WP.18
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (iii): Metadata and the Statistical Cycle

USING THE METADATA IN STATISTICAL PROCESSING CYCLE – THE PRODUCTION TOOLS PERSPECTIVE

Invited Paper

Submitted by Statistical Office of the Republic of Slovenia[1]

I. INTRODUCTION

1. In 2004, SORS started the action plan for the migration of statistical production from the mainframe computer to a local environment. The production tools used in the local environment offer much better possibilities for use (and reuse) of metadata. The paper will show the usage of different types of metadata in three main production environments used in the office: Blaise, SAS and Oracle.

2. In 2000 we've introduced Gentry tool, giving subject-matter experts possibilities to define the Blaise instrument (working Blaise survey specific data entry-editing application) – not by writing a code, but simply entering the parameters, describing the survey questionnaire and its data fields into the CAPI questionnaire. From this CAPI questionnaire, the survey questionnaire is generated. The "generated" metadata are later used through the process like metadata from other, not generated Blaise datamodels. Generation is based on some common in-house standards for questionnaires. It covers basic metadata (description of fields, blocks, etc.), but it is successful and efficient on a large number of surveys, and at the same time it enforces standardized processing.

3. In 2004 the Statistical Office of the Republic of Slovenia started the project of developing and implementing standard editing and imputation tool, written in SAS. The reason was first of all the growing demand for quick and efficient flow and results of the statistical process and the possibility of cost reduction. Metadata are used for the presentation of input matrices of the edit coefficients and for the presentation of the matrix. The plan is to set up such a system that these two matrices, together with the data, would be the only input for the whole system. Output metadata are quality indicators which indicate the number of values that have been changed during the process as well as the impact of these changes on final results. Through the values of variable indicators it is also possible to see in which stage of the process the data value was changed.

[1] Prepared by Matjaž Jug, Pavle Kozjek and Tomaž Špeh

4. In 2005 the new Oracle database, containing data and metadata for structural business statistics (based on several secondary and primary sources), was built. It is based on metadata rich model with classifications, source indicators, process indicators and variable definitions. The process requires a robust loading mechanism with the possibility to change data sources for existing variables and add new variables, general imputation and editing procedures, standard extraction procedures and analytical possibilities.

II. USING THE METADATA IN BLAISE APPLICATIONS

A. Metadata in Blaise environment

5. At SORS, the Blaise system is used as a general data collection and editing tool, completely covering CATI and CAPI mode of data collection, as well as all data entry and editing from paper forms (about 110 forms) and other data sources, where the microediting phase is necessary.

6. There are two general rules about metadata in Blaise (Pierzchala - Manners, 2001):
   a. all metadata defined once
   b. no data without metadata.

7. The additional Blaise capability of metadata manipulation allows generating customized data export routines and downstream data definitions from the metadata in the instrument (working questionnaire application). Thus it is possible to accomplish multiple modes of data collection (e.g. CATI, CAPI) and data editing with the same system, and export data including metadata in any way you need them, all from one metadata specification.

8. Metadata generated from the Blaise datamodel are in general the basic metadata to describe data structure and support processing. Some components are not mandatory; they depend on the application developer: if a certain part of metadata (e.g. question text) is defined in a datamodel, then it is used for generation of metadata for the next phases in a process. Only the key elements that define the datamodel are required.

B. Gentry, an in-house developed tool for high-speed data entry

9. In general, most of Blaise capabilities concerning metadata are used at SORS. But as a first step to get the survey metadata, the correct Blaise datamodel (as a source of metadata, used later in the process) should be prepared. Usually this should be done by the Blaise application developer. We prefer to have it generated, at least the major part. Currently, our central metadata repository Metis can not provide all information needed to generate the survey Blaise datamodel, and there is no general procedure available to use the metadata from Metis.

