STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 13
English only

CONFERENCE OF EUROPEAN STATISTICIANS

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

Topic II: Propagation of knowledge to users

DATA EDITING METHODS AND TECHNIQUES: KNOWLEDGE TO AND FROM USERS

Submitted by ISTAT, Italy[1]

Invited paper

I. A FUNCTIONAL MODEL FOR DISSEMINATING KNOWLEDGE IN THE DATA EDITING FIELD

1. The number and the variety of statistical production processes characterising any National Statistical Institute (NSI in the following) generally determines an intensive and diversified demand of specialised statistical and technological knowledge coming from the Production Units (i.e. the staff responsible for surveys). In this context, an important objective is to ensure the dissemination inside the NSI of all the available knowledge on data editing methods and techniques. In particular, all the available information about the best solutions (from the point of view of costs and quality) for particular classes of problems have to be disseminated to the statisticians responsible for surveys. This requires the management and optimisation of information flows to and from users (i.e. the statisticians responsible for the surveys).

2. A *functional model* based on a given *data editing method life cycle* is designed to ensure that knowledge and expertise are as widely shared as possible within the NSI. In that model, a fundamental assumption is that a centralised unit in charge of the search and the definition of the best methods does exist.

3. We define a *data editing method life cycle* as the following sequence:

i) *analysis of demand*: for any class of homogeneous statistical information production processes:
   - identification of actual and potential needs (requirements);
   - analysis of problems and limits due to the current existing procedures, if any;

ii) *search for best solutions* already available: in the market, in the academic environment, in other NSIs;

iii) if best solutions are not available, *research and development of new methods and techniques*;

iv) *software acquisition* (if already available) or *development* (otherwise);

v) *testing of the selected method*, by applying it to selected situations, representative of the entire class;

vi) *evaluation*, and, if pNSItive,

vii) *generalised dissemination* of the method to the entire class of production processes.

The need for information exchange arises in steps i) and vii). In step i) the flow is mainly from internal users to methodologists in the central unit. On the contrary, there is an oppNSIte flow in step vii).

---

[1] Prepared by Orietta Luzi and Antonia Manzari

4. In the following paragraphs we will analyse more in depth some of the above-mentioned phases: in section II some aspects of the first phase of the functional model are discussed; section III deals with the search of the best solution for a particular class of surveys; in sections IV and V some aspects related to the development and the evaluation of the selected solutions are analysed, respectively. Section VI describes some approaches that seems to be appropriate for disseminating the acquired knowledge about the evaluated methods to the entire class of surveys. Section VII contains some concluding remarks.

II. Analysis of demand

5. In any NSI there are several kinds of statistical information production processes, characterised by different features and objectives. Different classes of survey processes can be defined depending on:
- target population,
- characteristic of investigated phenomena,
- survey organisation,
- available resources and time,
- statistical objectives and kind of produced figures,
- data processing features.

6. Generally speaking, the common problem of any of these situations is the *optimisation of the overall survey process and, more specifically, of the data processing phase, in terms of quality, cost, time, respondent burden*. In many NSIs there exists a central unit in charge of collecting and organising the demand of methodological and/or technical solutions coming from users (i.e. the statisticians responsible of the surveys), whose main goal is studying, evaluating and providing the "best solutions" for each class of problems. In other words, all the actual (and potential) requirements have to be gathered and all the possible solutions have to be evaluated in order to find (classes of) best solutions for that requirement.

7. In particular, the identification of operational, technical or methodological solutions to the problem of designing best data E&I strategies, can be simplified if surveys are grouped in subsets (or 'clusters') that are homogeneous with respect to some classification criteria. For example, clusters could consists of statistical survey processes similar in terms of target population (business or household surveys), type of surveyed variables (quantitative or qualitative), survey typology (censuses, periodic sample surveys, panel surveys, administrative surveys, etc.), survey organisational features (mode of data collection, mode of data-entry, available budget, technological and human resources, timeliness requirements, and so on). This makes it possible to build a sort of *classification tree* in which pre-defined criteria are used as branching rules and each node contains survey processes similar with respect to all the criteria considered in the higher levels. An example of a survey process classification tree is shown in figure 1, where the *Business Survey* root node generates several more specific nodes at lower levels corresponding to particular sets of surveys identified on the basis of some simple criteria.

