

**CONFERENCE OF EUROPEAN STATISTICIANS**

**UN/ECE Work Session on Statistical Data Editing**

(Rome, Italy, 2-4 June 1999)

Topic (i): Measuring the impact of editing in various phases of statistical survey processing

**ADEQUATE EDITING – GUARANTEE OF STATISTICAL DATA QUALITY**

Submitted by the National Commission for Statistics, Romania<sup>1</sup>

**Contributed paper**

**I. General principles of editing**

1. A statistical survey asks a subset of a population questions in order to obtain information about certain groups or to obtain information about the population on a sample basis. Its purpose is to provide data regarding the country's resources and characteristics at a given point in time. These data are used to measure the existing conditions for small areas and also to plan and implement programs and activities for education, health facilities, administration and other needs. If there are missing, incorrect and/or inconsistent data, the statistics derived from these data may offer an inaccurate snapshot of a population. Therefore, before any tabulation programs are run, the data should be checked for identification of erroneous data and amended so that important data items are valid and consistent.

2. One of the definitions of editing is the following: *Editing is the process of maximising the quality of data in a limited time while minimising the introduction of new errors.* The process involves a number of sequential, interrelated activities. It is almost impossible to produce a data file that is 100 percent error free, but every effort at accuracy should be made.

3. Errors may occur during each stage of a survey such as:

- questionnaire design
- training of interviewers
- data collection
- field editing
- coding
- data capture
- edit specifications
- computer editing
- tabulation
- publication.

For this reason, continuous quality control through all phases of survey processing is extremely important.

4. Firstly, the questionnaire should have an appropriate content and structure, a clear formulation of the questions and be of a reasonable length. The graphic design of the questionnaire is also very important.

---

<sup>1</sup> Prepared by Gabriela Tihohod.

The selection from a list of answers, rather than asking for full text entries, can reduce the respondent's burden.

5. The interviewers must be properly trained and made to understand that their job is important and how the errors in the collection stage may influence other stages. The interviewers must know the way to ask the questions in order to improve the quality of the responses from individuals. The interviewers' training is often the crucial factor in survey processing and therefore should be very rigorous.

6. Many errors in data collection can be easily corrected (either by interviewers themselves or by supervisors) as long as the forms are close to the source of the data. For this reason, field editing is an important phase in survey processing.

7. Precise and detailed instructions for manual and computer coding must be determined, the coders must be trained and efforts must be made to eliminate coding errors, especially the valid but incorrect codes.

8. If the method used for data capture is keying, verification (rekeying or double keying) can reduce the errors introduced at this stage. However, an intelligent data entry system ensures that the value for each data item is within the permissible range of values. This kind of system increases the chance that the data entry operator will key in reasonable data and relieves some of the burden on later stages of the data preparation process.

9. The subject-matter specialists should write complete, concise and clear specifications for the computer editing, concerning what checks and imputations should be made in the data. These specifications should be developed at the same time as the questionnaire itself. The subject-matter specialists also must be involved in testing the edit programs, that is, in providing exhaustive test data and reviewing the outputs to ensure that all the necessary edits were included in the edit specifications.

10. Computer editing requires a good communication between the methodological group who supplies data edit specifications and the programmers. The programmers should review the specifications and work closely with the specialists for complete clarification. It is the programmer's responsibility to produce an edit program free of errors. For this purpose, the programs must be adequately and completely tested. Programme development should follow once the questionnaire is finalised in order to prevent their re-designing in case of the modification of the questionnaire.

11. Conclusions:

- The staff involved in the whole survey processing must be carefully selected and trained;
- It is better to spend more time preparing and testing each phase to avoid possible problems which may occur and which may be hard to solve later.

## II. Computer editing

12. Computer editing reduces the time for data checking and correcting errors. The purpose of editing is to make the data as nearly representative of the real life situation as possible by eliminating omission and invalid entries and by checking inconsistent entries.

13. For the population census in 1992 and the agricultural pilot survey in 1995, the Census Department of the National Commission for Statistics used CONCOR (CONSistency and CORrection) - a data editing component of IMPS (Integrated Microcomputer Processing System). The system was developed by the International Statistical Programs Center of the U.S. Bureau of the Census. The editing programs for other surveys were written by ourselves in FoxPro language.

14. CONCOR is defined by its authors as *an integrated system of computer programs which can identify and change invalid and inconsistent data being prepared for tabulation and analysis*. CONCOR is

written in COBOL language. A CONCOR program may be run after the data have been entered (batch mode) or it may be run at the time of data entry (interactively) through CENTRY, the data entry module of IMPS.

15. The CONCOR package can make imputations by different methods such as 'hot deck' and 'cold deck' techniques. Arrays (which may be either single or multi-dimensional) can be used for storing values used in cold deck imputations or to maintain the values needed for hot deck imputations. These arrays can be saved for use by the next CONCOR run.

16. The advantages of a CONCOR program are:

- it can handle a data file with multiple record types, each type being identified by a record type code;
- the user has a great deal of flexibility in developing techniques for editing and changing the data;
- the structure checks and consistency checks may be easily done because the entire questionnaire is accessible at a given time;
- it can generate comprehensive statistics about the edit tests performed and the number of changes made to the data and it can also provide reports of the statistics;
- it automatically keeps track of the number of questionnaires and records processed and makes this number available for reference;
- if the edits include changes to the data file, these changes are applied to a separate file (Output Data File) and so the original file is not modified;
- it has the ability to produce hot deck imputations;
- IMPS requires that a Data Dictionary be created which defines the characteristics of the data file to be processed. The user must define:
  - the number of different record formats on the file;
  - the code which distinguishes one record type from another and the code which distinguishes one questionnaire from another;
  - a name for each record type;
  - a name for each data item;
  - the starting position within the record and the length in characters for each data item;
  - a list of values for each data item if necessary.

This Dictionary represents an advantage because it provides a file description in a single place, the use of a common file description reduces the workload for IMPS users since they do not have to redefine the data items for each program they write and it also allows to create programs which makes the data file more understandable for a broader audience;

- with the aid of the CONCOR package, edit programs can be developed more rapidly, because the statements are more simple, the CONCOR controls the reading and writing of the records and the user has the possibility to refer to one record type only or to the entire questionnaire;
- an abbreviated form of the data item names (called the short name) is made available in order to permit easy reference to these names;
- when a data entry application is developed using CENTRY, the characteristics of the data items and the list of values accepted for each data item are obtained from the Data Dictionary and the quality of data entry improves.

17. The disadvantages of a CONCOR program are:

- it can process only one input data file at a time;
- the data file to be processed must be sequential ASCII files (that is, it may not have indexed sequential or relative organisation and they may not be database files);
- data files may be accessed only sequentially;
- the seeking of a record is time-consuming (in case of a large quantity of data);
- the entered files may contain duplicate records and a procedure must be performed for handling these records so only one is placed on the output-file;
- all data must be in character format;

- it has the capability to match the primary input file against one different file only;
- it does not handle floating point numeric data.

18. Using our own programs, the advantage is the possibility to process as many files (even indexed) when necessary and to eliminate other limitations of CONCOR, but the disadvantage is much more time spent to obtain statistics on the frequency of specific imputed values or ranges of values.

19. For more rapidity and efficiency we need software which combines the capability of the CONCOR package with other facilities.