookup tables) and allows automatic editing process (once specified, editing rules can be re-launched with other or new datasets). EBB produces reports on the rule execution and alternatively allows interactive review of messages.

17. The Outlier Detection Building Block (ODBB) provides tools for detecting outliers. Outliers can appear due to errors, faults in the assumptions that generated the expected values (false outliers), or it can simply be the case that some real cases happen to be a long way from the characteristic values of the data. ODBB includes basic, statistical methods to identify univariate outliers: Hidiroglou-Berthelot and σ-gap supported by numerical expressions using absolute numbers, distribution percentiles and number of observations. Outliers can be detected for top and/or bottom of the distribution. ODBB produces reports. Output files indicated detected outliers. Multivariate tests are to be added in future.

C. Some actions to improve "validation culture"

18. Standardization of methods and tools is not sufficient if a "validation culture" is not in place in the organization. The tacit knowledge buried in the organization should become explicit and a common set of shared values should be put in place.
A series of actions have been undertaken or planned in this respect, to raise awareness on validation issues and to contribute to diffuse some concepts:

- an informal network, called the "network of methodologists" has been created, this group is composed by persons with knowledge, experience or interest in data validation related issues. This subgroup acts as a steering group for methodological activities in the field of data validation and proposes, discusses, coordinates and evaluates all activities in this domain in Eurostat.
- an analysis of all the data flows and validation processes in place in Eurostat is ongoing to propose guidelines to harmonize and make the validation process more efficient.
- the definition of a common strategy and a reference manual of validation techniques is foreseen as a result of these actions
- awareness raising actions such as the "validation day" are organized. This event, to be held in November 2009, is intended as a forum where the advantageous for the organisation practices in editing and outlier detection will be popularised and discussed by methodology, IT and production units.

19. The main difficulties to move to a more integrated process are related to the difficulty to integrate the common tools in the current processes and to the necessary organizational adaptations. Some of the processes are complex: validation process takes place at different stages and frequent revisions add complexity. It is more and more obvious that the integration of process should start with the definition of standard processes drawing on a set of standard services provided by common tools in communication with the data process. Consequently, the simplification of individual validation processes is expected to happen.

D. Extension of the CVD concepts to ESS statistics production.

20. European statistics are developed, produced, and disseminated on the basis of a set uniform standards and harmonised methods. The level of standardisation and harmonisation depends very much on the type of statistics. This way of producing statistics is, however, no longer adapted to the changing environment which calls for even more integration, improved efficiency and overall cost, and respondent burden reduction.

21. Eurostat has developed a vision for reforming the production method of European statistics and is proposing to seek a common business architecture of the ESS. The new business architecture draws heavily on the standardisation of processes and sharing their implementations. The extension of CVD at the ESS level should apply the common metadata for the ESS, the protocols for a common ESS IT (service oriented) architecture and the definition of standard processes for ESS statistics production supported by carefully chosen implementations of statistical methods.

22. A first step towards integration of data validation processes in Eurostat and MS could be to develop pre-validation services hosted on a server allowing MSs to validate their data before transmitting them to Eurostat. The service would both use and create common standards and be compatible with EBB. Any application using common standards to communicate with other applications and process handler could be shared among MSs for integration and standardization in production processes.

E. Data analysis and data validation, possible synergies

23. Eurostat is considering the use of data analysis and visualization techniques to aid production processes. Directors of Methodology of the European Statistical System supported during their last meeting this approach. Data analysis and visualization techniques can be useful in the cycle of production of official statistics: well targeted analysis techniques can in fact allow better understanding the data, improving data gathering processes and improving quality: e.g., visualization based outlier detection techniques.

IV. CONCLUSIONS

24. Data validation at Eurostat is one of the core processes. It differs significantly from data validation in Member States. Data validation is considered as a business case for fostering integration of statistical processes at the Eurostat but also at the ESS level, promoting the sharing of tools and standards. The CVD project in Eurostat has reached significant achievements so far with the development of standard tools for performing validation. However, its introduction in production is cumbersome because processes are not standardised. The next steps planned to implement are the promotion of the developed tools and their methodological improvement. The extension of the CVD concepts at ESS level is planned as a step by step approach trying to link more closely data validation at Eurostat and in Member States.

REFERENCES

EDIMBUS Project (2007). *Recommended practices for editing and imputation in cross-sectional business surveys* (http://edimbus.istat.it/EDIMBUS1/document/RPM_EDIMBUS/RPM_EDIMBUS.pdf)

Eurostat Internal Document (2007). *CVD Masterplan.*, Eurostat, Luxembourg.

Eurostat Internal Document (2005). *Introduction to data validation.* Unit B2 (Methodology and Research), Eurostat, Luxembourg.

UNECE (2006). *Statistical data editing, vol. no. 3, Impact on data quality*, Geneva.