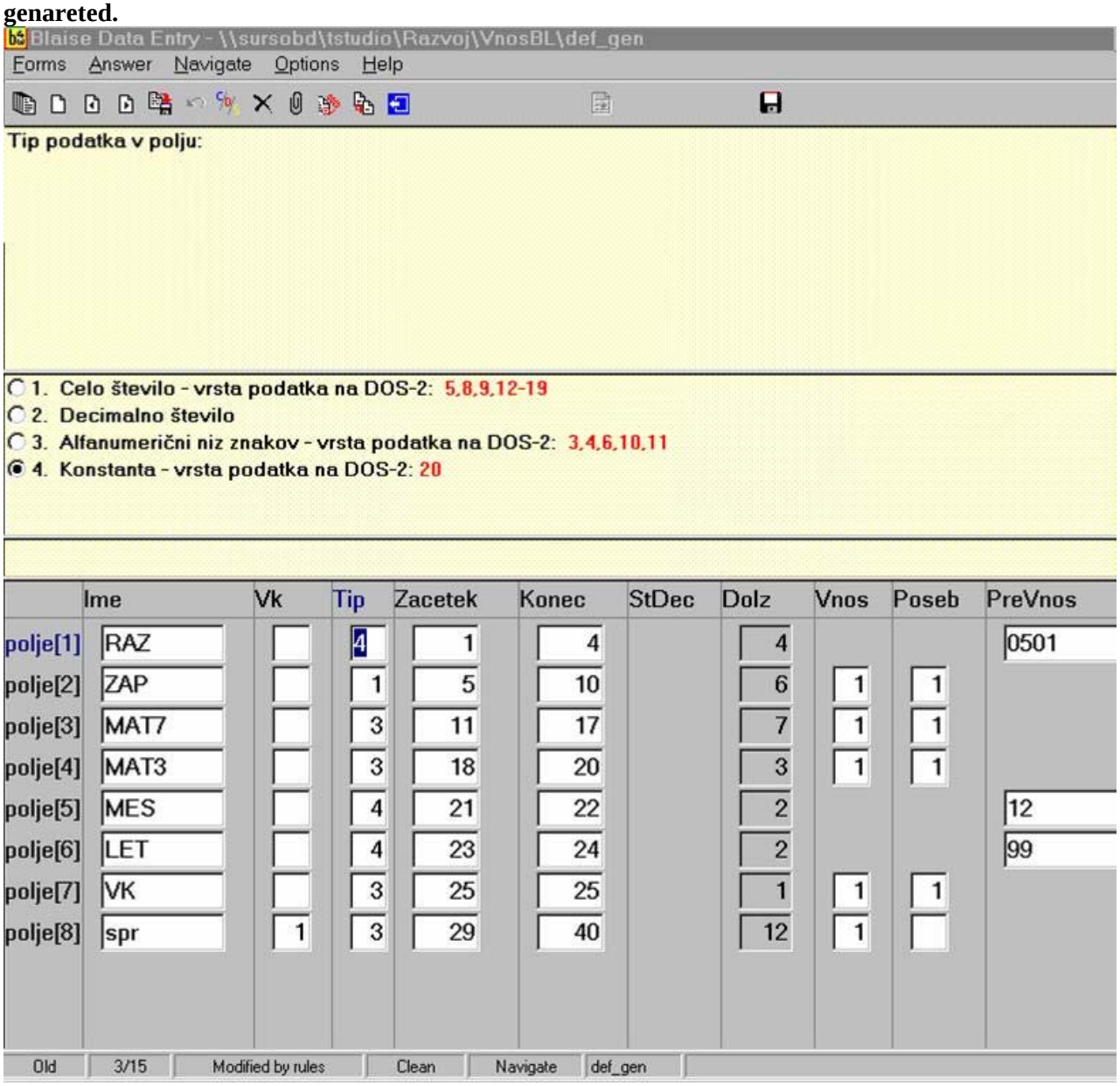
10. To bridge the absence of a structured metadata source, we offered subject-matter people possibility to generate the Blaise instrument (survey specific data entry-editing application) – not by writing the code, but simply entering the parameters, describing the survey questionnaire and its data fields into the CAPI questionnaire. From such a collected database (which is in fact a repository of basic metadata of all surveys in the system), the survey data entry applications can be generated. Used tools: Blaise, Manipula and VB.

