CRP.3
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (v): Quality indicators and quality reporting

QUALITY INDICATORS FOR EVALUATING AND DOCUMENTING EDITING AND IMPUTATION

Submitted by ISTAT, Italy[1]

I. Introduction

In National Statistical Offices (NSOs) the analysis of the effects on data of any processing activity has progressively assumed a central role. In this paper we concentrate on the problem of evaluating and documenting data editing and imputation processes (E&I), and in particular the E&I activities performed at the post-data collection stage.

The results provided by evaluation studies on E&I procedures are widely recognized as an important source of information on the quality of the data resulting from data capturing, on the quality of the E&I process itself and on the reliability of other aspects of the survey process. The definition of the *quality of an E&I process* depends on the particular E&I aspect that we want to investigate and evaluate (Granquist, 1997). From a general point of view, different evaluation studies are performed at different stages of the E&I *life-cycle* in order to gather different types of information on E&I (Di Zio *et al.*, 2001): information on the accuracy of an E&I method/procedure (in terms of its capability of correctly identifying and restoring "true" data) is generally collected *before* the application of the E&I method/procedure to survey data (e.g. through experimental studies); information on the effects of an E&I procedure on data for monitoring its performance and tuning it is generally obtained *during* the data processing; information needed for documentation and survey management purposes (e.g. for monitoring over time the changes of the E&I impact on data) is obtained *after* the data treatment. Furthermore, other elements are to be taken into account when assessing the quality of an E&I process, like its cost, timeliness, burden on respondents.

In the area of evaluating E&I, the most recent activities at Istat have been mainly focused on the following areas (Di Zio *et al.*, 2001; Brancato *et al.*, 2004): 1) evaluating the performance of E&I methods through the analysis of their statistical effects on data distributions and relations; 2) providing the subject-matter experts with detailed documentation on the process impact on data in each survey occasion in order to permit the monitoring of process quality over time; 3) providing NSO with information about the quality level of each survey in order to allow comparisons; 4) providing final users with information about the main characteristics of the E&I process and the data modifications due to the E&I process.

In order to meet the above mentioned objectives, an evaluation framework in which both users and producers needs are taken into account has been developed at Istat. In this context, crucial problems like identifying appropriate indicators, standardizing their computation, supporting the survey managers in computing them have been faced. The activities have been carried on taking into account the increasing attention paid to the identification of standard quality indicators and the production of standard quality reports at both NSOs level and Eurostat level (Chambers, 2001; Linden *et al.*, 2004; Eurostat, 2000).

[1] Prepared by Giorgio Della Rocca, Orietta Luzi, Marina Signore, Giorgia Simeoni.

Supporting survey managers is an important aspect because of the costs and burden due to the additional charge of producing standard quality reports for assessing and documenting the data processing activities.

In this paper we describe the recent advancements reached at Istat in the area of developing a comprehensive framework for evaluating E&I procedures and documenting their impact on survey data. This framework consists of a system of statistical measures corresponding to:

- performance indicators for evaluation purposes;
- standard quality indicators for documentation purposes.

It is worth mentioning that the latter are managed in the *Information System for Survey Documentation (SIDI)*, which is the centralized system for standard documentation of all Istat surveys. SIDI is based on the integrated management of metadata and standard quality indicators. It covers all the different stages of the data production process and includes a specific subset of standard indicators related to the E&I process.

All the proposed indicators, including those defined in the SIDI context, have been implemented in a generalized tool called IDEA (*Indices for Data Editing Assessment*). IDEA has then multiple goals, namely to allow survey managers to compute all the proposed indicators in a standardized and controlled way as well as to simplify and make less expensive the evaluation task thus stimulating survey managers in performing the evaluation of data processing.

A common feature of these indicators is that they can all be derived by comparing edited data and the corresponding raw data.

In addition, they can be classified as *high level indicators* and *low level indicators*. High level indicators (e.g. the standard quality indicators required by SIDI) are computed taking into account all the edited variables and units. The low level indicators are computed on subgroups of edited variables and/or subsets of units, in particular, the subset of data modified during the E&I task. Considering the subset of modified data is particularly useful when more detailed analyses are to be performed, or when the percentage of the modified values is low with respect to the overall observed values.

All the indicators can be computed on different *data domains* identified by a user-defined *stratification* item. For sample surveys, weighted indicators can also be obtained.

The paper is structured as follows: in section 2 the indicators defined for evaluating the quality or the effects of E&I processes are described. Section 3 focuses on the standard indicators required for documenting E&I processes in the SIDI system. Issues related to the relationships between SIDI and IDEA are also illustrated. Concluding remarks are summarized in section 4.

2. evaluating the IMPact of editing and imputation procedures

In this section we describe the approach and the indicators proposed for measuring the impact on a given set of survey data of an E&I method/procedure.