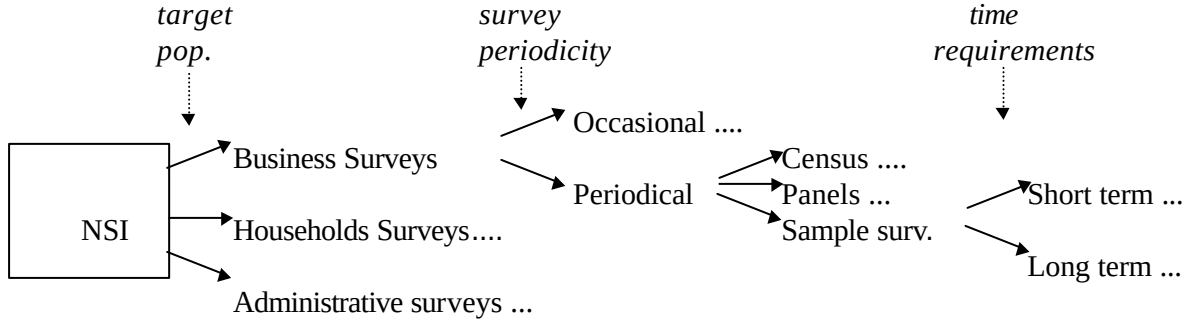

Figure 1 - An example of structured tree of survey processes

8. Even if the improvement or the re-design of the E&I strategy of a specific survey implies in any case the development of *ad hoc* data processing flows, the task of identifying possible solutions can be simplified if the above-mentioned survey and methods classification structure has been defined.

III. SEARCH FOR BEST SOLUTIONS

9. The general scheme described in the previous section also makes it possible to classify the already available or known E&I methodologies and techniques on the basis of an evaluation of their usefulness and applicability to the specific context of each node. This classification facilitates the search for the best solutions for each class of survey process, in the sense that the set of possible solutions is restricted to a particular subset of alternatives. It should be noted that, in the above tree structure, some of the possible methodological or technical solutions could be 'inherited' from a higher level node to one or more of its lower level nodes: for example, the use of generalised automatic systems[2] for numeric variables can be considered appropriate in any subsequent node of the general *Business surveys* class (see figure 1). On the other hand, editing and/or imputation techniques requiring the use of historical information on the same respondent unit are peculiar to the *Panel business surveys* node; similarly the use of macro or selective editing criteria, that implies follow-up activities, are more appropriate in *Short-term business surveys*, where timeliness and costs are the most important production constraints.

10. It should be emphasized that a 'best solution' for a given survey or class of surveys generally consists of a *set of combined single 'best solutions'*, referred each to a particular step or to a specific problem constituting the whole E&I process. In this sense, in order to identify an overall best solution for a given node, it is required to analyse each step of its peculiar processing flow in order to define the corresponding (set of) best method(s). Therefore, in the following, we will denote with 'best solution' a given set of best methods or software.

11. In any case, it is also possible that, in the choice of the best solution for a given survey belonging to a certain class, the constraints represented by costs, time, human and available technological resources may be really strong. In extreme situations, these limits could determine even the impossibility to adopt the best solution identified for a given class to a survey belonging to that class.

12. Once the best solution has been found, further problems arise relating to its actual availability. In particular, it is necessary:

- to verify if either the best solution has been already developed as consolidated methodology, or an original theoretical development is needed;
- to verify if either the identified solution is already included in some generalised or *ad hoc* software, or its implementation has to be planned (inside or outside the NSI);
- to verify if either the eventually available software including the solution is already available in the NSI, or it has to be acquired from outside,
- to assess if either the solution has been already tested, or an experimental activity has to be designed,
- to evaluate if either the solution is easy to transfer to users, or its dissemination requires that training activities, documentation production and other activities have to be planned and performed.

