

**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

**EVALUATING THE QUALITY OF EDITING AND IMPUTATION: THE SIMULATION
APPROACH**

Invited Paper

Submitted by ISTAT, Italy¹

I. INTRODUCTION

1. In statistical surveys non-sampling errors correspond to deviations of the collected data of the surveyed variables from the corresponding actual (*true*) values. These errors may derive from different sources, like response, measurement, data processing. It is well known that non-sampling errors may account for the greater proportion of the total error. For this reason, non-sampling errors are to be detected and eliminated from data. *Editing and imputation* (E&I) generally indicate the set of activities done at the post-data capturing stage for the identification and treatment of errors.

2. In order to evaluate the quality of an E&I process we need to define which component we are interested in and the indicators to quantify it. In this paper we focus of the capability of an E&I process of correctly detecting errors and eliminating them by restoring the corresponding true values without introducing new errors in data (*accuracy*).

3. The evaluation of the accuracy of an E&I method can be performed only when, for a given set of raw data, the corresponding *true* values are known. In this situation the accuracy of E&I can be assessed by comparing the *final* (edited and imputed) data with the corresponding *true* ones.

4. In this setting, a first problem is how to obtain true and raw data. One of the possible approaches consists in artificially generating them. In this paper we discuss some aspects relating to the use of the simulation approach for obtaining artificial data to be used for evaluation purposes.

5. The second main problem is identifying appropriate statistical measures, based on the comparison of final data with the corresponding true ones, for the assessment of the E&I capability to deal with the specific types of non-sampling errors contaminating data. Different measures can be defined depending on the evaluation aim: indicators measuring the capability of E&I of correctly identifying and correcting errors (*micro* level indicators), and other measures for assessing the statistical impact of E&I on target estimates (*macro* level indicators). In this paper a set of micro level indicators are proposed and discussed.

6. A software for evaluating E&I procedures based on the simulation approach has been developed at ISTAT. This software, called ESSE, allows simulating sets of raw data by the controlled contamination

¹ Prepared by Marco Di Zio, Ugo Guarnera, Orietta Luzi, and Antonia Manzari.

of given sets of true values based on predefined error models/mechanisms. Micro level indicators are directly provided based on the adopted simulation strategy and the E&I results.

7. The paper is structured as follows. The experimental framework based on the use of simulation, as well as its statistical implications are illustrated in section II. The artificial generation of synthetic data is discussed in section III. Some aspects relating to obtaining *true* and/or *raw* data are discussed respectively in sections IV and V. A set of quality indicators measuring the E&I accuracy at micro level are discussed in section VI. In section VII the error models available in the software ESSE for obtaining artificial raw data are illustrated.

II. EVALUATING EDITING AND IMPUTATION METHODS: THE SIMULATION APPROACH

8. In specialized literature, E&I is viewed not only as the set of methods, activities, and tools aiming at identifying and eliminating errors from a set of statistical survey data, but also as a possible source of non-sampling errors that may cause an estimate not to be in perfect agreement with the parameter it is supposed to estimate (Rancourt, 2002), and, if it contains stochastic elements, as an additional source of data variability. Therefore, the problem of measuring the *quality* of E&I is twofold: 1) assessing its capability of correctly dealing with non-sampling errors affecting data, i.e. its accuracy, and 2) measuring its impact on target estimates.

9. When dealing with the problem of designing and implementing an optimal E&I procedure, the natural question to be answered relates to how it is possible to evaluate the quality of E&I. Let X_1, \dots, X_p be the p variables of interest, i.e. the set of phenomena to be investigated by a specific survey process. The quality of an E&I method can be evaluated when it is possible to compare the final data with the corresponding true ones, i.e. when the following three sets of data are available:

- 1) the *true* data, i.e. the $n \times p$ matrix X of the actual values of the p variables on the n units;
- 2) the *raw* data, i.e. the $n \times p$ matrix Y corresponding to contaminated data;
- 3) the *edited* data, i.e. the $n \times p$ matrix X' of final (edited and imputed) values resulting from the application to Y of the E&I method under evaluation.

10. In this framework, given the target phenomenon, under predefined assumptions on the error typologies possibly contaminating them, designing an evaluation study for a specific E&I method implies the definition of the following main elements:

- 1) how to obtain true data;
- 2) how to obtain raw data;
- 3) how to measure the E&I accuracy.

11. Relating to the first two aspects, different approaches have been proposed in literature. The choice among them mainly depends on the available resources and evaluation objectives (Granquist, 1997; Norbotten, 2000).

