WP. 3
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

WORK SESSION ON STATISTICAL DATA EDITING
(Madrid, Spain, 20-22 October 2003)

Topic (i): Development and use of data editing quality indicators

EVALUATING, MONITORING AND DOCUMENTING THE EFFECTS OF EDITING AND IMPUTATION IN ISTAT SURVEYS

Invited Paper

Submitted by ISTAT, Italy[1]

I. INTRODUCTION

1. In large-scale surveys conducted by National Statistical Offices (NSOs), the treatment of non-sampling errors represents a critical problem because of its impact on the quality of final results. There is no generally adopted formal definition of editing, depending on the goals of the editing operation (Granquist, 1995). We adopt the UNITED NATIONS (2000) definition: “an editing procedure is the process of detecting and handling errors in data, including the definition of a consistent system of requirements, their verification on given data, and elimination or substitution of data which is in contradiction with the defined requirements. Imputation is a procedure for entering a value for a specific data item where the response is missing or unusable”. In this paper, we will refer to Editing and Imputation (E&I) as the integrated set of actions and procedures aiming at producing complete and coherent datasets by identifying and eliminating non-sampling errors from statistical data.

2. Measuring and documenting the effects on data of any data processing activity performed during the survey production process, including E&I, has become a mandatory requirement in many NSOs.
As relating to E&I, the importance of gathering information during this data processing phase has been widely recognised (Granquist, 1997b). In this context, required information relates to the quality of the E&I process (its suitability with respect to data problems, process costs and timeliness), the data accuracy (amount of errors, their characteristics, their possible sources), and the E&I performance (its statistical impact on data).

3. This information is generally acquired through different evaluation and assessment activities at different stages of the E&I life-cycle (Granquist, 1997a; Di Zio et al., 2001): *before* the application of E&I methods, for testing their quality in terms of their suitability with respect to the specific survey quality requirements; *during* E&I, for monitoring performance aspects (e.g. time and resources spent, impact on data and error characteristics) and tuning the procedure; *after* the data processing, for documentation and survey management purposes.

4. In the area of evaluating, monitoring and documenting the effects of E&I, during last years ISTAT researchers worked mainly in two directions:

[1] Prepared by Giorgio Della Rocca (dellaroc@istat.it), Orietta Luzi (luzi@istat.it), Emanuela Scavalli (scavalli@istat.it), Marina Signore (signore@istat.it), Giorgia Simeoni (simeoni@istat.it).

a. identifying appropriate sets of statistical measures (quality and performance indicators), either at micro (elementary data) and/or macro (marginal and joint distributions) levels, for evaluating the quality of E&I or assessing its impact on data distributions and relations. A great deal of research activity has been performed in these areas (among others, see Whitridge, 1999; Madsen, 2000; Nordbotten 1997 and 2001; Cirianni et al., 2001). In the area of evaluating the quality of E&I procedures/methods, the EUREDIT project[2] established a general framework in the evaluation field (Charlton, 2003; Chambers, 2001). Exploiting and integrating the EUREDIT research activity, some standard indicators associated to specific quality criteria have been defined at variable level;
b. providing survey managers with the *Information System for Survey Documentation – SIDI.* SIDI is aimed at documenting and assessing the quality of survey production processes (including E&I) and data quality through the integrated management of metadata and standard quality indicators (Brancato et al., 1998). The assessment of quality and the standardization of quality reporting are central problems not only at NSO level, but also in an international perspective, particularly in the context of the European Statistical System (Lyberg L. et al., 2001, Eurostat, 2000).

5. The activity was then concentrated in developing the generalized tool *IDEA* (*Indices for Data Editing Assessment*) described in this paper. Through the computation of different measures, IDEA allows survey managers to carry on the evaluation analysis under some specific contexts: the evaluation of the quality of E&I procedures; the assessment of the statistical effects on survey data of E&I activities; the production of standard quality indicators for SIDI. While the last two contexts imply the computation of standard indicators through the comparison of *raw* (original) and *final* or *clean* (edited and imputed) data, in the first context these indicators are computed through the comparison between *true* and *final* data.

6. The paper is structured as follows: in section II the problem of evaluating the quality and the effects of editing and imputation processes is discussed. The identified solutions are described as well as the indicators implemented in IDEA. Section III focuses on the standard documentation of E&I processes and on the evaluation of the overall impact of E&I on survey data as managed in the SIDI system. Issues related to the implementation of the SIDI system and the relationships with IDEA are also presented. In section IV the main operational characteristics of the software IDEA are illustrated. Concluding remarks are in section V.

II. EVALUATING THE QUALITY AND THE EFFECTS OF EDITING AND IMPUTATION PROCESSES

7. E&I methods are generally evaluated for two main purposes:
   a. measuring the quality of E&I methods in terms of their capability of correctly identifying and restoring “true” data in sampling units;
   b. assessing the E&I statistical effects on data.

