

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 4 of the provisional agenda

**NATIONAL REPORT: PRESENT STATE IN STATISTICAL DATA EDITING AT
THE CZECH STATISTICAL OFFICE**

Submitted by the Czech Statistical Office¹

¹ Prepared by Václav Lehner.

I. INTRODUCTION

1. This report provides information on recent developments in statistical data processing in the Czech Republic during the last year (VI/96 - VI/97). It is a continuation of projects described in the paper "Future Development of the DataMan System" prepared for the Work Session on Statistical Data Editing, Voorburg, the Netherlands 4-7 November 1996, Working Paper No. 2.

2. In the area of design of statistical surveys, their software, data collection and processing, the activities of the Czech Statistical Office (CSO-CZ) were focused on the following topics:

- further development of the systems ProjektMan and DataMan;
- electronic data collection (electronic data entry);
- improved evaluation of data quality;
- client version of DataMan.

II. Further Development of the Systems ProjektMan and DataMan

3. For several years, systems ProjektMan and DataMan constitute the software background of most surveys conducted by the CSO-CZ. The design of survey indicators, graphic design of questionnaires and their preparation for print, definition of checking rules, sampling criteria for respondents, and defining a system of checking printed outputs are implemented using these systems. Publishing outputs from statistical surveys and other stages of data processing (e.g. imputation) are defined by independent projects, again using ProjektMan software.

4. The main benefits of this system are the possibility to organise team work through a computer network and the modularity of projects. The complete information on all projects is permanently available, including the timetables for preparatory, testing and routine work for every survey. The database on projects carried out in the past is also available. Users can generate printed documentation on individual projects at any time (ProjektMan works in a graphic mode of WYSIWYG).

5. Users can export from the system different outputs, e.g.:

- checking algorithms;
- graphic forms of questionnaires along with information on the layout of fields;
- data definitions;
- sampling criteria for the sets of statistical units;
- checking programs;
- programmes for printing tables and configurations.

III. Electronic Data Collection (Electronic Data Entry)

6. Another “branch” of the technological series ProjektMan - DataMan was introduced in the past year. This is a software designed for graphic data entry that is user-friendly intuitive, and supported by extensive help. The software is sent to respondents on floppy disks or via Internet. It is equipped with a system of checks appropriate to the questionnaire. As much as possible, the checks are identical with the checking system used at the CSO-CZ. As a whole, this software is entitled “EDE” (electronic data entry) - in the Czech language “EPV”.

7. EPV functions in MS WINDOWS. The questionnaire is presented on the screen graphically, in the same form as the printed questionnaire. This graphic layout is generated automatically by EPV based on information exported from ProjektMan, as well as information on the appropriate data structure. As long as methodological or explanatory notes for individual fields, or advice about how to complete the questionnaire are defined in the project, they can also be retrieved from EPV by clicking the right mouse button on the appropriate element of the screen form.

8. The individual checks are -just like for DataMan - either interactive (reporting errors as early as at data entry), or batch ones (working only at the complex evaluation of a questionnaire). Programmes for checking, printing, etc., are shared with DataMan and can be run by either software products at the same time. Both EPV and DataMan can work with the same libraries of programs. EPV is not equipped with translators, so that the developing environment of applications for EPV is DataMan.

9. As a rule, the lifetime of an individual EPV copy for a certain questionnaire is 1 year. When the respondent installs the software and runs it the first time on his computer, EPV will create the occurrence of questionnaires for all periods for a given periodicity.

10. While entering data, the respondent can carry out the following activities:

- switch between different periods;
- trigger data checks arbitrarily;
- display messages from the CSO-CZ, methodological instructions, explanatory notes, or help;
- write their own messages and comments for the CSO-CZ.

11. The respondent exports the data and sends a message to the CSO-CZ periodically. The data are coded and compressed. In this form, they can be sent to the CSO-CZ (over Internet or on a diskette). Some respondents also enter and check data in EPV, then print out the completed questionnaire and send it to the CSO-CZ by mail.

12. Data entered by EPV are added to data files on a given questionnaire at the CSO-CZ. Both data types are handled in DataMan in an identical manner. The only difference is that the specific message attached to an EPV record by the respondent can be displayed by the CSO-CZ staff at any time also in DataMan.

13. Because of technical reasons in CSO-CZ, it is impossible for the moment to add information concerning a particular respondent while recording EPV from Internet. However,

this will be implemented in 1998 at the latest.

IV. Improvements in Data Quality Evaluation

14. Another area that has required much work is introducing more advanced technology in the surveys proper. This applies to changes necessary for imputation and for improving data quality evaluation.

15. The former versions of DataMan did not allow to distinguish the following three types of missing values in the questionnaire:

- respondent's answer is: 0 ("zero");
- no value attached (non-response);
- the field is not completed within the framework of a given version of a given questionnaire, while another questionnaire version expects a value in this field - all the questionnaire versions are recorded in a common table (file) of the database.

The current versions of DataMan and EPV are already able to process this „triple nature of zero“ which is very important for imputation algorithms, and is also used to improve the checks.

16. Concerning the evaluation of the correctness of a questionnaire, every completed questionnaire was given a “quality code”: CORRECT/INCORRECT/ALLOWED ANOMALY. Now, each check is evaluated in this manner. The state of the checks is recorded in the database. A statistician can (e.g. after a communication with the respondent) let through any implausible anomaly, to suppress a check, or to activate the suppressed check again.

17. Every check has defined its “static” or “dynamic degree of importance”. The dynamic degree of importance is based on the weight of respondent's performance.

18. If any of the checks in a certain record are suppressed and a change takes place in some of the arguments of its algorithm, then the check is automatically activated at the next checking on a given record. It is not important where the changed value of argument of the checking algorithm can be found in the database, and whether there was a change on-line or off-line. The changes in the values of arguments of suppressed checks are monitored separately for data editing checks, and separately for batch checks. The reason is that only at batch checks the number of repeated data record checks is maintained at a certain overall quality of the record completing. This overall quality code has now six degrees instead of the above mentioned three.

19. The above technology with the “triple zeros”, and with substantially improved algorithms for data adjustment and imputation, is already in routine operation for quarterly survey of production statistics in 1997.

V. Client Version of DataMan

20. Another area where intense work is being done is the extension of DataMan to the client-server platform, where DataMan works as a client of the relational database (with an emphasis to ORACLE7). The aim is to implement the well-proved technology of data processing used at the CSO-CZ also in the CSO-CZ main database, to eliminate data conversions between DataMan and Oracle7, and to make available the performance possibilities of Oracle7 also for DataMan technology.

21. An important objective is the sharing of information sources between ProjectMan and DataMan, as well as a direct access of both products to the system of the CSO-CZ indicators and other sources.

22. In June 97, the client version of DataMan is not yet ready for operation, because the main development capacity between June 96 and June 97 was focused on work mentioned under points III. and IV. It is expected to be introduced in the summer of 1997.