An E&I process generally consists of many sub-phases, each performing a particular step of the whole data treatment process (e.g. treatment of outliers, editing of systematic or stochastic errors on either categorical or continuous data, and so on.). As a consequence, the evaluation task can consist in the assessment of the impact on data of the overall E&I process, or it can be split in the evaluation of simpler sub-problems, each of them focused on a specific E&I sub-phase.

Furthermore, the evaluation of the impact on statistical survey data of E&I activities can be performed with different goals. One of them is monitoring and tuning the E&I process: the modifications produced on raw data are analysed in order to identify possible problems in data and data processing, and improve the efficiency of the E&I process during the data treatment itself. This type of analysis is generally performed with respect to sub-phases of the overall E&I process, in order to optimise the tuning process.

The analysis of the impact of E&I on raw data is also performed for documentation purposes, for producing information for final users, and for data analysis. For example, for a given survey, the availability of information on the statistical effects of the E&I activities over time permits a comparative analysis that could highlight structural modifications in the surveyed phenomena or occasional organisational problems requiring permanent or occasional adjustments to the survey organisation or to the E&I procedure.

Under specific conditions, the approach and the indicators illustrated in this section can also be used to measure the quality of an E&I method/procedure in terms of its capability of correctly deal with errors, e.g. the capability of an error localization process of correctly identifying not acceptable data, or the capability of an imputation technique of restoring *true* data in sampling units. This kind of evaluation implies the knowledge of *true* data for each unit[2]: in this case, the proposed indicators are to be computed by comparing the final data to the corresponding *true* ones. This type of evaluation study is typically performed at the E&I design and test phases, when the aim is assessing the suitability of a given method or approach for a specific data problem.

The performance indicators illustrated in this section have been defined starting from the results obtained during the EUREDIT project[3] (Charlton, 2003; Chambers, 2001), in which a general framework for evaluating the quality of E&I methods in an experimental context was developed. On the basis of the quality (or performance) criteria defined in EUREDIT, a set of statistical measures and graphical techniques are proposed. In particular, the impact of an E&I method/procedure is evaluated on the basis of the following performance criteria:

- impact on individual data;
- impact on marginal and joint distributions;
- impact on aggregates;
- impact on relationships between variables.

Depending on the evaluation purpose, each criterion assumes a different meaning. For example, if the aim is assessing the effects on data of an E&I method, indicators measuring the impact on individual data values provide information on the amount of changes produced on elementary data by the E&I activities, while if our goal is evaluating the quality of an E&I method, these indicators provide information on the method capability of recovering the *true* values for missing or erroneous items.

In general, the relevance and priority of the above mentioned criteria mainly depend on the investigation objectives, the investigation characteristics and the nature of the analyzed variables. For example, if our aim is to verify the quality of an imputation method and micro data have to be provided to final users, the most important quality criteria that has to be met is the preservation of both micro data and (marginal or joint) distributions. As a further example, the distributional accuracy can assume a relevant role in the case that the distributional assumptions (univariate or multivariate) on observed variables are to be analyzed or taken into account in subsequent statistical analyses.

The evaluation indicators proposed for each one of the so far introduced criteria are low level indicators. In defining these measures, an effort was required in order to identify a common set of suitable indicators 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 to the fact that we are comparing either raw and clean data, or true and clean data. This task was not simple, particularly when the reference data correspond to raw datasets: in this case, the unacceptable or out of range data might affect the computation of indicators.

The proposed evaluation indicators can be applied for evaluating either the overall E&I process, or each single E&I step, depending on the evaluation purpose and the available indicators.

In the following subsections the indicators implemented in IDEA for each of the performance criteria mentioned above will be illustrated.

2. 1 Preservation of individual data

The indicators proposed for measuring the impact of an E&I process on individual survey data depend on the nature of the investigated variable (categorical or continuous).
Let Y be the variable subject to E&I, and let Y^R_i and Y^F_i be respectively the reference and final values of Y in the i^{th} unit (i=1,...,n), where n is the number of responding units.

a) Categorical Variables (nominal or ordinal)

[2] A low cost consuming approach that permits this situation is based on the use of the simulation approach, also adopted in the EUREDIT project for the comparative evaluation of competitive E&I methods.
[3] The EUREDIT Project was funded under EU Fifth Framework research program (www.cs.york.ac.uk/euredit/).

a.1) If *Y* is a categorical variable, straightforward information on the overall amount of individual data in which the value of *Y* changed from one category to another category due to E&I is provided by the following *imputation rate*:

$$D_1(Y^R, Y^F) = \frac{\sum_{i=1}^{n} w_i \times I(Y_i^R, Y_i^F)}{\sum_{i=1}^{n} w_i} \times 100$$

where w_i are the possible sampling weights and $I(Y_i^R, Y_i^F)=1$ if $Y^R_i \neq Y^F_i$ and 0 otherwise. $D_1(Y^R, Y^F)$ simply measures the percentage of units in which *Y* has different categories in the reference and final data sets. $D_1(Y^R, Y^F)$ is 0 if all the categories are equal in the two compared data sets, while it reaches its maximum value 100 when all units assume a different category in the two data sets.

a.2) The *net imputation rate* $D_{1i}(Y^R, Y^F)$ indicates the percentage of units in which the value of *Y* changed from a *blank* value in the reference data set to a non *blank* value in the final data set. It is computed like $D_1(Y^R, Y^F)$ where in this case $I(Y_i^R, Y_i^F)=1$ if [$Y_i^R \neq Y_i^F$ and Y_i^R = *blank* and $Y_i^F \neq$ *blank*], and 0 otherwise.