It is obvious that the first type of evaluation implies the knowledge of “true” data for each unit: a low cost consuming approach allowing this situation is based on the use of the simulation approach (Schulte Nordholt, 1998; Manzari et al., 1999), also adopted in the EUREDIT project for the comparative evaluation of competitive E&I methods. This kind of evaluation is typically performed in the *E&I design and test* phase, for assessing the suitability of a given strategy for a specific application/problem.

8. The second kind of evaluation is typical of those phases of an E&I process (*process monitoring and tuning, analysis of E&I results*) in which it is important not only to measure the modifications produced on statistical properties of data, but also to verify that the process meets the expected quality and cost requirements. Possible problems in data and data processing can be identified during E&I by *monitoring* its performance and comparing it with expected results. In this way, the process efficiency can be improved during the data processing itself by modifying the appropriate parameters. Once E&I has been completed, its impact on raw data is generally assessed through an *analysis of E&I results*, in order to both documenting the process and data characteristics, and planning possible future improvements of the survey process.

[2] The EUREDIT Project was funded under EU Fifth Framework research program (www.cs.york.ac.uk/euredit/).

9. In order to build up a general environment in which both evaluations a) and b) were possible, we started from the research work done during the EUREDIT project, in which the following performance criteria were adopted for evaluating the quality of E&I methods (Chambers, 2001):

- preservation of elementary values;
- preservation of marginal and joint distributions;
- preservation of aggregates;
- preservation of relations.

10. It is obvious that all criteria assume different meanings depending on the evaluation purpose. For example, in the evaluation context a), the preservation of values criterion has to be interpreted as the capability of recovering the true value for each item either missing or erroneous, while in case b) it simply measures the amount of changes produced by the E&I procedure. In the latter case, this information is useful for example, when our aim is to preserve data coherence and completeness while minimizing the amount of data changes, like in case of treatment of random errors.

11. The relative importance of evaluation criteria varies depending on the investigation objectives. The individual accuracy of data could be not required if the survey objective is to publish parameters estimates for the investigated phenomena: in these cases, the estimates accuracy could be the only quality requirement needed. On the contrary, in the case that micro data have to be provided to end users, the preservation of values could become the most important criterion to be met. The distributional accuracy can assume a relevant role in the case that the distributional assumptions (univariate or multivariate) on observed variables are to be analysed or taken into account in subsequent statistical analyses. In general, the relevance and priority of the performance criteria mainly depend on the investigation objectives, the investigation characteristics, the nature of the analysed variables.

12. Starting from the so far introduced criteria, we had to take into account the two different evaluation purposes a) and b) mentioned above in order to identify the appropriate performance indicators. In other words, an effort was required in order to identify a common set of indicators suitable for assessing both the quality and the effects of E&I (in the following we will use the term *performance* for indicating both purposes), regardless of the fact that we are comparing either raw and clean data or true and clean data. This task was not simple, particularly in the computation of some indices when using raw datasets, in which unacceptable or out of range data are generally present.
In the following sub-sections the evaluation indicators implemented in IDEA for each performance criterion are illustrated.

II.1 Preservation of individual data

13. The performance of an E&I process in terms of its overall impact on individual data can be measured in terms of how much and in what direction each variable has been modified by the procedure itself. Different indicators can be used depending on the variable nature (categorical or continuous).

14. Let *Y* be the variable subject to the E&I process, and let $Y^R{}_i$ and $Y^F{}_i$ be respectively the reference and final edited and imputed values of *Y* in the i^{th} unit (i=1,...,n).

If *Y* is a nominal variable, the index $D_1(Y^R, Y^F) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i^R, Y_i^F)$, where $I(Y_i^R, Y_i^F)$ =1 if $Y^R{}_i \neq Y^F{}_i$ and 0 otherwise is used.

If *Y* is an ordinal variable, we use the index $D_2(Y^R, Y^F) = \frac{1}{n \times m}\sum_{i=1}^{n} w(Y_i^R, Y_i^F)$, where

$$w(Y_i^R, Y_i^F) = \begin{cases} 0 & \text{if } \; Y_i^R = Y_i^F \\ \left|Y_i^R - Y_i^F\right| & \text{if } \; Y_i^R \neq Y_i^F \quad \text{and } Y_i^R, Y_i^F \neq blank \\ \text{m} \quad \text{if } Y_i^R \neq Y_i^F \text{ and } Y_i^R = blank \;\; \text{or } Y_i^F = blank & \end{cases}$$