13. As for the first two points, if the identified solution has not been already developed and/or implemented in any software, an evaluation of the costs and resources required for these activities has to be done. In particular, if a useful software exists but has to be acquired from outside, the above-mentioned evaluation represents a benchmark for the costs and resources needed to obtain or buy it. In the Italian National Statistical Institute (ISTAT in the following) there have been very different experiences in this context. For example, to satisfy the general demand in the *Household surveys* area for an automatic system in order to improve quality, timeliness and costs of final results, the use of

[2] Several generalised software handling economic variables have been implemented by various NSIs: GEIS (Kovar et al. 1988) by Statistics Canada, CHERRYPI (Ton De Vaal, 19??) by Statistics Netherlands, AGGIES (Todaro,1999) by the U.S. NASS, SPEER (Winkler, 19??) by the U.S. Bureau of the Census.

automatic generalised systems implementing the Fellegi-Holt probabilistic methodology (Fellegi et Holt, 1976) was identified as best solution. In that situation, the development and implementation of that software (called SCIA, *Automatic System for data Check and Imputation*) (Riccini et al. 1995) has been carried out with existing ISTAT resources (already available budget and human resources). The same process characterised the life cycle of two other ISTAT automatic systems: the software RIDA for missing data integration based on the lowest distance donor imputation method (Abbate, 1997), and the software ESSE (*Editing Systems Standard Evaluation*) (Luzi and Della Rocca 1998; Manzari and Della Rocca, 2000) for evaluating the quality of E&I procedures. A different life cycle characterised the acquisition of the software GEIS (*Generalised Editing and Imputation System*) developed by Statistics Canada and dealing with numerical, continuous and non-negative variables. Recently, the demand for methodological and technical solutions coming from the *Business Surveys* field, and regarding the problem of dealing with large amounts of stochastic non-influential errors, has rapidly increased. Also, in this case, the best solution has been considered an automatic generalised system implementing the Fellegi-Holt methodology. But in this case, for organisational, costs and resources reasons, it was preferred to acquire such a software instead of developing it with internal resources.

IV. DEVELOPMENT OF THE SOLUTION

14. When an identified solution has not already been implemented in any existing software, either developed inside or available outside the NSI, the only chance is to develop it. Statistical software development requires performing the following steps:

- Algorithm identification,
- Resource allocation and cost evaluation,
- Implementation,
- Training to users,
- Maintenance.

15. When dealing with computational problems, the knowledge of mathematical methodology, then of the algorithm, is a fundamental prerequisite. The algorithm has to be fully defined, as well as input and output data. Once the algorithm has been identified, its optimisation from a computational point of view, is also a crucial aspect to consider: software must be not only reliable, but also highly performing in terms of time and resource allocation. This phase requires strong and productive relationships between methodologists/statisticians and software developers.

16. Software development requires a careful planning and evaluation of resources (in terms of people, environment and time needed to develop the software).

17. Software implementation consists of three main phases: *design*, *code implementation* and *performance test*. The design of the application requires the definition of technical specifications: target platforms (Unix, Windows, etc.), type of users interface (graphically, by command line, etc.), format and structure of both input/output data and diagnostic reports tracing specific E&I actions. The code implementation requires the availability of expert programmers and development environments (machine, programming languages and compilers, debugging tools, etc.). The development phase implies also that developers write user manuals (user guide, installation guide, readme files, etc.). Once the code has been developed and debugged by the programmers, a suitable test phase needs to be performed by users. This test is aimed to evaluate: programme quality (detect residual bugs), requirements satisfaction, performance (computational time, system resource used) and ease of use. Based on the feedback from testing, an iterative reworking of the code is generally performed.

18. In order for users to start using the new application effectively, training is necessary, including theoretic and practical (hands-on) sessions.

19. Finally, future maintenance of the programmes also has to be guaranteed. The need for maintenance comes from: residual bugs in the code, interaction conflict with additional applications

installed later, operating system upgrade. Maintenance is also related to the implementation of small programme enhancements solicited by users. Maintenance also requires a careful "skill transfer" stage to people in charge of it, in case they are different from developers. To facilitate maintenance of an application code, all the tasks of the development process as well as all the technical characteristics of the code itself are to be properly documented.

V. EVALUATION OF THE SOLUTION

20. When a solution has been developed and/or implemented in a software, testing of its performance and evaluation of its impact on the survey process is demanding for every search of "best solution" process. The evaluation process is, in fact, an essential support for choNSIng or discarding an E&I method or strategy for its application in a specific class of surveys.