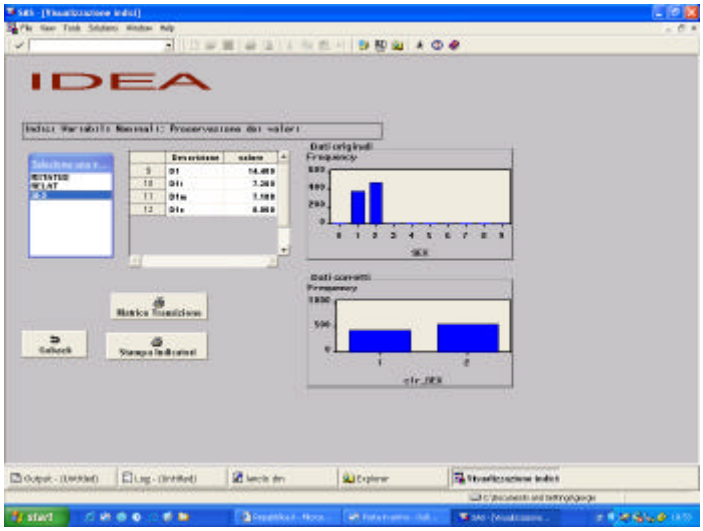
a.3) The *modification rate* $D_{1m}(Y^R, Y^F)$ indicates the percentage of units in which the value of *Y* changed from a non *blank* value in the reference data set to another non *blank* value in the final data set. It is computed like $D_1(Y^R, Y^F)$ where in this case $I(Y_i^R, Y_i^F)=1$ if [$Y_i^R \neq Y_i^F$ and $Y_i^R \neq$*blank* and $Y_i^F \neq$*blank*], and 0 otherwise.

a.4) The *cancellation rate* $D_{1c}(Y^R, Y^F)$ indicates the percentage of units in which the value of *Y* changed from a non *blank* value in the reference data set to a *blank* value in the final data set. It is computed like $D_1(Y^R, Y^F)$ where in this case $I(Y_i^R, Y_i^F)=1$ if [$Y_i^R \neq Y_i^F$ and $Y_i^R \neq$*blank* and $Y_i^F =$*blank*], and 0 otherwise[4].

a.5) Graphical representations of the frequency distributions of *Y* in the two compared data sets permit the visual analysis of the overall data changes. In figure 1 an example of graphical representations that can be directly performed in IDEA is shown for variable *Sex*.

Figure 1: Frequency distribution in reference and final data for variable *Sex*

[4] it is obvious that, when the *blank* value is not in the domain of *Y*, $D_{1c}(Y^R, Y^F)$ highlights a problem in the E&I process.



a.6) Useful information on changes in categories due to E&I 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.

b) <u>*Ordinal Variables only*</u>

b.1) If *Y* is an ordinal variable, information about how much categories of *Y* have been changed due to E&I is given by the following index:

$$D_2(Y^R, Y^F) = \frac{1}{m \times \sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \times d(Y_i^R, Y_i^F) \times 100$$

where Y_i^R and Y_i^F are the coded categories of *Y* in the reference and in the final data sets, w_i are the possible sampling weights,

$$m = \begin{cases} (max_Y - min_Y) + 1 & \text{if the category } blank \text{ is in the domain of } Y \\ (max_Y - min_Y) & \text{if the category } blank \text{ is not in the domain of } Y \end{cases}$$

where max_Y, and min_Y are the higher and lower categories of *Y*, and

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

$D_2(Y^R, Y^F)$ varies between 0 (all categories are equal in the two compared data sets) and 100 (maximum difference among categories).

b.2) If the value *blank* is not in the domain of variable *Y*, the index $D_2(Y^R, Y^F)$ is also computed on the subset of the $n_g \leq n$ of units in which *Y* assumes non *blank* values. In other words, once the item non-responses are discarded from the computation, the new index $D_{2n}(Y^R, Y^F)$ measures the *net* percentage of changes on *Y* values from not *blank* categories to other non *blank* categories. $D_{2n}(Y^R, Y^F)$ is useful particularly when *Y* is affected by a large amount of item non responses, in this situations in fact $d(Y_i^R, Y_i^F)$ assumes it s maximum value *m* and $D_2(Y^R, Y^F)$ results amplified by large amounts of these values.