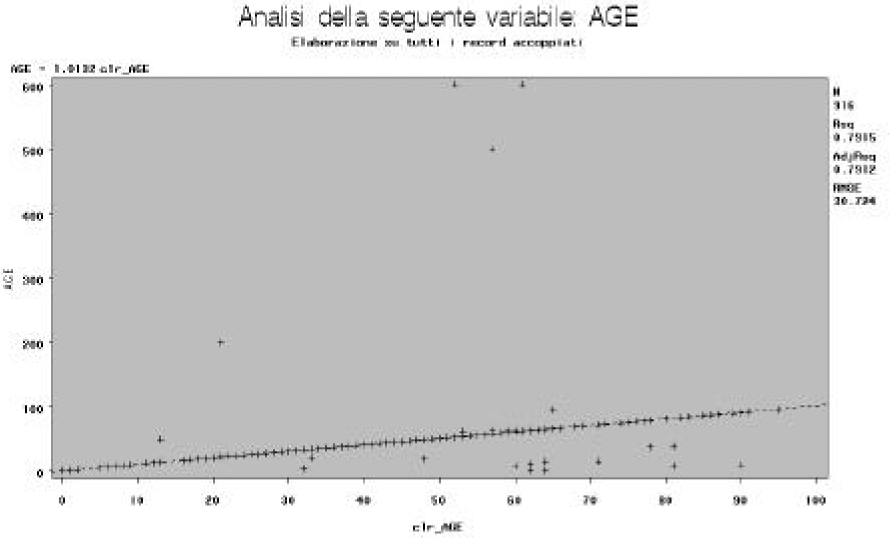
and $m=(max_Y - min_Y)+1$ if the category *blank* is in the domain of *Y*, while $m=(max_Y - min_Y)$ if the category *blank* is not in the domain of *Y*, and max_Y, min_Y are the maximum and minimum categories of *Y*.

15. For both nominal and ordinal variables, useful information on changes in categories due to the E&I phase is obtained by analysing the *transition matrix* obtained by building up a contingency table in which the categories of *Y* in the two compared data sets are crossed together. The frequencies of cells outside the main diagonal represent the number of changes due to E&I. Anomalous frequencies indicate possible biasing effects of E&I.

16. If *Y* is a continuous variable, the proposed indices take into account both the number and the amount of changes. They all belong to the class of measures $D_{La}(Y^R, Y^F) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| Y_i^R - Y_i^F \right|^a \right\}^{1/a}$ where a>0 is chosen in order to give the appropriate importance to high differences. The indices corresponding to α=1,2, $\propto$ have been implemented in IDEA.

17. For continuous variables, indices directly obtained by the regression model $Y_i^F = b \times Y_i^R$ are also used, such as the *slope* β, the R^2 and adjusted R^2, the *Root Mean Squared Error* (*RMSE*). In this case, graphical representations of data greatly help in identifying not effective performances. In figure 1 an example of IDEA scatter plot and indices obtained by regressing raw and final data for variable *Age* is shown.

Figure 1: Results of the regression between raw and final data for variable *Age*



II.2 Distributional accuracy

18. Generally, the evaluation of the E&I effects on (marginal or joint) distributions can be measured by means of *descriptive* statistics (indicators, techniques of multivariate analysis) or *test of hypothesis* techniques. *Descriptive* indicators miss the idea of generalisation of the conclusion (inference), however they furnish a simple first measurement of the distance of the distributions, and then they can give some clues to understand if the distribution of the compared datasets are quite similar. With regard to *test of hypothesis*, we believe that, in the context of Official Statistics, non-parametric statistical methods should be preferred: it is always difficult to introduce a model in survey investigations related to Official Statistics, and in addition non-parametric methods require few assumptions about the underlying population from which data are obtained. Furthermore, most of the classical distribution-free tests are based on the assumptions that the random variables to be tested are independent random samples, and this property is not always satisfied in complex survey design often adopted by NSOs. For these reasons, we selected only descriptive statistics.

19. In case of categorical variables, for univariate distributions the following dissimilarity indices are provided: $I_1 = \frac{1}{2}\sum_{k=1}^{K}\left| f_{Y_k{}^R} - f_{Y_k{}^F} \right|$, $I_2 = \left\{ \frac{1}{2}\sum_{k=1}^{K}\left| f_{Y_k{}^R} - f_{Y_k{}^F} \right|^2 \right\}^{\frac{1}{2}}$, where $f_{Yk}{}^R$ and $f_{Yk}{}^F$ are respectively the frequencies of category k in the reference and in the final datasets. It is obvious that information on changes of marginal distributions is also provided in the transition matrices so far introduced.

