

CONFERENCE OF EUROPEAN STATISTICIANS

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

Item 3 of the provisional agenda

**REPORT ON DEVELOPMENT AND PROGRESS IN DATA EDITING IN THE  
ITALIAN STATISTICAL INSTITUTE (ISTAT)**

Submitted by the Italian National Institute of Statistics <sup>1</sup>

---

<sup>1</sup> Prepared by Giulio Barcaroli and Antonia Manzari.

## **I. THE CURRENT SITUATION: DATA EDITING METHODS AND TECHNIQUES USED**

1. The Italian National Institute of Statistics (ISTAT) conducts annually 218 surveys, among which 138 are characterized by a dominance of quantitative variables (mainly business and administrative surveys), while the remaining 80 collect mainly qualitative data (surveys on individuals and households).

2. As for the characteristics of data editing procedures, a great difference can be found between the two types of surveys. Editing procedures are almost completely automated, with few exceptions, and generalized software (namely SCIA and DAISY) (Riccini et al. 1995) (Barcaroli, Venturi 1992), based on the Fellegi-Holt methodology (Fellegi, Holt 1976), has been applied to the most important surveys on individuals and households: we can cite the Population Census of 1991, the Labour Forces Survey, the Multipurpose Surveys on Households, the Survey on Households Expenditures, the Panel Survey on Graduates and the Survey on Marriages. The editing procedures of remaining surveys are still automated, but based on the old deterministic approach.

3. However, in the field of business surveys, data editing procedures are often of the interactive type, or mixed (partly automated and partly interactive), as greater care is put in editing operations concerning enterprises. Until now, for business surveys no generalized software was available. A number of experimental applications have been carried out, involving macroediting techniques and the use of GEIS (Kovar et al. 1991), but at present no advanced technique or software is used.

## **II. HARDWARE AND SOFTWARE ENVIRONMENT**

4. Until 1995, the EDP system in ISTAT was characterized by two different and scarcely communicating environments: the mainframe and the PC. The only linkage between them was the 3270 emulation capability of PCs, that were not even linked in LANs. As for the software, the only DBMS was ADABAS (installed on mainframe, not on PCs), while SAS and SPEAKEASY were the statistical packages (both operating also in this case only on mainframe).

5. Since 1994, a decision was taken to change this situation, switching from the mainframe to a departmental solution based on UNIX machines. The current situation concerning hardware is the following:

- the mainframe is still operating, but its importance is decreasing: many applications are being moved to the UNIX environment;
- UNIX machines are being massively introduced in ISTAT: recently more than 400 IBM RISC 6000 have been acquired, and others are planned to be acquired in the near future;
- PC platforms are also being strengthened by acquiring new units (near 200) and by linking them in LANs together with UNIX workstations.

As for the software, SAS is now available on every platform, while ORACLE has been chosen to substitute ADABAS, also in this case on every system.

### III. PROJECTS UNDER DEVELOPMENT

6. The main target of the different projects under development is the change in the general philosophy of the traditional editing procedures applied in some residual surveys on individuals and households and in almost every survey on enterprises.

#### *Business surveys*

7. As the automatic probabilistic approach has been introduced in the most important surveys on households, priority has now been put on the application of advanced techniques and methods in the field of business surveys. Among them, we can mention the surveys that have been chosen to be completely redesigned:

- Survey on the Structure of Agricultural Enterprises;
- System of Enterprises Balances;
- Sampling Survey on Small Enterprises;
- Survey on Value Added of Enterprises.

8. The Survey on the Structure of Agricultural Enterprises involves some 91,000 enterprises and is based on a rather complex questionnaire. The basic choice is to apply GEIS in order to develop a probabilistic editing procedure substituting the previous deterministic one.

9. The System of Enterprises Balances is a survey conducted annually and which collects numerous data (more than 200 variables in the questionnaire) from 50,000 enterprises employing 20 workers and more. In this case, statisticians are considering the possibility of applying macroediting techniques (that have been experimented in the past) jointly with the use of the current deterministic automatic procedures.

10. The Sampling Survey on Small Enterprises is an annual survey collecting data from near 70,000 enterprises employing up to 19 workers. The idea is to use GEIS together with techniques of selective editing.

11. The so-called “quick” Survey on Value Added is a survey of about 7,500 enterprises that employ more than 100 workers, in which timeliness plays a dominant role. For this reason we are considering applying advanced CASIC techniques, together with macroediting methods that allow to limit the time required by both phases of data collection and control.

#### *Individual and households surveys*

12. As mentioned earlier, the probabilistic approach has been adopted in the most important surveys on individuals and households, by using SCIA. However, a number of surveys still adopt traditional automatic editing procedures based on deterministic approach. Among them, the highest priority of development of new procedures concerns:

- the Survey on Births (near 400.000 records every year);
- the Survey on Deaths (about 600.000 records every year).