c) <u>Continuous Variables</u>

c.1) If *Y* is a continuous variable, indicators $D_1(Y^R, Y^F)$, $D_{1i,}(Y^R, Y^F)$ $D_{1m}(Y^R, Y^F)$ and $D_{1c}(Y^R, Y^F)$ are still applicable, with similar meaning, and they provide information about how many data of *Y* have been changed due to E&I.

c.2) In order to evaluate both the number and the amount of changes on variable *Y*, indices belonging to the class of measures

$$D_{La}(Y^R, Y^F) = \left\{ \frac{\sum_{i=1}^{n} w_i \times \left| Y_i^R - Y_i^F \right|^a}{\sum_{i=1}^{n} w_i} \right\}^{1/a}$$

are proposed, where w_i are the possible sampling weights, and a>0 is chosen in order to give the appropriate importance to high differences between values of *Y* in the reference and in the final data sets. The indices $D_{L1}(Y^R, Y^F)$, $D_{L2}(Y^R, Y^F)$ and $D_{L\infty}(Y^R, Y^F)$ have been implemented in IDEA, where:

$$D_{L\infty}(Y^R, Y^F) = \frac{\max_i \left| Y_i^R - Y_i^F \right|}{\sum_{i=1}^{n} w_i}$$

It is straightforward that the higher is the distance between corresponding values in the two compared data sets, the higher is the value of each one of these indices.

c.3) A set of simple statistics directly provide information on some aspects of changes in individual data, like the *number of missing and non missing observations* in the compared data sets.

c.4) A set of indices can be directly obtained by the following regression model:

$$Y_i^F = b \times Y_i^R$$

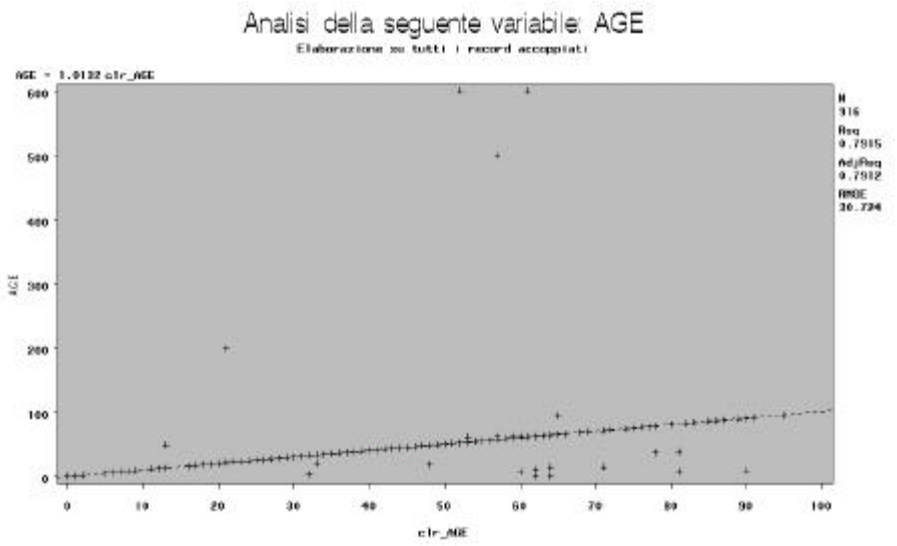
In particular we use the following ones:

- the *slope* β,
- the R^2 and the adjusted R^2,
- the *Root Mean Squared Error* (*RMSE*).

The graphical representation of regression results greatly helps in identifying not effective performances: in Figure 2 an example of scatter plot produced by the IDEA software for the regression analysis between raw and final data for variable *Age* is shown.

c.5) Graphical representations of the *Y* values in the two compared data sets permit a deepest analysis of changes in the individual data due to E&I. Box plots and histograms are useful tools in this context. In figure 3 an example of graphical representations of data through SAS Insight that can be directly performed in IDEA is shown for variable *Age*. Useful information on the main distributional statistics is directly provided by these representations.

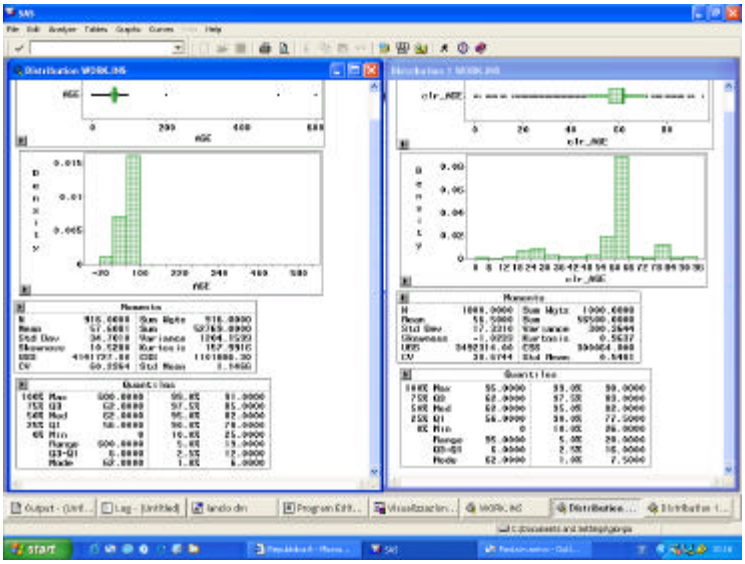
Figure 2: Results of the regression between raw and final data for variable *Age*