20. For bivariate distributions, the following index is used: $I_3 = \frac{1}{2}\sum_y\sum_x\left| f_{yx} - \tilde{f}_{yx} \right|$, where f_{yx}, $\tilde{f}_{yx}$ are respectively the frequencies of the contingency table obtained by crossing the categories of Y and X. The index assumes values in the interval [0,1] and can be easily extended to any set of k variables (k≥2) to make analyses on multiple distributions. Note that for variable combinations assuming few categories, I_3 can have a value lower than the value corresponding to variable combinations having many categories. Therefore, this index is useful when comparing results produced by different E&I methods on sets of variables with the same categories.

21. For a continuous variable *Y*, the Kolmogorov-Smirnov index (*KS*) is used to assess the difference between the marginal distributions of the variable in the compared data sets. Let $Y^R{}_n$, $Y^F{}_n$ be respectively the values of *Y* in the reference and final data sets containing n units. The *KS* distance is defined as:

$$\mathrm{KS}(F_{Y_n^R}, F_{Y_n^F}) = \max_t \left| F_{Y_n^R}(t) - F_{Y_n^F}(t) \right| \text{ where } F_{Y_n^R}(t) = \frac{\sum_{i=1}^{n} w_i I(Y_i^R \le t)}{\sum_{i=1}^{n} w_i}, F_{Y_n^F}(t) = \frac{\sum_{i=1}^{n} w_i I(Y_i^F \le t)}{\sum_{i=1}^{n} w_i}$$

and w_i are the sampling weights. Obviously, *KS*=0 only when $F_{Y_n^R}(t) = F_{Y_n^F}(t)\,\forall\, t$.

II.3 Preservation of aggregates

22. The impact of E&I on statistical aggregates is generally evaluated in terms of: 1) distance between the final estimate of the aggregate and the corresponding original one, and 2) non-sampling components of the variance of the aggregate estimate due to non sampling errors and E&I activities.
The first aspect does not take into account the different mechanisms or models underlying the generation of errors (missing and inconsistent values). It can be simply evaluated by observing the differences between point estimates before and after E&I. To this aim, totals, means, variances and other (either weighted or not) statistics can be computed and analysed through IDEA. Synthetic distances (Chambers, 2001) among these statistics will be implemented in the software.
It is well known that when estimating statistical parameters in presence of non-response and imputation, the variance of estimates gets inflated due to these non-sampling variability components. The most used approaches for correctly estimating variance in presence of this factors are re-sampling techniques (see, among others, Lee et. al., 2001; Rao, 2001; Beaumont et al., 2002), and multiple imputation (Rubin, 1987; Schafer, 1997). These aspects are not considered in the software IDEA, they represent a critical area for possible future developments.

II.4 Preservation of relations

23. One of the main drawbacks in using imputation relates to the fact that imputation has effect on multivariate relations between variables. Discussions about effects of imputation on data relationships can be found in Kalton et al. (1982), Kalton et al. (1986) and Little (1986). Recent research relates particularly to regression imputation (Shao et al., 2002), but more research is needed with respect to other imputation techniques. One traditional way of evaluating the effects of E&I on data relations consists in analysing changes produced on the covariance or correlation or association structure of reference data, depending on the variables nature.

24. For categorical variables, the preservation of the bivariate relation between items *Y* and X subject to E&I (either both or only one of them), is measured by analysing the *Cramer contingency coefficient* (Kendall et al., 1979) before and after the E&I process:

$$C=\left\{\frac{c^{2}}{n\times min(r-1,c-1)}\right\}^{1/2}$$

were the χ2 is the traditional squared contingency index based on the differences among the frequencies of the two-way contingency table crossing the r categories of Y and the c categories of X, and the corresponding theoretical frequencies (i.e. frequencies corresponding to no association), and n is the number of observations. C is 0 when Y and X are not associated, while in case of complete association its value is 1. Note that the I_3 index previously introduced (section II.2) provides summary information about overall E&I effects on multivariate relationships.

25. For continuous variables, IDEA allows to measure only the preservation of bivariate relations. This is done by analysing changes in usual measures (covariances and correlations) before and after the E&I process. For each couple of items subject to E&I, IDEA provides the covariance and the Pearson correlation indices.

III. DOCUMENTING EDITING AND IMPUTATION PROCESSES: THE SIDI SYSTEM

III.1 Quality indicators

26. Since 2001, the Information System for Survey Documentation SIDI is implemented in ISTAT. In SIDI, survey managers document survey processes in terms of metadata and quality indicators in a standard and integrated way.

27. Documentation in SIDI is process oriented, in the sense that it follows the different phases of the data production process. Considering the editing and imputation phase, survey managers are asked to provide information on the editing technique (e.g. manual, interactive, automatic) and on the methodology for detecting and correcting the errors (e.g. deterministic or stochastic methods), as well as one or more sets of quality indicators on the impact of editing and imputation procedures on data (Fortini et al., 2000, Fortini et al, 1999).