11. This is the basic principle of Gentry, an in-house developed tool for high-speed data entry from paper forms. It generates ready-to-use applications for data entry (including the double-keying). Of course, difficult and complex datamodels can not be generated: Gentry datamodel specification is based on some common in-house standards for paper questionnaires. In the system only basic process metadata are used (description of fields, blocks, etc.), but the basic needs of the process are fulfilled. The system has been used in production for about six years and successfully covers all SORS data entry from paper forms. Using standardized and straightforward solutions it was also a good example for later developed system for data editing.

12. The system for data editing was made on the same principles as Gentry and using the same tools. For better connection between the two systems, the generator for data restructuring was developed. (Typical transformation from “form=many records” to “form= one record”, and vice versa). Again, some parameters need to be entered in a CAPI form, and upon them the new meta descriptions are generated, using Manipula (the Blaise subsystem that covers batch processes) and VB interface.

13. The most used options to prepare data for further processing (data export from Blaise) are simple ASCII or ASCIIRELATIONAL (with tables separated) files, both with generated meta descriptions for import into target systems (at SORS mostly SAS and Oracle). To support the large, frequent and changing data flows, the stronger integration of Blaise applications with other applications can be used, using BCP (Blaise Component Pack) which also facilitates access to relational databases through OLE DB. To conclude, generating metadata (although only basic, to support the processing) is one of the key features of data entry and editing systems at SORS. Although implemented on a rather basic level, it contributes a lot to efficiency and transparency of the systems, and probably most important, it enforces standardized approach to application development.

Figure 1: GEntry (Generator of applications for high-speed data entry): a part of Blaise questionnaire for survey datamodel specification. Based on values entered, the high-speed data entry application is genareted.



III. THE ROLE OF METADATA IN AUTOMATIC EDITING SYSTEM IN SAS

C. Metadata in SAS environment

14. In 2004 the Statistical Office of the Republic of Slovenia started the project of developing and implementing a standard editing and imputation system. The reason was first of all the growing demand for quick and efficient flow of the statistical process and the possibility of cost reduction. This project is part of the statistical infrastructure project, the goal of which is to provide standard statistical tools and methods for supporting the statistical process at SORS.

15. The statistical process is about relating multiple data sources and bringing them together to produce statistical data. Metadata provide the definition across data sources that make this possible. In addition, metadata enable you to trace what moved when, how it was changed, what business rules were applied, and what impact those changes might have. These are critical issues. Failure to place enough emphasis on metadata will result in problems later on; often at great cost.

16. Our system is based on two well-known methods – the Hidiroglou-Berthelot method for detection of outliers (H-B method) and the Fellegi-Holt method for errors localization and data imputation (F-H method). It has been applied in the case of two short-term business surveys: the Monthly Survey on Turnover, New Orders and Value of Stocks in Industry and the Monthly Survey on Wages.

17. This paper handles the automatic data editing process and its metadata within the statistical infrastructure, and its relationships with the other statistical processes. It considers its contribution to the management and evaluation of the processes themselves, and to the measurement of the quality of the statistical data. It was concluded that two kinds of process metadata are needed:

a. those relating to its progress through the statistical process - process metadata and
b. those relating to the options and parameters which need to be applied to the automatic editing system - methodological metadata

