STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 21
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Cardiff, United Kingdom, 18-20 October 2000)

Topic III: New techniques and tools for editing imputation

EDITING AND IMPUTATION IN EUROSTAT

Submitted by Eurostat[1]

Contributed paper

I. The Research Programmes

1. Eurostat Unit A4 monitors a number of projects which are researching into statistics and which are funded by the Commission. There are essentially two ways that these projects happen. The first is where the Commission has a particular interest in a topic, proposes the research, and then issues calls for tender for contractors to carry out the research. These are known as SUP.COM (activities of scientific and technical SUPport of a COMpetitive nature) projects. The other type is where researchers have to propose their own projects, the proposals are then evaluated by independent experts and those which are considered to be of sufficient quality receive a contribution towards the projects' costs from the Commission. Such projects are currently being funded under the EU's present (the fifth) and previous (fourth) Framework Programmes on research. The projects funded under the 4th Framework Programme are known as the DOSIS (Development of Statistical Information Systems) projects, and those funded under the 5th Framework Programme are known as the EPROS (European Plan for Research in Official Statistics) projects.

2. There are two projects directly researching editing and imputation methodology. These are AUTIMP (AUTomatic IMPutation software for business surveys and population censuses) which is a DOSIS project, and EUREDIT (The Development and Evaluation of New Methods for Editing and Imputation) which is an EPROS project.

3. AUTIMP is the smaller project. It was initially envisaged as having 15 months duration, but this has recently been extended by six months (this is still short compared with most projects, which are either of 24 or 36 months duration). There are five partners in the project (projects have to be proposed by international consortia, consisting of at least two partners from at least two different countries, but most proposals contain at least five partners), four of which are National Statistical Institutes and the other is a University.

4. The AUTIMP consortium considers that often in business surveys, too much effort is spent in correcting minor flaws in the data, and that imputation of erroneous or missing fields should be as automated as possible. The project is therefore developing innovative imputation software for business surveys and population censuses, for both numerical and categorical data. The project is progressing well. At the end of March there was a review meeting, where a review of some existing imputation software was presented, and the software developed, in the project, to impute categorical data was compared with

[1] Prepared by Harald Sonnberger and Nick Maine.

existing software. The project reviewer attended the meeting and her report concluded that the project is making good progress and the deliverables so far received are of a good standard.

5. EUREDIT is a much bigger project than AUTIMP. It will last for three years and has twelve partners, including four of the five partners from AUTIMP. Thus EUREDIT can be considered as a continuation of AUTIMP, but is wider in its scope. It will begin by a review of editing and imputation software currently available (including the software produced by AUTIMP) for data arising from the social sciences (household surveys, business surveys, population censuses, panel surveys etc.) and establish best practices among these reviewed methods.

6. The project will then develop and test new editing and imputation methods. These new methods will use new developments in statistical theory and computer science. In particular, advances in computing capabilities have made possible the application of the more complex statistical modelling techniques. A new comprehensive software package, complete with documentation, will be produced incorporating the successful new methods which have been developed.

7. An important difference between EUREDIT and AUTIMP is that it is intended to exploit commercially the software resulting from EUREDIT. To this end, two commercial software companies are included in the consortium. They are Numerical Algorithms Group (NAG) from the UK, and Insiders GmbH from Germany.

8. This direct exploitation will involve NAG taking the software developed during the project and producing commercial standard software along complete with accompanying documentation and support materials. The resulting system can then be incorporated into NAG's existing and future statistical software products. In addition, the software can be sold directly to companies creating their own in-house software systems.

9. Insiders GmbH plans to implement the most appropriate imputation algorithms evaluated by EUREDIT in a financial risk management system that it is currently developing. Inclusion of algorithms that best utilise historic information in an efficient manner is a vital step for improving the scope and precision of risk analysis that can be performed. This will significantly improve the market opportunities for the Insiders product as there are currently no other commercial risk management systems that utilise such sophisticated imputation methodologies as those investigated in EUREDIT.

10. Indirect exploitation will arise through providing the expertise and basic software to other software producers enabling them to incorporate the techniques into their software products and services.

11. The project started on 1 March this year, and so far no review meetings have taken place.

12. AUTIMP and EUREDIT are the only two projects researching editing and imputation methods, but editing and imputation are important steps in any data analysis. Editing and imputation are included in some capacity, therefore, in many other projects. A SUP.COM project (lot 14 of SUP.COM 95 'Seasonal variations') developed new methods and produced software for time series analysis. One of the new methods and corresponding program produced is called 'TRAMO' (Time series Regression with ARIMA noise, Missing observations and Outliers). Hence, this project, which was proposed by the Eurostat because it considered that its existing time series software was obsolete, developed techniques for imputing missing values in time series. This imputation software has since been used in two other DOSIS projects on time series analysis. These are FORCE4 which was looking at forecasting, and TESS which is researching automatic seasonal adjustment.