2.2 Preservation of marginal and joint distributions

The evaluation of the E&I impact on (marginal or joint) distributions can be measured by means of *descriptive* statistics (indicators, techniques of multivariate analysis) and/or *test of hypothesis* techniques. *Descriptive* indicators miss the idea of generalisation of the conclusion (inference), however they provide a simple first measurement of the distance of the distributions, and then they can give some clues to understand to what extent the compared distributions are similar.

Figure 3: Box plot and histogram in reference and final data for variable *Age*



Relating to the use of *test of hypothesis* in the context of Official Statistics, it has to be underlined that it is always difficult to introduce a model in survey investigations, 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 designs often adopted by NSOs. For these reasons, we propose only descriptive statistics.

d) <u>Categorical variables (nominal or ordinal)</u>

d.1) If Y is a categorical variable, the evaluation of the impact of E&I on its univariate distribution can be performed by using the following dissimilarity indices:

$$I_{m1} = \frac{1}{2}\sum_{k=1}^{K}\left| f_{Y_k}^{R} - f_{Y_k}^{F} \right|$$

$$I_{m2} = \left\{ \frac{1}{2}\sum_{k=1}^{K}\left| f_{Y_k}^{R} - f_{Y_k}^{F} \right|^2 \right\}^{\frac{1}{2}}$$

where $f_{Y_k}^{R}$ and $f_{Y_k}^{F}$ are respectively the frequency of category k (possibly weighted) in the reference and in the final datasets. Both these measures assume their minimum value 0 only if the two distributions are exactly the same, while they assume the maximum value 1 when in each distribution all units have the same category of Y but that category is different in the two distributions.

It is obvious that information on changes of marginal distributions is also provided in the transition matrices so far introduced.

d.2) Graphical representations of the frequency distributions of Y in the two compared data sets permit the visual analysis of the changes of distributions due to E&I (see point a.5 in subsection 2.1).

d.3) If Y and X are two categorical variables, the evaluation of the impact of E&I on their joint distribution can be performed by using the two following indices:

$$I_{j1} = \frac{1}{2}\sum_{y}\sum_{x}\left| f_{yx} - \tilde{f}_{yx} \right|,$$

$$I_{j2} = \left\{ \frac{1}{2} \sum_y \sum_x \left| f_{yx} - \tilde{f}_{yx} \right|^2 \right\}^{\frac{1}{2}}$$

where f_{yx} and $\tilde{f}_{yx}$ are respectively the frequencies (possibly weighted) of the contingency table obtained by crossing the categories of *Y* and *X*. The indices assume values in the interval [0,1] and they can be easily extended to any set of *k* variables ($k \geq 2$) to make analyses on multiple distributions. Note that for variable combinations assuming few categories, I_{j1} and I_{j1} can have a value lower than the value corresponding to variable combinations having many categories. Therefore, these indices are useful when comparing results produced by different E&I methods on sets of variables with the same categories.

e) <u>Continuous variables</u>

e.1) If *Y* is a continuous variable, 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, and let

$$F_{Y_n^R}(t) = \frac{\sum_{i=1}^{n} w_i \times I(Y_i^R \leq t)}{\sum_{i=1}^{n} w_i}, \quad F_{Y_n^F}(t) = \frac{\sum_{i=1}^{n} w_i \times I(Y_i^F \leq t)}{\sum_{i=1}^{n} w_i}$$

be the weighted empirical distribution functions in the reference and final datasets. The *KS* distance is defined as:

$$KS(F_{Y_n^R}, F_{Y_n^F}) = max_t \left| F_{Y_n^R}(t) - F_{Y_n^F}(t) \right| = max_j \left| F_{Y_n^R}(t_j) - F_{Y_n^F}(t_j) \right|$$

where the $\{t_j\}$ values are the ordered values of *Y*. It results *KS*=0 only when $F_{Y_n^R}(t_j) = F_{Y_n^F}(t_j) \ \forall \{t_j\}$.

e.2) A set of statistics can be used in a straightforward way to obtain information on distributions changes, like the variable *mean* and *standard deviation*, the distribution *quartiles*, the *minimum* and *maximum* values of the variable *Y*.

e.3) A deepest analysis of changes in the marginal distribution of *Y* in the reference and final data sets can be performed through graphical representations of data like those mentioned in subsection 2.1 (see point c.5). Useful univariate statistics are directly provided in graphs.