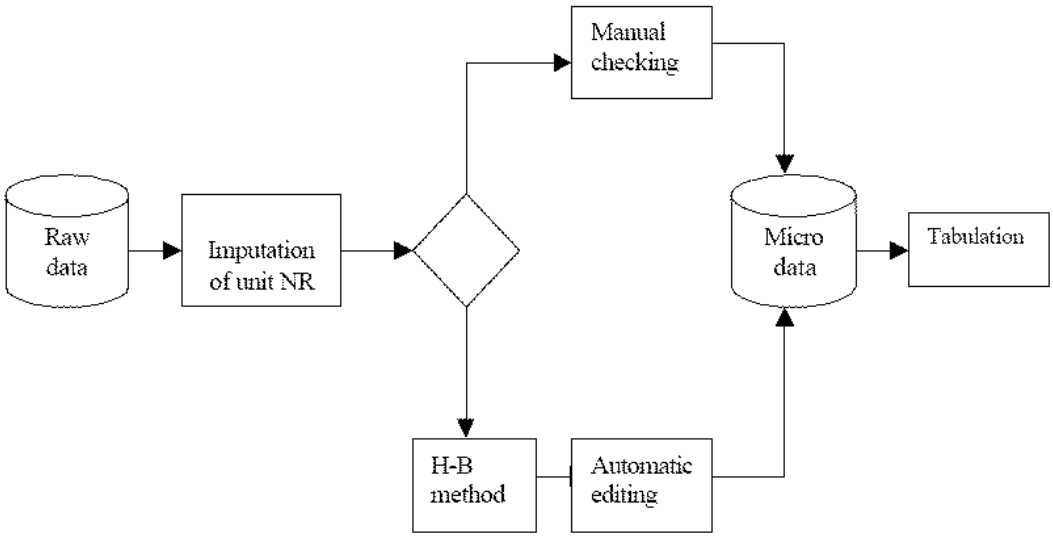
D. Process metadata

18. There should be no need for any human intervention in deciding which process should be applied to the data when, or in what order. This therefore needs to be determined by metadata accompanying the data, and interpretable by the data management systems and the statistical tools. This tool must be able to: recognize the data which are due to be subjected to editing and/or imputation; recognize which editing method should be applied, and with what parameters. In our case metadata are used for the presentation of input matrices of the edit coefficients and the matrix on the correlates. These two matrices, together with the data, are the only input for the whole system. We are planning to set the computer program for derivation of implied edits out of the basic set of edits. For the case of the two discussed surveys this work has still been done partly manually. The output metadata are quality indicators which indicate the number of values that have been changed during the process as well as the impact of these changes on the final results. Through the values of variable indicators it is also possible to see in which stage of the process the data value was changed. The quality indicators are classified into three levels. The first level denotes the data source, the second level denotes if the data were changed and the third level denotes how they were changed.

E. Conceptual metadata

19. The process of exchanging data from Blaise to SAS or Oracle is based on metadata as described in the section II. On the other hand, SAS and Oracle tables have common attributes as the names, types and lengths of variables, which simplify the data and metadata flow between SAS and Oracle.

Figure 2: Statistical process in the case of Monthly Survey on Turnover, New Orders and Value of Stocks in Industry. Data items imputed or changed in each phase are marked with proper process indicator (see annex).



IV. METADATA CONNECTED WITH THE DATA IN ORACLE DATA WAREHOUSE

F. Datawarehousing at SORS

20. At SORS we've started to build a common statistical data warehouse with the Foreign Trade database, implemented in Oracle RDBMS. During several projects between 1999 and 2005 additional databases for final statistical microdata were built in SORS, all of them based on dimensional data model and Oracle technology. Users have access with different analytical tools, for example Oracle Discoverer, Microsoft Access and SAS Enterprise Guide.

21. In the dimensional model data are structured in a form suitable for on-line analysis and tabulation. As long as the data warehouse was used only as a supplement to traditional production systems (situated on mainframe) the basic metadata (mainly classifications) were good enough. However, with the need for corporate data management solution in LAN we have started to add additional metadata into the star-schema model.