12. For true data, a number of alternative solutions can be adopted. As errors arise mainly at the data collection and data entry phases, one possible way to obtain more accurate data could be to repeat these activities under better conditions (e.g. using professional interviewers, computer assisted interviewing, reconciliation of current answers with previous ones, etc.) (Lesseler *et al.*, 1995). The main disadvantage of this approach is that it is rather expensive, in terms of both resources spent and respondent's burden. Furthermore in spite of whatever action taken, a given amount of errors will continue contaminating data. For these reasons this approach can seldom be adopted.

13. A set of *true* data might be obtained by linking information coming from appropriate alternative sources of information (statistical or administrative) (Poulsen, 1997). The problem here is the difficulty of obtaining data from the same sampling units, on the same phenomena of interest and for the same period of time.

14. An alternative solution is to define as *true* an artificially generated set of data. In this case, the quality of the performed analyses strongly depends on the models chosen to generate data, in particular on how well these models represent data.

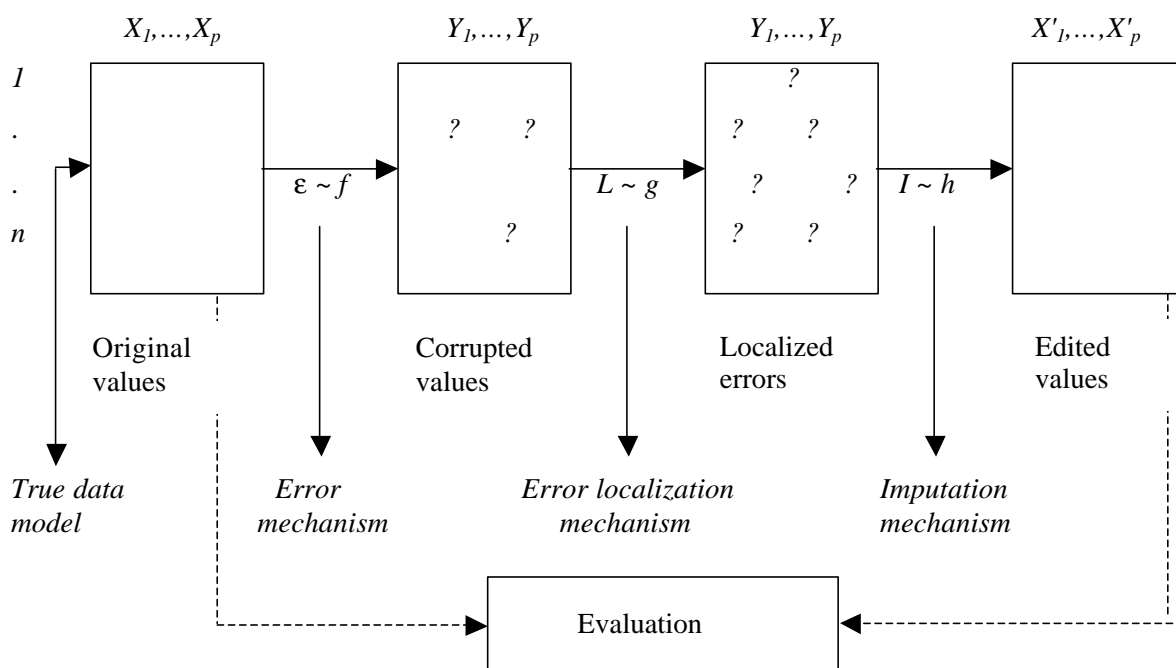
15. Concerning raw data, these can correspond to either observed survey data, or to artificially generated data. In the latter case, starting from a set of true data, pre-defined rates of pre-specified types of errors are introduced based on specific error mechanisms (Barcaroli et al., 1997).

16. For both true and raw data, when using the simulation approach the goal is investigating the phenomenon of interest having a certain control on the mechanism ruling some key aspects of the phenomenon itself. It means that we simulate the situation we want to investigate, and use the obtained results for assessing the quality of either a single technique or an overall E&I procedure in the specific simulated situation. The main advantages of the simulation approach with respect to other ones are low costs and the possibility of controlling some critical aspects of the applications. Drawbacks mainly relate to the effort required for modelling the data and/or errors. The simulation approach is also useful for the comparative evaluation of different E&I methods or procedures. This is the case of the research activities carried out in the context of the EUREDIT project (<http://www.cs.york.uk/euredit/>) where, starting from predefined sets of data assumed as true, raw data sets have been artificially generated in order to perform comparative evaluations of different E&I techniques.