21. Comparison of different solutions requires a metrics definition, in other words, standard evaluation criteria have to be defined and proper indicators have to be computed for each solution. A universal definition of quality of E&I methods is a difficult task because of the numerous aspects of quality. Generally speaking, a main distinction is made between accuracy and efficiency characteristics. Quantifying the *accuracy* of the E&I procedure means to provide a measurement of the closeness between the actual (true) values and the output determined by the investigated procedure. The output of an editing method is the classification of each observed value as correct or erroneous, while the output of an imputation method is the new assigned value. Whatever the output (classification result, individual data items, frequency distribution or parameter estimates), the computation of quality indicators of E&I methods requires comparisons among *actual* data set (true values), *raw* data set (collected values) and *clean* data set (edited and imputed values). Generally speaking, the availability of *actual* values corresponding to *raw* values is quite rare because of the high cost for carrying out careful re-interviews of respondents (professional interviewers, computer assisted interviewing, reconciliation of current answers with previous ones, etc.). In case of non-availability of true data, it is possible to arrange *actual* and/or *raw* data by a simulation approach. Instead of simulating both artificial true data and corresponding raw data, ISTAT strategy is based on the use of production data as true data, and therefore only the simulation of a *raw* data set is needed. In experimental applications, *actual* data are obtained as a result of processing a set of observed data according to an E&I procedure (straightforward and less expensive solution), while corresponding *raw* data are obtained by inserting artificial errors in *actual* data. To perform the simulation step the generalised ESSE software has been developed (Luzi and Della Rocca, 1998). ESSE performs controlled generations of artificial *erroneous values*, according to predefined error generation models (the ones most commonly occurring in the phase of the compilation of the questionnaire and in the data entry phase), and provides reports containing accuracy indices to assess the capability of the E&I process in localising errors and restoring the true values without introducing new errors in data. Accuracy indices are computed on the basis of the number of detected, undetected and introduced errors. An example of such indicators is given in Manzari and Della Rocca, 2000.

22. For the sake of clarity we point out that the only comparison between a *raw* data set and *clean* data set does not provide indicators about quality of the E&I method in terms of its accuracy. It just allows to measure the effect of E&I method on *raw* data and gives insights into the quality of the data collection procedure and, therefore, of the collected data set. An example of indicators to evaluate the quality of the collected data set is given by the quantitative information provided by the Italian Information System for Survey Documentation (SIDI) (Fortini et al., 2000).

23. Quantifying the *efficiency* of the E&I procedures means to provide a measurement of costs reduction in terms of *time* of data processing, *technical* (hardware and environment) and *human* resources required in applying the given E&I method.

24. Further information useful for the evaluation of the impact of an E&I method on survey process are its *reproducibility*, its *flexibility* to changes (in number of variables, codes, and edit rules), and the automatic *availability of standard documentation* of the E&I actions (diagnostic reports).

25. Finally, in relation to the evaluation issue, it is recalled that, when in a periodic survey a new E&I procedure replaces a traditional one (because of changing in organisational aspects or in resources availability, or because a new methodology suggests a solution better than the current one), the comparative evaluation of the quality of the new procedure against the traditional one should be followed with an evaluation of the impact of the new procedure on the time series of events. The impact measurement requires processing a set of raw data according to both the previous and the new procedure, and it is generally performed in terms of distances between estimates computed from the two sets of clean data.

VI. STATISTICAL E&I KNOWLEDGE DISSEMINATION

26. This phase is one of the most critical of the entire life cycle, because at this stage all the collected information about the identified best solutions, their effectiveness, their usefulness, their requirements and implications have to be transferred to users. In particular, the main objective here is not only to transfer the acquired knowledge, but also to allow the users be able to evaluate all the consequences produced by the possible use of that particular E&I technique in its survey (in terms of quality of data and organisational impact).

27. There are different ways to fulfil the above-mentioned information requirements. The most effective ones can be considered the following:

(a) developing a *knowledge base* containing all the available information on the particular methods, techniques and software representing the best solutions; in particular it has to be stressed to what type of data and processes they can be applied, and what requirements must be considered (for example in terms of environment and resources);
(b) preparing and providing documentation of experiments and past experiences in the application of a given method: quality indicators and estimates of cost should be available in an information system for the documentation of surveys (SIDI for example);
(c) in order to allow the user to apply the method, user manuals should be available, and training courses should be organised on a regular basis (*indirect assistance*); in any case, *direct assistance* given by the methodologists in implementing first applications is the best way to disseminate knowledge (under certain conditions), as a form of *training on the job*.