28. The SIDI set of standard quality indicators on E&I are described in table 1. As it can be seen by observing the indicators' formulae, all the required numbers can be obtained comparing the raw and the final data matrices, following the process oriented approach.

29. While the main aim of collecting metadata in a centralized system is documentation, the presence of quality indicators could be useful also for evaluation purposes. In fact, the first group of indicators (1 to 15) concerns the overall impact of E&I procedures on the data matrix. In particular, the first three ones provide the data matrix dimension, while the remaining ones measure specific aspects of E&I performance. For example, the imputation rate is the percentage of survey data modified in some way by the E&I procedure. It can be interpreted as "how much E&I have modified the original data". The percent composition of imputation rate suggests what are the major problems in the overall quality of collected (raw) data. High percentages of net imputation indicate an item non-response problem. Otherwise, high percentages of modification indicate that erroneous values are the main problem and it can be then investigated if these errors are originated by data collection or data entry. Indicators 16 to 23 concern the distributions of imputation rate by variables and by records. The indicators in this subset are voluntarily not very sensitive. They are meaningful only in the case that the impact of E&I on data is quite strong and they work as an alarm bell that indicates problems on the original data or on the E&I procedure. In fact, through the analysis of these indicators, it is possible to discover, for example, if the E&I procedure tends to work heavily on a few variables. In general terms this sub-set of indicators can be useful to understand the behaviour of the E&I procedure: if the values of these indicators are generally low, it means that the E&I procedure doesn't modify so much the data and there aren't groups of variables or records more affected than others.

30. As already mentioned, a survey manager can provide one or more sets of quality indicators. In fact, in a survey, different techniques and methods for E&I can be used and there can be more than one data matrix, for example related to different statistical units. Typically, household surveys have a matrix in which each record is an household and it contains only the information collected on the household itself, and another data matrix in which each record refers to an individual (household component). Furthermore, most ISTAT surveys are sampling ones, and obviously in order to evaluate the real impact of E&I on estimates it's necessary to calculate weighted standard indicators which take into account the sample weight of each statistical unit. In detail, a survey manager, for each survey occasion, has to provide a "principal" set of quality indicators, un-weighted, related to the most important survey data matrix (e.g. statistical unit), and to the most relevant E&I methodology (if it is possible to split the different E&I steps). Then, he/she can provide a weighted set, describing the weighting scheme and indicating the total number of un-weighted units. Finally, he/she can provide other sets of quality indicators related to other statistical units and/or E&I techniques.

Table 1. Indicators on the quality of editing and imputation phase and their formulae

N	INDICATORS	FORMULAE OR DEFINITIONS
1	Total Records	
2	Total Variables	
3	Total Imputable Variables	Number of potentially imputable variables by the editing procedures [3].
4	Imputation Rate	Values modified by E&I / potentially imputable values [4]
5	Modification Rate	Changes from a value to a different imputed value / potentially imputable values
6	Net Imputation Rate	Changes from blank to a different imputed value / potentially imputable values
7	Cancellation Rate	Changes from a value to an imputed blank / potentially imputable values
8	Non Imputation Rate	Values not transformed by E&I / potentially imputable values
9	Blank Unmodified Values Rate	Blank unmodified values / potentially imputable values
10	Non Blank Unmodified Values Rate	Non blank unmodified values / potentially imputable values
Percent Components of Imputation Rate		
11	% of Modification	Changes from a value to a different imputed value / imputed values
12	% of Net Imputation	Changes from blank to a different imputed value / imputed values
13	% of Cancellation	Changes from a value to an imputed blank / imputed values
Percent Components of Non Imputation Rate		
14	% of Blank Unmodified Values	Blank unmodified values / non imputed values
15	% of Non Blank Unmodified Values	Non blank unmodified values / non imputed values
Indicators referred to Imputation Rate Distribution		
16	First Quartile of Imputation Rate Distribution by VARIABLE	Value of the imputation rate leaving the 25% of the ordered <u>variables</u> to the left
17	Third Quartile of Imputation Rate Distribution by VARIABLE	Value of the imputation rate leaving the 75% of the ordered <u>variables</u> to the left
18	Number of Variables with an imputation Rate greater than 5%	
19	Number of Variables with an Imputat ion Rate greater than 2%	
20	First Quartile of Imputation Rate Distribution by RECORD	Value of the imputation rate leaving the 25% of the ordered <u>units</u> to the left
21	Third Quartile of Imputation Rate Distribution by RECORD	Value of the imputation rate leaving the 75% of the ordered <u>units</u> to the left
22	Number of Records with an Imputation Rate greater than 5%	
23	Number of Records with an Imputation Rate greater than 2%	