G. Metadata–rich production warehouse

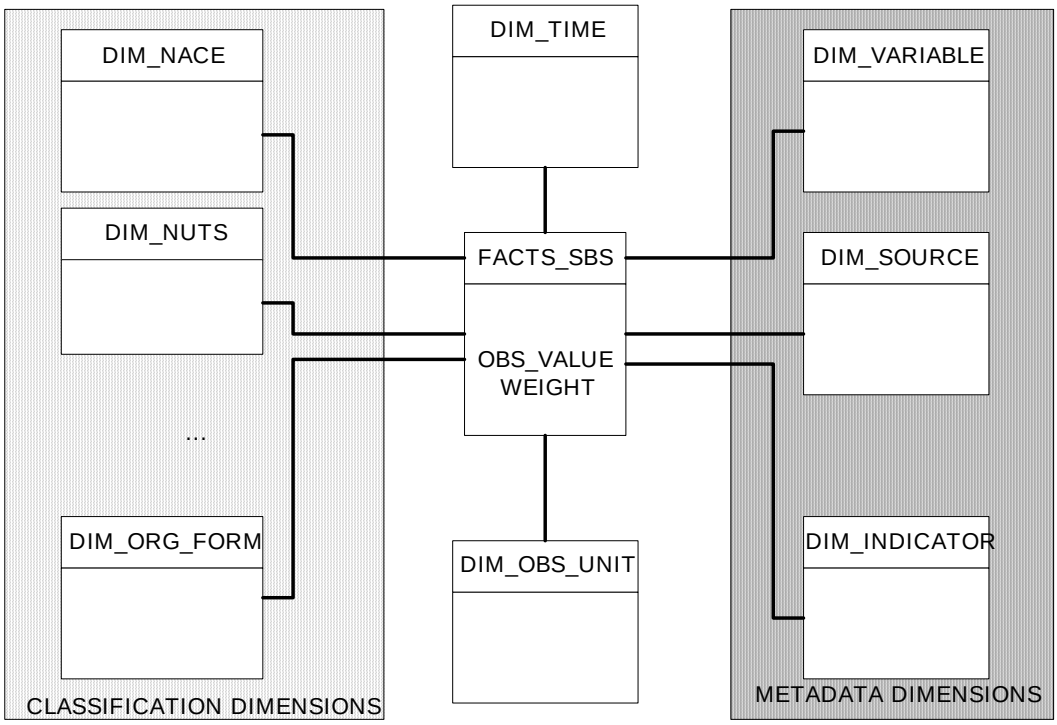
22. In 2004 with the small pilot covering five surveys in the field of waste statistics the new metadata rich dimensional model was introduced. The idea was based on similar data models in NASS and ABS (Yost, Cotter, 2003). The next project (started in June 2005) was modernisation of structural business statistics. The database, built in one of the earliest data warehousing projects as the analytical database for final statistical data, was upgraded in the productional data warehouse containing metadata linked with data.

23. The main requirements included a metadata rich model with classifications (for on-line tabulating), process indicators (describing whether observation was reported, derived, edited or imputed; indicating the type of data source and method of collection/derivation or imputation), variable definitions (with variable code, short name, description, unit of measure, etc.) and data source indicator (with the possibility that some particular continuous variable can be loaded from more than one source, depending on period or part of population). Thin fact table design was selected since in structural business statistics the described metadata elements are observed on the level of a single observation value. In the central fact table there are only the

columns for weighted and unweighted observation value and foreign keys for dimension tables populated with metadata.

24. Such a model with metadata directly linked with data in a structure suitable for analytical tools offers powerful analytical possibilities for users (Lubulwa, 2003), better possibilities for using metadata in data processing modules (for example SAS) and better control over the statistical process.

Figure 3: Conceptual data model of metadata-rich structural business statistics star schema



V. LESSONS LEARNT

25. Metadata usage in productional tools has shown many opportunities but also a lot of new challenges.

H. The role of central repositories for metadata

26. Central classification server (Klasje) and central metadata repository (Metis) are natural places to store metadata used by production tools. It is much easier to build the production process architecture if there are centrally managed metadata stores. However, using metadata stored in a central repository in production tools means that all metadata in the repository have to be exact, complete and collected and managed in a very consistent way. In the opposite case the production process will not run smoothly or will even break. The needed level of quality can be achieved using corporate business rules, good user interfaces and permanent control. In SORS classification server Klasje is used as a primary source for production tools and there are activities to implement the similar process for other types of metadata (variables, questionnaires etc.).

27. Metadata suitable for usage in production tools of course need to be in a structured form, not in plain text. That is one of the reasons that central repositories should be used instead of template-based documentation. However, not all types of metadata are feasible to put into a metadata repository. For process metadata observed on observation value level, the most suitable place to be stored is together with data. For some other types of metadata (variable definitions, classifications, etc.) it is very convenient to store it in a data

warehouse with direct connection with the data, but the place from where it should be loaded into a data warehouse are central repositories.