28. Among the possible ways of implementing a *knowledge base* on existing available E&I methodologies, the following seems to be the most effective ones:

- producing internal documentation like methodological and/or technical manuals describing the selected E&I approach, method or software, its field of application, its technical characteristics, possible advantages and disadvantages in its usage and so on. In general, this documentation can consist of methodological monographic volumes dealing with specific subjects in the case of an E&I method or technique, or it can be represented by *user manuals* and/ or *application guidelines* in the case of software or generalised systems. All these materials have to be organised in such a way that users could easily access to it, in order to guarantee the higher level of its diffusion and sharing;
- producing *statistical guidelines* dealing with the process underlying the design, implementation, test and application to statistical survey data of E&I methods and techniques. These manuals have to point out the theoretical and operational steps to be performed, the essential factors and the major distinctive elements to be taken into account in building an E&I strategy;
- developing an Intranet site, dedicated to the specific area of data E&I, structured in such a way that all information (internal and external references, related topics, meetings, etc.) related to new solutions is available and easy to find. In this context, the creation of a *discussion group* could facilitate the sharing not only of knowledge but also of problems and related solutions regarding the use of the various best methods.

29. The work of sharing all the acquired information on new E&I methodologies and approaches for a given class of surveys can be greatly facilitated by documenting and disseminating experiments and past working experiences performed using these methodologies on surveys belonging to that class. In this way,

it is possible to highlight not only the peculiar characteristics and properties of any selected technique, but also to provide users with important additional information, allowing a more precise evaluation of the advantages and disadvantages of using it in terms of quality of results, costs, timeliness, impact on data and survey organisation. For each experience, documentation should be available in a standard format in order to allow the evaluation of the E&I method by analysing its performance on surveys the method was applied to. The availability of quality indicators and cost estimates could be useful, particularly if stored in an information system for the documentation of surveys. This is the case of the already mentioned SIDI system developed by ISTAT, providing users with standard quality indicators on the E&I phase related to a number of surveys.

30. A useful way for disseminating knowledge about new E&I solutions (editing methodologies, imputation techniques, generalised software, outlier detection approaches and so on) consists in supplying *training courses* on these specific topics. In general, these courses are planned to be held more than once, in order to allow as many people as possible to attend them. Generally speaking, as these courses have very specific objectives in terms of discussed subjects, they have a particular target, for example people with some specific background and experience, or subject matter involved in particular survey areas (e.g., business, households or administrative surveys). Training courses in the area of E&I have to be not only *information processing courses*, where only theoretical and/or technical knowledge about the discussed methodologies, techniques or software is provided to the participants, but also *experience processing courses*, where the users have to play a key role. In other words, the course should be the result of the continuous interaction between participants and teacher (generally a methodologist), in the sense that the users expert contribution can have an impact on the degree of accuracy assigned to the discussion and the development of each topic dealt with during the course.

31. An important feature of training courses is their capability of generating a *feed-back* process from the users to the teacher, in order to both verify the course quality in terms of its contents and structure, and evaluate its impact on final users and survey processes. In particular, information could be collected from participants in order to evaluate the course usefulness in terms of:

- comprehension degree of discussed topics by the participants,
- degree of users autonomy in experimenting the acquired methods,
- potential impact of integrating the discussed techniques in survey processes,
- potential demand from production processes by monitoring the number of users interested in adopting the new methodologies or tools in their surveys.

32. Another important characteristic of training courses on E&I solutions is the portion of time reserved to practical applications and experimentation of the considered method and/or software, allowing the user to better understand and acquire knowledge on it. By adopting this pragmatic strategy, it is also possible to anticipate in this phase an important amount of knowledge transferring work that otherwise will have to be performed as a *training on the job activity* (also called *direct assistance*), i.e. in the context of working co-operation activities devoted to the improvement or the re-design of the survey E&I processes. In this case, a given amount of time and costs has to be spent by subject matter people in acquiring the new E&I solution, in order to be able to eventually introduce it in the survey process. In any case, the direct assistance provided by the methodologist to the survey statisticians in the form of training on the job can be in some situations the best or the only way of transferring knowledge in this area. This is the case, for example, of initial design or re-design of E&I strategies for complex or strategic surveys (like population or business census and other important exhaustive or sampling surveys). In these cases, an important initial investment in terms of planning, implementing, testing and evaluating the overall E&I strategy is needed. In this context, the methodological contribution is not only possible, but necessary in all the phases of the E&I strategy building process, and the user training on the newly introduced techniques and/or tools become an essential activity incorporated and distributed in the overall co-operation process.