31. For the principal set of indicators, the system, besides the indicators' computation and tabular representation, provides various specifically designed functionalities for further analysis. Firstly, the indicators on the overall impact of E&I procedures on the data matrix can be analysed with regard to geographical detail. The system offers graphical representations (maps) of the indicators that allow subject matter experts to identify troubles in particular geographical areas. Through this functionality it is possible to perform an high level territorial monitoring of the quality of collected data. Secondly, through time series graphical representations, the system allows survey managers to monitor over time the values of various

[3] Some variables might be excluded from the imputation process (e.g. identification codes)

[4] Potentially imputable values= Total records* Total imputable variables

indicators. Using this functionality, a subject matter expert is able to evaluate the performance of the E&I procedure and its impact on data through consecutive survey occasions. Finally, it is possible to make different types of comparisons to better evaluate the quality of survey data:

a. For a given survey, the system permits to compare the value of an indicator with a general mean value, obtained averaging the values of the same indicator for all surveys.
b. Furthermore, the comparison could be done with particular averages calculated within subgroup of surveys that use the same E&I methodology.
c. At last, it is possible to compare the values of the same indicator in different surveys.

32. At the moment, the mentioned functionalities are available only for the main set of indicators. They will be soon implemented even for the weighted set, for which now only the indicators' computation is available. For the further sets of indicators the system builds reports containing all the indicators values and the metadata necessary to correctly interpret the indicators.

III.2 Implementing the SIDI system

33. Implementing and maintaining information systems is a demanding task. Even if documentation is recognised as an important aspect of quality, it is time-consuming and survey managers do not usually consider it as part of their current production activity (Blanc M. et al., 2001). Therefore, it is important to have a strategy for populating and keeping information systems updated.

34. In particular, the SIDI system has a considerable impact on current statistical activity. In fact, survey managers are required not only to provide metadata on the information content and the production process, but also to calculate a set of standard quality indicators for the main phases (frame, data collection, data entry, editing and imputation, timeliness and costs) of each survey occasion. Both aspects (documentation of survey metadata and calculation of standard indicators) require specific training and knowledge of the system definitions and functionalities. However, the calculation of quality indicators has the greatest impact from a technical and organisational point of view. With regard to survey metadata, the bigger effort is needed the first time the survey manager has to document her/his survey. Once the survey documentation has been completed, the metadata only need to be updated when a change occurs. Differently, the standard quality indicators have to be calculated for each survey occasion. Thus implying an additional amount of work for the survey managers. Furthermore, for certain surveys it might be necessary to review some procedures in order to be able to calculate the SIDI indicators.

35. The awareness of such problems has brought us to define a strategy for the implementation of the SIDI system. Three main aspects of such a strategy are to be mentioned:

a. The net of quality facilitators inside Istat. It is a new role for Istat which has been designed for supporting the release of SIDI. The quality facilitators are experts of quality issues and of the SIDI system whose task is to document and update the survey metadata and to calculate the standard quality indicators. After attending an especially designed training course, the quality facilitators are formally appointed. It is foreseen to train a quality facilitator for each Istat survey, thus creating a net inside the Institute. Up to now, some 50 people have already been trained. The training of the quality facilitators is going to be completed by 2003.
b. The development of generalised software to support the survey managers in the calculation of standard quality indicators. As already mentioned, the quality indicators managed into SIDI are process oriented. This means that they could be obtained as a by-product of the survey production process itself. Furthermore, different production processes are homogeneous with respect to the evaluation of a given set of standard quality indicators. An example are the “data entry” quality indicators for all those surveys which use external companies. For these surveys, Istat applies the same quality control procedure for assessing the quality of data entry. Therefore, it is planned to develop a generalised procedure for the calculation of SIDI indicators related to this phase. The availability of generalised software for calculating quality indicators is one of the major support that could be provided to survey managers in order to simplify and speed up their work. In fact, a main purpose is to integrate as much as possible the quality activity into statistical production processes. The software IDEA has been developed also for these purposes and it is now currently used to provide quality indicators for the SIDI system. The relationships between IDEA and SIDI are better described below.

c. The integration between SIDI and other local information systems or data bases where relevant information for the calculation of quality indicators is stored. Examples are the information system for monitoring the data collection phase for business structural statistics and the data base with data collection information related to multipurposes surveys.