2.3 Preservation of aggregates

The impact of E&I on statistical aggregates is generally evaluated in terms of: 1) non-sampling components of the variance of the aggregate estimate due to non sampling errors and E&I activities, and 2) distance between the final estimate of the aggregate and the corresponding original one.

Concerning the first aspect, 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 these 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 at this stage of research, they represent a critical area for possible future developments.

Concerning point 2), it does not take into account the different mechanisms or models underlying the generation of errors (missing and inconsistent values). In this context, the evaluation can be carried on by simply analysing the differences between point estimates of the target parameters, before and after E&I.

For continuous variables, some standard aggregates (*sum of variable values, standard deviation, mean,* etc.) are directly provided by IDEA (see point c.5 in subsection 2.1, and points e.2 and e.3 in subsection 2.2).

2.4 Preservation of relations

Measuring the effects of E&I on data relations is a relevant problem to be considered when evaluating E&I. In the area of data imputation, fundamental references are the papers from Kalton *et al.* (1982), Kalton *et al.* (1986) and Little (1986). Recent research relates particularly to regression imputation (Shao *et al.*, 2002), but further research is needed with respect to other imputation techniques. One traditional way of evaluating this aspect consists in analysing changes produced by E&I on the structure of data relations of reference data.

f) Categorical variables (nominal or ordinal)

f.1) If *Y* and *X* are categorical variables subject to E&I (either both or only one of them), the *Cramer contingency coefficient* (Kendall *et al.*, 1979) before and after the E&I process is proposed for evaluating changes in the bivariate relation between them used. For each compared data set, the *Cramer contingency coefficient* is based on the χ^2 index computed on the basis of 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.
The index is equal to 0 when *Y* and *X* are not associated, while it is equal to 1 in case of complete association. Note that the index I_3 previously introduced (see subsection 2.2, point d.3) provides summary information about overall E&I effects on multivariate relationships.

g) Continuous variables

g.1) For the evaluation of the effects of E&I on bivariate relations between two continuous variables *Y* and *X* the usual measures (covariances and correlations) in the reference and in the final data sets can be used. In each data set, for each couple of items subject to E&I the following measure are then proposed:
- the *matrix of covariances*,
- the *matrix of Pearson correlations indices*.

3. documenting editing and imputation processes

The quality indicators described in section 2 refer to the evaluation of the E&I process. However, further standard quality indicators on E&I have been defined at Istat with documentation purposes. These indicators can also be calculated by IDEA, but they should then be imported and stored in the Information System for Survey Documentation (SIDI). As already mentioned, SIDI is the centralized information system for standard documentation of the whole data production process (from the design to the dissemination stage) for each Istat survey. Both a qualitative and a quantitative perspective are considered in SIDI. The documentation of E&I procedures and the standard quality indicators related to E&I are included among the more extensive information available in the SIDI system. In order to understand the framework for documentation implemented at Istat, the SIDI system is described in the next subsections. Besides a summary description of the whole system (subsection 3.1), particular attention is devoted to the SIDI standard quality indicators related to E&I and the interaction between SIDI and IDEA.

3.1 The SIDI system

Improving and documenting the quality of survey processes and products is a major concern for many NSOs.

The SIDI system has been designed and implemented in order to support the survey managers in such activities (Brancato *et al.*, 2004).

The main purposes of the SIDI system are:

- to support the survey managers in monitoring and documenting the quality of their surveys;
- to compare the quality of different surveys;
- to provide Istat management with periodic reports concerning the quality of Istat surveys.

A very important characteristic of the SIDI system is that it manages both metadata and quality indicators in an integrated way. Indeed, besides the documentation purposes, metadata are also necessary in order to correctly interpret the quality indicators. For instance, the quality of the editing and imputation phase can be better analyzed if quantitative indicators are integrated by some information concerning the editing and imputation procedures which have been applied.

SIDI manages the following classes of metadata: i) metadata related to the information content of the survey (e.g. observed phenomena, statistical units,…); ii) metadata related to the production process, in particular the description of all the performed survey operations and of the survey quality control activities which have been carried out in order to prevent nonsampling errors, to correct them and to estimate their impact on final estimates; and iii) relevant documentation concerning specific surveys (e.g. internal documents, instructions for manual editing,…) as well as general documents (e.g. Eurostat regulations and documents on quality, …).

It is worth underlining that the quality indicators managed in SIDI are process oriented. In particular, they have been defined in order to allow the survey managers to monitor and to analyze over time the quality of the main survey phases. Indeed, the process-oriented indicators are less expensive to be calculated and are therefore a good starting point in order to create a common and standard level of quality measurement within an Institution. It can be added that sometimes process indicators might provide some insight on product quality and that improving the process quality has been largely recognized as one tool to achieve a better product quality (Lyberg et alt., 2001). To this purpose a set of quality indicators has been defined for the following survey aspects: frame; data collection, data entry, editing and imputation, timeliness, costs, and coherence. The quality indicators related to the editing and imputation phase are described in detail in subsection 3.3. Furthermore, the quality indicators are to be calculated for each survey occasion. Therefore, the system manages time series of each group of indicators and is equipped with a large set of graphical and tabular representations.