17. Concerning the problem of how to measure the E&I accuracy, the comparison of the final data with the corresponding true ones allows to compute indicators that summarize the performance of the E&I process at the *micro* or reported level, as well as the *macro* or aggregate level (Garcia et al., 1994; Stefanowicz, 1997; Chambers, 2001). In the first case the measures are based on the changes occurred in the individual data items while in the second case the focus is on the changes occurred in the estimates and/or distributions. Both types of measures can be computed and a hierarchy (in terms of relevance) generally depends on the level at which the data are released.

18. In order to understand better the evaluation problem in a simulation context, as well as its statistical implications, it is useful to graphically represent the overall data and process flow characterizing the simulation-based evaluation framework. In Figure 1 (Di Zio et al., 2002) all the stochastic elements influencing the final results are depicted.

Figure 1: Evaluating editing and imputation processes in a simulation context



19. Suppose for the sake of simplicity that the initial data are fixed. In our simulation context, the X' values can be thought of as a realization of a random vector from a multivariate probability distribution depending on different random mechanisms: the corruption of the original values with probability law f (*error and missing data mechanism*), the error localization method (with probability law g when it is not deterministic) and the imputation method (with probability law h when it is not deterministic).

20. To statistically evaluate the quality of the E&I process we generally compare two quantities $Q(X)$ and $Q(X')$, where $Q(\cdot)$ is a generic function of the data (possibly vector valued), through a suitable generic loss function $d(Q(X), Q(X'))$ taking into account all the different stochastic mechanisms affecting the results. One natural way of doing that is to compute the expected value:

$$E_{e, L, I} [d(Q(X), Q(X'))] \quad [1]$$

where (e, L, I) represents the random mechanisms as in Figure 1. For instance, for X univariate, if we define $Q(X)=X$ and, for each pair of n -dimensional vectors \mathbf{u}, \mathbf{v} , $d(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^n d(u_i, v_i)$, where

$d(s, t) = \begin{cases} 1 & \text{if } s \neq t \\ 0 & \text{otherwise} \end{cases}$, we obtain the basic elements of the class of micro indicators described in section 6. An analytical computation of the expected value [1] is generally not feasible because of the difficulty of modelling the joint distribution of the random variables X' . On the other hand, we are often able to draw observations from the joint distribution of the random variable $X'(e, L, I)$. In this context we can use a Monte Carlo integration method to approximate the expected value [1] (Rubinstein, 1981).

21. Given the hierarchical process represented in Figure 1, if we are interested in assessing the impact on final results of each single stochastic mechanism, we can decompose the expected value [1] in parts corresponding each to the effect induced by each mechanism present in the process. Most of the methods proposed in literature deal with the problem of estimating the components due to missing data and to imputation (see for example Lee *et al.*, 2002; Rubin, 1987). Concerning editing, Rancourt (2002) and Ghosh-Dastidar *et al.* (2003) discuss the problem of evaluating the component due to error localization.

22. One way of estimating the effects of each random mechanism affecting final results (including E&I) is based on the use of a Monte Carlo approach, consisting in iterating the editing and/or the imputation methods under evaluation conditionally to the other stochastic mechanisms that hierarchically precede them. For example, if we aim at measuring the components due to a specific missing mechanisms and to a given imputation model, we have to iterate k times the contamination process under the assumed missing model (thus generating k raw data sets) and, conditionally to each contaminated data set, we have to apply j times the imputation method under evaluation. Analyzing the resulting distributions of the predefined quality indicators, the different contributions (e.g. in terms of bias and variance) of the various mechanisms taken into account in the experimental study can be measured.

III. SIMULATION OF SYNTHETIC DATA

23. In this section we discuss issues related to data simulation. Let (X_1, \dots, X_p) be a random variable following the probability function $F(x_1, \dots, x_p; \mathbf{q})$.

If we know $F(x_1, \dots, x_p; \mathbf{q})$, by simply generating random sample from it, it is possible to use Monte Carlo techniques as discussed in section 2 (Devroye, 1986).

Of course, more real and interesting is the case when $F(x_1, \dots, x_p; \mathbf{q})$ is unknown. In this case we can replace $F(x_1, \dots, x_p; \mathbf{q})$ with its estimate $\hat{F}(x_1, \dots, x_p; \mathbf{q})$ (having a random sample $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$, where $\mathbf{x}^{(i)} = (x_1^{(i)}, \dots, x_p^{(i)})$, available) and draw samples using this estimated p.d.f.. Different techniques can be used to estimate F , and we mainly divide them in parametric and non-parametric.