36. SIDI is actually made up of two different subsystem: SIDI1 is the management system for inserting, modifying and updating metadata and quality indicators; and SIDITOP is the display system for querying and navigating through metadata and quality indicators. At the moment SIDITOP is available on ISTAT intranet. To assure that the indicators are calculated in the same standard way by each survey, it has been decided to ask survey managers to provide in SIDI1 the numerators and the denominators (*numbers*) needed for calculating the indicators. The system itself calculates the quality indicators. By means of SIDITOP, the quality indicators can be analysed (time series and/or geographical analyses) and compared (among different surveys and with general and specific mean values). With regard to E&I indicators the numerators and the denominators of the formulae in table 1 are to be inserted in SIDI1. To this purpose, survey managers need to compare raw and clean data sets for the same survey occasion. This would have required the survey managers to prepare some ad hoc programmes in order to calculate the *numbers* for SIDI1, implying possible errors in computations and loss in timeliness of the system update. As mentioned before, such a work was needed for every survey using an E&I procedure, regardless to the procedure used (i.e. deterministic or stochastic imputation). Therefore, the best way to support survey managers in this specific task was to develop a generalised software which could automatically produce the input numbers for SIDI1 by comparing raw and clean data sets. The software IDEA easily provides such numbers in a standard way for all Istat surveys. In conclusion, it is worth mentioning that IDEA has an added value because the documentation activity required by the SIDI system for the E&I phase can be done with the same generalised software that can be used to tune and monitor the E&I process.

IV. THE SOFTWARE IDEA: OPERATIONAL ASPECTS

37. The generalised software IDEA has been implemented in SAS/AF, and allows the computation of standard indicators under different evaluation contexts through the comparison of appropriate couples of data sets. When measuring the *effects* of E&I, IDEA allows two types of evaluations based on the comparison of *raw* and *clean* data:

a. evaluating effects at “high” level, i.e. by considering all variables and units subject to E&I;
b. evaluating effects at “low” level, i.e. by considering single items (or subsets of items) and/or subgroups of units.

Different indices are used in the two approaches. In the first case, IDEA allows the computation of the SIDI standard quality measures. In the second approach, IDEA provides different types of indices depending on the particular investigated aspect (preservation of elementary data, distributions, aggregates, relations).

38. When evaluating the *quality* of E&I, the capability of the procedure of correctly deal with errors is evaluated through indicators computed by comparing *true* and *clean* data. The evaluation level is the ”low” one, i.e. single items (or subsets of items) and/or subgroups of units are considered. The statistical measures used in this case are the same as those used for evaluating the effects of E&I at “low” level.

39. As so far mentioned, IDEA allows survey managers computing indicators on particular subsets of units: i) at both “high” and “low” level, separate analyses can be performed on different *data domains* identified by a *stratification* item; ii) at “low” level, for each item subject to E&I, E&I effects or quality can be evaluated by considering only the subset of data modified during E&I. This last aspect is particularly useful when we want to perform more detailed analyses, or when the percentage of modified values is low with respect to the overall observed values, thus analysing all observations should mask the real E&I impact on data. In case of sample surveys, at both “high” and “low” level weighted performance indicators can be computed.

V. CONCLUDING REMARKS AND FUTURE WORK

40. Information on E&I quality and impact represents not only a requirement at NSOs level, but also a powerful tool that survey subject matters can use for better understanding data and process characteristics. Improvements in the short and medium run can be produced on the basis of indications provided by the analysis of the performance of E&I activities on specific items, units, errors. IDEA is a generalized software allowing different types of evaluations aiming at satisfying different needs of subject matter experts. At the present stage of development, IDEA provides a set of statistical measures partially inspired to the evaluation measures used in the EUREDIT project, and allows the computation of the standard quality indicators required by the ISTAT SIDI system. IDEA is currently used by ISTAT survey managers because of its usefulness in terms of standardization and simplification of the available indicators calculation.

41. Further software developments are planned, particularly relating to the evaluation of E&I effects on data relationships and multivariate distributions. Indices for evaluating the quality of editing like those proposed in Manzari et al. (1999) will be also implemented in the software. Appropriate synthetic measures are also needed to better evaluate the E&I impact on estimates like totals, means, variances. Finally, additional data and data distributions graphical representations could improve the evaluation effectiveness.

REFERENCES

Beaumont J.-F., Mitchell C. (2002) The System for Estimation of Variance due to Nonresponse and Imputation (SEVANI), *Proceedings of the Statistics Canada Symposium 2002, Modeling Survey Data for Social and Economic Research* (to appear).

Blanc M., Lundholm G., Signore M. (2001), “LEG chapter: Documentation”, Proceedings of the International Conference on Quality in Official Statistics, Stockholm 14-15 May 2001, CD-ROM.

Brancato G., D'Angiolini G., Signore M. (1998) Building up The Quality Profile of ISTAT Surveys, Proceedings of the Joint IASS-INEGI-IAOS Conference " Statistics for Economic and Social Development, Aguascaliente, Messico, 1-4 September, CD ROM

Cirianni A., Di Zio M., Luzi O., Palmieri A., Seeber A.C. (2001) Comparing the effect of different adjustment methods for units with large amounts of item non-response: a case study, Proceedings of the International Conference on Quality in Official Statistics (First version), Stockholm, Sweden, May 14-15.