Finally, we note that in order to allow comparisons among different surveys both metadata and quality indicators have been highly standardized.

3.2 Implementing the SIDI system

The implementation of the SIDI system started in 2001. As it is well known, implementing and maintaining up-to-date information systems is a demanding task. Even if documentation is recognized as an important aspect of quality, it is a time-consuming activity and survey managers do not usually consider it as part of their current production work (Blanc *et al.*, 2001). Therefore, it is important to have a strategy for populating and keeping information systems updated.

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 groups of standard quality indicators for 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, their calculation implies 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.

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:

1. The net of quality facilitators. The quality facilitator is a new professional role for Istat which has been introduced for supporting the release of SIDI. More in detail, they 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 in the Institute. Up to now, some 65 people have already been trained.
2. The development of generalized software to support the survey managers in the computation 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 can be considered similar with respect to the evaluation of a given set of standard quality indicators. For example; we can mention 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 generalized procedure for the calculation of the SIDI indicators related to this phase. The availability of generalized software for calculating quality indicators is one of the major supports 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 in order to satisfy also these purposes and it is now currently used to provide quality indicators for the SIDI system (see subsection 3.4)
3. 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 information on data collection related to the multipurpose surveys.

3.3 SIDI Quality indicators for Editing and Imputation

With regard to the editing and imputation phase, survey managers are asked to provide metadata 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).

The SIDI set of standard quality indicators on E&I is described in table 1.

As a result of the process oriented approach, all the required indicators can be obtained by comparing the raw and the final data matrices. This means that every survey that uses an E&I procedure can provide the SIDI indicators, regardless to the characteristics of the procedure (i.e. deterministic or stochastic).

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 indicators 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 the major problems are in the overall quality of collected (raw) data. A high percentage of net imputation indicates the presence of item non-response problems. Otherwise, a high percentage of modification is a sign that erroneous values are the main problem; it can then be 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 subset of indicators can be useful in order to understand the behaviour of the E&I procedure: for example, 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.

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 a 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 is 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 two principal sets of quality indicators, one weighted and one 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 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[5].
4	Imputation Rate	Values modified by E&I / potentially imputable values[6]
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 variables to the left
17	Third Quartile of Imputation Rate Distribution by VARIABLE	Value of the imputation rate leaving the 75% of the ordered variables to the left
18	Number of Variables with an imputation Rate greater than 5%	
19	Number of Variables with an Imputation Rate greater than 2%	
20	First Quartile of Imputation Rate Distribution by RECORD	Value of the imputation rate leaving the 25% of the ordered units to the left
21	Third Quartile of Imputation Rate Distribution by RECORD	Value of the imputation rate leaving the 75% of the ordered units 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%	

For the principal sets 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

[5] Some variables might be excluded from the imputation process (e.g. identification codes)
[6] Potentially imputable values= Total records* Total imputable variables

geographical detail. The system offers graphical representations (maps) of the indicators that permit subject matter experts to identify troubles in particular geographical areas. Through this functionality it is possible to perform a 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 indicators. Using this functionality, they are able to evaluate the performance of the E&I procedure and its impact on their data through consecutive survey occasions. Finally, it is possible to make different types of comparisons for better evaluating the quality of survey data:

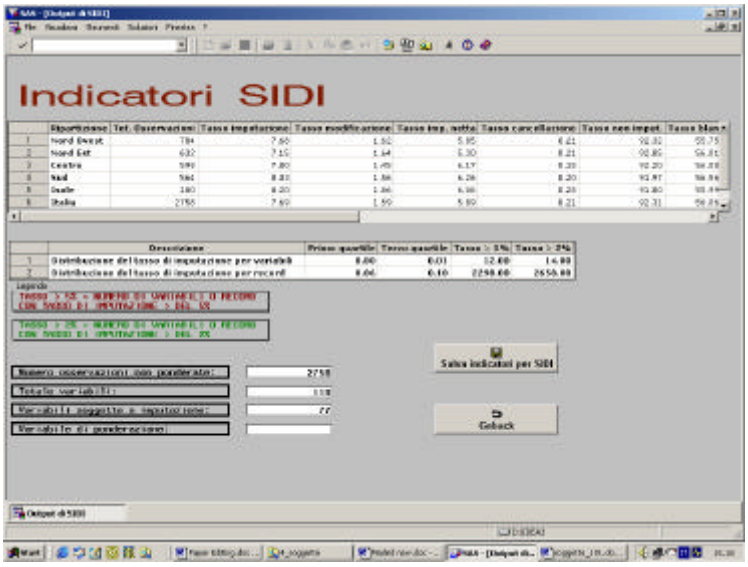
1. For a given survey, the system permits to compare the value of an indicator with a general mean value, obtained averaging the values of the same indicator for all surveys.
2. Furthermore, the comparison can be done with particular averages calculated within subgroup of surveys that use the same E&I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