33. Two examples of different approaches in disseminating knowledge, deriving from ISTAT experience, are represented by the processes of knowledge sharing, in the case of two generalised software: SCIA and GEIS. In the case of SCIA, with its first applications performed on very important

and strategic surveys (1991 Census of Population, Labour Forces survey, Multipurpose survey, Household Expenditures survey), the training of users on that software has been performed very often by direct assistance provided by methodologists of the Methodological Studies Office in designing and developing the overall data E&I strategies. A different approach has been followed in the case of GEIS: in order to disseminate as much basic knowledge about that system as possible, specific training courses, addressed only to subject matter working in the *Business surveys* area, have been planned. In order to stress not only the advantages (for example, in terms of quality of final results, completeness, possibility of monitoring each data processing step, and so on), but also the limits of the software, a large amount of practical applications on experimental data were performed during the course. As feed-back, after the course, the participants were asked to perform an evaluation of the system in terms of:

i) its potential usefulness in their operational and statistical context,
ii) its potential overall impact on their survey processes,

in order to make a first evaluation of the future potential demand of direct assistance coming from that area.

VII. CONCLUSIONS

34. The proposed *functional model* for disseminating the best data editing methods and techniques inside National Statistical Institutes is characterised by a strong interaction between the users (i.e. the statisticians responsible of the surveys) and the methodologists working in the centralised unit that are in charge for identifying and analysing those methods. This holds particularly in the *analysis of demand* and *knowledge dissemination* phases, where the information flows to and from users has to be optimised. In the other phases of the data *editing method life cycle,* critical activities are represented by the *evaluation* steps of the identified best solution in terms of:

- the quality of produced results and costs related to its implementation or acquisition from outside,
- the overall impact of its introduction in the statistical production processes.

35. Another critical point relates to the identification of the best way of transferring the acquired knowledge on the best solution to users, i.e. to allow the users to know:

- if the method is suitable to their particular survey,
- the advantages in terms of costs and/or quality in using the new method in their survey,
- how to apply the method.

36. Different solutions have been presented and discussed, that can be simultaneously adopted in order to improve the knowledge dissemination process. Among others, the production of a *knowledge base* on the selected methodologies or software and the *training* of users by specific courses or by direct assistance can be considered the more appropriate in terms of effectiveness and dissemination extent.

References

Abbate C. (1997) “La completezza delle informazioni e l’imputazione da donatore con distanza mista minima”, ISTAT, *Quaderni di Ricerca*, n.4/1997.

De Vaal, T. (1996) “CherryPi: a computer program for automatic edit and imputation”, Report presented at the *Work Session on Statistical Data Editing of the United Nations Statistical Commission and Economic Commission for Europe*, Voorburg, 4-7- November 1996.

Fellegi I.P., Holt D. (1976), “A systematic approach to edit and imputation”, Journal of the American Statistical Association, Vol. 71, pp.17-35.

Fortini M., Scanu M., Signore M. (2000) “Measuring and analysing the data editing activity in ISTAT Information System for Survey Documentation”, in *ISTAT Essays* n. 6/2000, 17-29.

Kovar J.G., MacMillian J.H., Whitridge, P. (1991), “Overview and strategy for the generalised edit and imputation system”, Statistics Canada, Methodology Branch.

Luzi O., Della Rocca G. (1998) “A Generalised Error Simulation System to Test the Performance of Editing Procedures”, *Proceedings of the SEUGI* 16, Prague 9-12 June 1998.

Manzari A., Della Rocca G. (2000) "A generalised system based on a simulation approach to test the quality of editing and imputation procedures", *in ISTAT Essays* n. 6/2000, 83-103.

Riccini E., Silvestri F., Barcaroli G., Ceccarelli C, Luzi O., Manzari A. (1995) "The editing and imputation methodology for qualitative variables implemented in SCIA", ISTAT Internal Report.

Todaro T. (1999) "Overview and evaluation of the AGGIES automated edit and imputation system", Report presented at the *Work Session on Statistical Data Editing of the United Nations Statistical Commission and Economic Commission for Europe*, Rome, 2-4 June 1999.

Winkler W. E. and Draper L. R. (1997) "The SPEER Edit System", in *Statistical Data Editing, Volume 2*, U. N. Statistical Commission and Economic Commission for Europe, Geneve, Switzerland, 51-55.