Chambers R. (2001) Evaluation Criteria for Statistical Editing and Imputation, National Statistics Methodology Series no 28, Office for National Statistics.

Charlton J. (2003) First results from the EUREDIT project – Evaluating Methods for Data Editing and Imputation, Proceedings of the 54th ISI Session, Berlin, 13-20 August (to appear).

Di Zio M, Manzari A., Luzi O. (2001) Evaluating Editing and Imputation Processes: the Italian Experience, UN/ECE Work Session on Statistical Data Editing, Helsinky, Finland, May 27-29.

Eurostat (2000), *Standard Quality Report*, Eurostat Working Group on Assessment of Quality in Statistics, Eurostat/A4/Quality/00/General/Standard Report, Luxembourg, April 4-5.

Fortini M., Scanu M., Signore M. (2000) Use of indicators from data editing for monitoring the quality of the survey process: the Italian information system for survey documentation (SIDI), Statistical Journal of the United Nations ECE, n.17, pp. 25-35.

Fortini M., Scanu M. and Signore M. (1999) Measuring and Analysing the Data Editing Activity in ISTAT Information System for Survey Documentation, Statistical Data Editing: A selection of papers presented at the UN/ECE Work Session on Statistical Data Editing, 2-4 June, Roma, Essays, ISTAT, pp. 17-29

Granquist, L. (1995) Improving the Traditional Editing Process, in Business Survey Methods, eds. B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, P.S. Kott, New York,: Wiley, pp. 385-401.

Granquist, L. (1997a) An overview of Methods of Evaluating Data Editing Procedures, Statistical Data Editing Methods and Techniques Vol. II, Conference of European Statisticians, United Nations, 1997.

Granquist L. (1997b) The New View on Editing. International Statistical Review. 65. No.3. pp. 381-387.

Kalton, G., Kasprzyk, D. (1986) The treatment of missing survey data. Survey Methodology, 12, No 1, 1-16.

Kalton, G., Kasprzyk, D. (1982) Imputing for missing survey responses. Proceedings of the section on Survey Research Methods, American Statistical Association, pp. 22-31.

Kendall M., Stuart A. (1979) The Advanced Theory of Statistics, Vol II: Inference and Relationship. Griffin, London.

Lee H., Rancourt E., Särndal C.-E. (2001) Variance Estimation from Survey Data under Single Imputation, in Groves R.M., Dillman D.A., Eltinge J.L., Little R.J.A. (eds), Survey Nonresponse, New-York:John Wiley&Sons, Inc., pp. 315-328.

Little, R.J.A. (1986) Survey nonresponse adjustments for estimates of means, International Statistical Review, 54, pp. 139-157.

Lyberg L.. et al. (2001), Summary Report from the Leadership Group (LEG) on Quality", Proceedings of the International Conference on Quality in Official Statistics, Stockholm 14-15 May 2001, CD-ROM.

Manzari A., Della Rocca G. (1999) E.S.S.E. Editing Systems Standard Evaluation, Conference of European Statisticians, Work Session on Statistical Data Editing, Rome, June 2-4.

Madsen B., Solheim L. (2000) How to measure the effect of data editing, Conference of European Statisticians, Work Session on Statistical Data Editing, 2000.

National Center for Education Statistics (1992) "NCES Statistical Standards"

Norbotten S. (1997) Metrics for the Quality of Editing, Imputation and Prediction, Statistics Sweden Technical Report, SN, 3 Oct 1997.

Norbotten S. (2000) Evaluating Efficiency of Statistical Data Editing: A General Framework, United Nations, 2000.

Rao, J.N.K. (2001). Variance Estimation in the Presence of Imputation for Missing Data. Proceedings of the Second International Conference on Establishment Surveys (ICESII), pp. 599-608.

Rubin, D. (1987). Multiple Imputation in Surveys. John Wiley & Sons.

Schafer, J.L. (1997) Analysis of Incomplete Multivariate Data, Chapman & Hall

Schulte Nordholt, E. (1998) Imputation: Methods, Simulation, Experiments and Practical Examples. International Statistical Review, 66, 157-180.

Shao, J. and Wang, H. (2002) Sample Correlation Coefficients Based on Survey Data under Regression Imputation. Journal of the America Statistical Association, Vol. 97, pp. 544-552

Statistics Canada (1998) Statistics Canada Quality Guidelines.

UNITED NATIONS (2000) Glossary of terms on Statistical data Editing, Geneva, 2000.

Verboon P., Schulte Nordholt E. (1997) Simulation Experiments for Hot-deck Imputation, in Statistical Data Editing, Methods and Techniques, Vol. II, N.48, UN/ECE, pp. 22-29.

Whitridge P., Benier J. (1999) The impact of Editing on Data Quality, UN/ECE Work Session on Statistical Data Editing, Rome, June 2-4.