13. Another SUP.COM project (lot 23 of SUP.COM 98: Research on the usage of new technology for the control of data quality in the transmission system) has as its objective the generation of validation

rules in the process of data exchange. The aim is to generate two sets of validation rules for data which are to be transmitted from National Statistical Institutes of Member States to Eurostat. These will be rules to be implemented in the Member States before transmission, and rules which can only be implemented by Eurostat after transmission. These validation rules will include methods for editing the data and imputing missing values.

14. CHINTEX is an EPROS project which started earlier this year. It will investigate the effects, on data quality, of changes in survey methodology of three countries who are included in the EC Household Panel, and also it will look at other quality aspects of the ECHP. The project's 'Description of work' makes specific mention of the outputs from AUTIMP and EUREDIT, which will be used in this investigation of the general data quality.

15. Two other EPROS projects which started earlier this year, and will include editing and imputation methods are X-STATIS and VL-CATS. The X-STATIS project will develop an 'expert-system' computer package for data analysis. This will guide the user through his data analysis and provide advice on the most suitable analysis techniques for the data, and on the interpretation of results. The VL-CATS project will develop a web site to provide Internet based training in official statistics. It will include many modules, each one containing training for a particular aspect of official statistics.

II. Work ongoing within Eurostat

16. As well as monitoring research projects on editing and imputation, there is also other work carried out within Eurostat (in particular in Unit A4) in this field. Most of the editing and imputation carried out on data supplied to Eurostat is done by Member States, but some is also done within Eurostat, for example in the Continuing Vocational Training Survey.

17. Because most of the editing and imputation is done by Member States, the methods used must be harmonised between Member States before Eurostat can produce meaningful data at the EU level. For this reason editing and imputation has been addressed in the Working Group on the Assessment of the Quality of Statistics. This Working Group is run by Unit A4 of Eurostat. A project was tendered by the Unit, and work has been carried out by a consultancy firm on 'Editing and Statistical Imputation Methods'. The purpose of this project was to test software and methods, used by official statisticians, for editing and imputation.

18. Arising from this project, and in collaboration with colleagues from Statistics Canada, two papers were prepared for the meeting of the Quality in Statistics Working Group held in November 1999. These papers were 'A functional evaluation of edit and imputation tools' and 'Estimating variance in the presence of imputation'.

19. In September 1999 a Task Force was set up to analyse procedures for producing estimations in Eurostat. The Task Force, under the leadership of Unit A4, developed a questionnaire which was sent to all Eurostat Units. The questionnaire was designed to obtain information about the domains where adjustment and estimations are regularly carried out. The results show that adjustment and estimation procedures are carried out for most domains in Eurostat, and in particular imputations for missing data are carried out for a number of surveys. These include, for example:

i) The Community Innovation Surveys (responsibility of Unit A4),
ii) Regional GDP and regional Unemployment rates (responsibility of Unit E4),
iii) The European Community Household Panel (responsibility of Unit E2), and
iv) The Continuing Vocational Training Survey (responsibility of Unit E3).

20. Reports have been compiled on the methods used for editing and imputation within Eurostat, on these surveys. Some editing and correcting of data are also carried out by the Member States. It is also hoped that new methodologies developed, in particular by the AUTIMP and EUREDIT projects, will in future be incorporated into the procedures used by both Eurostat and the Member States.

21. However, other work is progressing within Eurostat via the Quality Working Group and other Task Forces. This work within the Eurostat groups is more about dissemination and implementation of new methodologies and best practices, rather than research into new methods.

22. A recommendation from the Quality Working Group was to set up a Task Force on variance estimation. A paper has now been written, proposing a mandate for this Task Force. A number of aspects of variance estimation were identified by the Working Group, and these are listed in the proposal for the Task Force. The first of these needs is 'to improve the existing guidance on estimating variance in the presence of imputation'. Hence the proposed Task Force is not specifically for variance estimation with imputed data, but this is recognised as a very important part of the Task Force's role.

23. Also the Quality Working Group proposed that discussion groups on methodolgy should be set up to look at areas where the quality of statistics could be improved. The members of the Working Group were asked for suggestions as to which areas should be included. A number of suggestions were made, one of which was to look at the quality of data editing.

24. Hence editing and imputation are taken very seriously by Eurostat. On the one hand, research into the fields is actively supported, and it is hoped that research results will be taken on board by Member States and also within Eurostat. On the other hand, a number of Eurostat groups (which include NSIs from Member States, and also other countries) are looking at editing and imputation along with other interdependent topics with a view to establishing and disseminating best practices and to disseminating information about improved methodologies as they become available.