I. **Harmonisation of metadata concepts**

28. In order to use them in the production process, metadata concepts should be harmonized. The usual problem is that people are using metadata already but they are using local versions (for example own variable names, etc.). With the introduction of central metadata management there is a tension to keep “legacy” metadata in the same form without harmonizing it. For efficient metadata – based production process also the cultural change is needed.

J. **Technical considerations**

29. From technical point of view the possibilities for metadata exchange and system integration are better and better. For example from SAS environment it is possible to generate a database in Oracle and populate it with data without the need to write single line of SQL code. The possibilities for metadata exchange are one of the main drivers for using the commercial tools instead of developing custom applications.

VI. CONCLUSION

30. At SORS we try to follow the golden metadata rules, where possible. In Blaise we are using conceptual metadata to define the data structure and generate data-entry applications. In SAS metadata are used as an input defining the rules for automated editing and process metadata are produced automatically. Both conceptual and process metadata are loaded in the Oracle metadata-rich datawarehouse environment to be accessible for statisticians.

REFERENCES:

Pierzschala, M., and Manners, T., Revolutionary Paradigms of Blaise, Proceedings of the 7th Blaise User's Conference, Washington D.C., 2001
http://www.blaiseusers.org/ibucpdfs/2001/Pierzchala_Manners_IBUC_Paper.pdf

R. Seljak, T. Špeh: Automatic Editing System for Two Short-Term Business Surveys,
http://www.unece.org/stats/documents/2005/05/sde/wp.43.e.pdf

P. Tate: The Role of Metadata in the Evaluation of the Data Editing Process,
http://www.unece.org/stats/documents/2005/05/sde/crp.5.e.pdf

M. Yost, J. Cotter: The Impact Of A Dimensional Data Warehouse On Survey Processing Systems, Statistical Input Data Warehouse Workshop, Canberra-Murramarang, 2003

G. Lubulwa: Towards a better IDW: What Every Analyst Wants, Statistical Input Data Warehouse Workshop, Canberra-Murramarang, 2003

APPENDIX: CLASSIFICATION SCHEME OF THE PROCESS INDICATOR

The process indicator is a 3-level classification (xy.zz). The first level (x) describes whether data was reported directly from reporting unit or in different way. The second level (y) shows status of the data after the statistical processing and offers information which data was changed. The third level (zz) is reserved for the information about the method, used for statistical processing. Classification is stored in classification server and used by production tools to automatically set the proper status. Data, stored together with »indicator dimension« in data warehouse offer the complete information about processing of each data value.

1st level (x)

Mode of data collection

1 data provided directly by reporting unit
2 data from administrative source
3 data computed from original values
4 imputed data – imputation of non-response
5 imputed data – imputation due to invalid values detected through the editing process
6 data missing because the unit is not eligible for the item (logical skip)

2nd level (y)

Data status

1 original value
2 corrected value

3rd level (zz)

Method of data correction

11 correction after telephone contact
12 data reported at a later stage

Reporting methods

11 reporting by mail questionnaire
12 computer assisted telephone interview(CATI)
13 telephone interview without computer assistance
14 paper assisted personal interview (PAPI)
15 computer assisted personal interview (CAPI)
16 paper assisted self interviewing
17 computer assisted self interviewing
18 web reporting

Imputation methods

10 method of zero values
11 logical imputation
12 historical data imputation
13 mean values imputation
14 nearest neighbour imputation
15 hot-deck imputation
16 cold-deck imputation
17 regression imputation
18 method of the most frequent value
19 estimation of anual value based on infraanual data
21 stochastic hot-deck (random donor)
22 regression imputation with random residuals
23 multiple imputation

Method of computation of derived values

11 production value calculated out of the deflated turnover and change of the stocks
12 production value calculated out of the deflated turnover
13 value calculated out of the given proportion
14 turnover calculated out of the tax authorities data