3.4 The role of IDEA for SIDI quality indicators

In order to assure that SIDI quality indicators are calculated in the same standard way by each survey, quality facilitators are asked to provide in SIDI the numerators and the denominators (input values) needed for calculating the quality indicators and the system itself calculates them. After that, the 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, the input values to be inserted in SIDI are reported in table 1. To this purpose, quality facilitators need to compare raw and clean data sets for the same survey occasion. This would have required the quality facilitators to prepare some ad hoc programmes in order to calculate the input values for SIDI, implying possible errors in computations and loss in timeliness. 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 quality facilitators in this specific task was to develop generalized software which could automatically produce the input values for SIDI by comparing raw and clean data sets. The software IDEA easily provides such values in a standard way for all Istat surveys.

A quality facilitator who wants to use IDEA to calculate SIDI indicators has only to provide the raw and clean SAS datasets with the following requirements: i) corresponding variables in the two datasets should have the same names; ii) both datasets should contain variable representing the records' id; iii) a variable that identifies the geographical area should be present in at least one of the two datasets. In addition, if the weighted set of indicators is calculated, a variable containing the weights should be present in the clean dataset. Such organized datasets are generally already available or easily obtainable for each survey. Starting from this input, IDEA provides an output file containing the SIDI input values, in a specific format that can be directly imported in SIDI. Furthermore, IDEA displays the SIDI indicators, and therefore the quality facilitator can control the values before importing them in SIDI (Figure 4).

Figure 4: Output of IDEA for SIDI



Other relevant typologies of quality indicators can be calculated considering the SIDI indicators computable with IDEA as a starting point.

In particular, the un-weighted imputation rate proposed by the Eurostat Task Force on Standard Quality Indicators (Linden and Papageorgiou, 2004) corresponds to the SIDI imputation rate obtained considering only one variable as "potentially imputable".

Furthermore, it is possible to use IDEA to compare datasets related to subgroups of units, or to specific intermediate sub-phases of the E&I procedure. The resulting indicators could then be used for more specific evaluation purposes by survey managers and could also be stored in local databases.

IDEA has gained a big success among the quality facilitators: all the ones that have provided the indicators on E&I in SIDI have used IDEA for all the survey occasions. In addition, the use of IDEA for SIDI stimulates them to perform further analysis. For example, when big changes in data were enhanced by SIDI E&I indicators, they tried to understand in which phase of the production process the errors corrected by E&I where mostly originated, and, if possible, took actions in order to avoid them in successive survey occasions.

Finally, it is worth mentioning the added value of having the possibility of performing the documentation activity required by the SIDI system for the E&I phase with the same generalized software that can be used to tune and monitor the E&I process. In fact, given the potentialities of IDEA, once it has been used to calculate SIDI indicators, the other available measures for evaluating the quality of E&I or assessing its impact on data distributions and relations are often calculated by the users, thus stimulating a quality assessment that would not have been performed without this powerful tool. Consequently, the availability of IDEA facilitates the standard documentation in SIDI and the fact that SIDI has to be populated stimulates the use of IDEA as a tool for the assessment of the production procedure.

5. ConcluDING REMARKS AND FUTURE WORK

Information on the quality and the impact of E&I is not only a requirement at NSOs level, but also a powerful tool that survey managers 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. In this paper we proposed an evaluation framework in which different indices and statistical measures are defined, based on specific evaluation (or performance) criteria, which can be used in different evaluation contexts. These measures are partially inspired to the evaluation measures used in the EUREDIT project. Further indices correspond to the standard quality indicators required by the Istat information system SIDI. In order to simplify and make more efficient the evaluation task, as well as for supporting subject matter experts, the generalized software IDEA for computing all the proposed indicators has been developed. IDEA is

currently used by Istat survey managers because of its usefulness in terms of standardization and simplification of the calculation process.

Further software developments are planned, particularly relating to the evaluation of E&I effects on multivariate distributions. Appropriate summary measures are also needed to better evaluate the E&I impact on estimates like totals, means, variances. Finally, additional graphical representations of data and data distributions will be implemented to improve the effectiveness of the evaluation.

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., Pellegrini C., Signore M., Simeoni G. (2004). Standardising, Evaluating and Documenting Quality: the implementation of Istat Information System for Survey Documentation – SIDI, *European Conference on quality and methodology in Official statistics*, Mainz, May 24-26 2004, CD-ROM

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.

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

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.

Linden H., Papageorgiou H. (2004). Standard Quality Indicators, *European Conference on quality and methodology in Official statistics*, Mainz, May 24-26 2004, CD-ROM

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.

National Center for Education Statistics (1992). *NCES Statistical Standards.*

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

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