13. Obviously, generalized software based on the probabilistic approach will be tested and used to develop the editing procedure of the Population Census in 2001. For this Census, advanced technologies for Optical Data Entry and Automatic Coding will also be tested during the planned Pilot Survey (October 1998).

#### **IV. PLANS FOR FUTURE DEVELOPMENT**

14. One of the most important targets of ISTAT in the field of data editing is the availability of generalized software to provide the most correct answers to the following problems related to automatic editing:

- treatment of quantitative variables;
- treatment of qualitative variables;
- joint treatment of both qualitative and quantitative variables.

Moreover, software is required to handle logic constraints on different units belonging to a given unit of a hierarchically higher level (*inter-units control*): this need is strong especially in the case of surveys concerning individuals and households.

##### ***Quantitative variables***

15. With regard to quantitative variables, we mentioned that we have planned to introduce GEIS in the most important business surveys. Until now, some experimental applications have been carried out (for instance in the Survey on Occupation and Wages in Enterprises).

##### ***Qualitative variables***

16. As for qualitative variables, we planned to strengthen the capabilities of SCIA, with particular regard to the following two points:

- the quality of the imputation;
- the feasibility of the probabilistic approach.

17. Concerning the first point, we planned to improve the quality of the imputation phase by integrating SCIA with a new technique for the search of the donor based on the concept of minimum mixed distance (Abbate 1996), in order to overcome limits that have been found in many experiences. This methodology allows to choose a donor, for any variable that must be imputed, by computing a mixed distance function in a given set of candidate records. The distance function is defined by weighting a set of elementary distance indicators, each of them computed accordingly to the typology of any variable. The weighting of the distance is made by considering as weights the values given by the  $\chi^2$  test between the variable to be imputed and all the variables statistically associated to it. For any variable to be imputed, the search of the donor is based on the computation of the distance function between the erroneous record and any candidate donor record, i.e. all

records with valid value for the variable to be imputed. The record with the lowest distance is selected and the value found in the corresponding field is imputed.

18. With regard to the second point, we know that the feasibility of the probabilistic approach it depends strictly on the number of edits that can be explicitly defined: if this number exceeds a certain limit, the generation of the complete set of edits with the existing algorithms is not possible (Fellegi, Holt, 1976). We are following the Bureau of the Census researches and developments in this field with great interest (Winkler, Petkunas 1997).

### ***Inter-units control***

19. Finally, as for the problem of the inter-units control, we are planning to test the New Imputation Technique from Statistics Canada (Bankier et al. 1994), primarily to verify its capability to deal with this problem, but also in general to assess the performance of this approach for treatment of errors in qualitative variables, in comparison with the Fellegi-Holt approach implemented in our SCIA system, and to evaluate the possibility to apply it for joint editing and correction of both qualitative and quantitative variables. To carry out the test we have requested the cooperation of the Statistics Canada team. Experimental applications should concern data from the last Population Census (1991) and the next Pilot Survey for the Population Census (October 1998).

### **References**

- Abbate C. (1996) “La completezza delle informazioni e l’imputazione da donatore con distanza mista minima - Il prodotto RIDA (Ricostruzione delle Informazioni con Donazione Automatica)”, ISTAT internal document.
- Bankier M., Fillion J.-M., Luc M., And Nadeau C. (1994) “Imputing Numeric and Qualitative Variables Simultaneously”, Proceedings of the Sections on Survey Research Methods, American Statistical Association, 242-247.
- Barcaroli G., Venturi M. (1997) “DAISY (Design, Analysis and Imputation System): structure, methodology and first applications”, Statistical Data editing Methods and Techniques, vol 2, United Nations Statistical Commission - Conference of European Statistician - Statistical Standards and Studies No 48.
- Fellegi I.P., Holt D. (1976) “A systematic approach to edit and imputation”, *Journal of the American Statistical Association*, vol. 71, pp.17-35
- Kovar J.G., Macmillan J.H., Whitridge P. (1991) “Overview and Strategy for the Generalized edit and Imputation System”, Working Paper n. BSMD-88-007E/F, Business Survey Methods Division, Statistics Canada
- Riccini E., Silvestri F., Barcaroli G., Ceccarelli C., Luzi O., Manzari A. (1995) “La metodologia di editing e imputazione per variabili qualitative implementata in SCIA”, ISTAT internal document.
- Winkler W. E.. Petkunas T. F. (1997) “The DISCRETE Edit System”, Statistical Data editing Methods and Techniques, vol 2, United Nations Statistical Commission - Conference of European Statistician - Statistical Standards and Studies No 48.