24. In the parametric approach, obviously it is needed to first specify a parametric form for the distribution $F(x_1, \dots, x_p; \mathbf{q})$, e.g. the Gaussian distribution. Then the parameters \mathbf{q} can be estimated, for instance using the maximum likelihood estimators, thus obtaining $\hat{F}(x_1, \dots, x_p; \hat{\mathbf{q}})$. Finally B random samples from $\hat{F}(x_1, \dots, x_p; \hat{\mathbf{q}})$ can be drawn in order to compute the quantity of interest.

In the non-parametric approach, we can obtain a random sample from the empirical distribution $\hat{F}_n(x_1, \dots, x_p; \mathbf{q})$ by drawing B random samples with replacement $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(B)}$ from the initial sample $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}$.

25. It is evident the connection of the above data generation methods with the re-sampling method techniques. In particular they are related to the bootstrap techniques, where theory is established to compute estimates and variance of estimates of target quantities (Efron *et al.*, 1993; Efron, 1994).

26. Following the scheme in Figure 1, $F(\mathbf{q})$ is the result of different random mechanisms that interact in a very complex way. It results that building step-by-step $F(\mathbf{q})$ is in turn a very complex task. The re-sampling methods allow starting from the end of the process (the sample observed from $F(\mathbf{q})$). For instance if we want to evaluate a method by varying the processes of generating data and non-responses, instead of specifying the two random mechanisms and the interaction between them, we can resample the initial sample (comprehensive of missingness) in order to reproduce the "real" mechanisms generating them.

IV. OBTAINING TRUE DATA

27. Following the previous section, the data generation can be made either through a non-parametric or a parametric approach. While the non-parametric data generation is clear, more words are needed in the case of parametric generation, especially when data has to mime "true" data. Apart from the classical Gaussian and multinomial distributions, we want to focus on the situation when data distribution is asymmetric. This happens in most of the surveys dealing with continuous variables, for instance incomes, expenditures. A first way of handling this issue is through the use of the log-normal distribution. The log-normal distribution refers to data that after the logarithmic transformation can be considered normal. From a practical point of view simulating this type of data consists in generating data from the normal distribution and transforming them through the exponential function. In multivariate context exponential transformations are generally component-wise and it is difficult to obtain data having the desired distributional characteristics in terms of joint relations between variables. However other asymmetric multivariate distributions have been introduced (see Johnson *et al.*, 1972), and it is worth mentioning them. For their tractability, a special mention is due to hyperbolic distribution Barndorff-Nielsen *et al.* (1983) and the multivariate skew-normal and skew-t distribution (Azzalini *et al.*, 1999; Azzalini *et al.*, 2003). In a practical context we remark that for the hyperbolic distribution, for skew-normal and for skew-t distributions specialized libraries have been implemented in the open source code R (<http://cran.r-project.org/>). In particular, packages are available for parameters estimates of these distributions. This is particularly useful when, in a simulation context, the parameters of the distributions to be generated have to be estimated from some set of available data and the log-normality cannot be assumed.

V. OBTAINING RAW DATA

28. One of the most critical issues in evaluating E&I procedures based on the simulation approach, is generating raw data. Since, as already noticed, a "best" E&I method in all practical situations does not exist, in the experimental context we have the problem of setting up "typical" situations in order to evaluate different methods for specific error typologies. To this aim we need modelling error as well as non-response, or, in other words, we have to model the distribution of the raw data conditional on the true ones.

29. As far as non-response is concerned, a first issue to deal with is the nature of the mechanism that generates missing data. In fact it is well known that most of the imputation procedures assume, more or less explicitly, at least MAR mechanism for non-response (Little *et al.*, 2003). One of the most common ways for simulating a MAR mechanism on a data-set with n observations and p variables consists of drawing from a suitably specified probability distribution $P(\mathbf{R}|\mathbf{Z};\mathbf{J})$ where \mathbf{R} is a random matrix whose entries represent the response indicators (i.e. $R_{ij}=1$ if the j -th variable on the i -th observation is observed and 0 otherwise), \mathbf{Z} is a set of covariates without missing values and \mathbf{J} is a set of parameters. Simulating a non-MAR mechanism could be also useful in order to check the robustness of E&I strategy with respect to departure from the ignorability assumption. Furthermore, in some special cases (Tang *et al.*, 2003), consistent data analyses can be performed even in the presence of a non-ignorable non-response mechanism.

30. Artificial perturbation of data through introduction of errors is an even more difficult task. In fact, unlike missing values, measurement errors are characterized not only by their occurrence pattern (i.e. the set of items in error), but also by their “intensity”, that is the extent (generally expressed in terms of ratio or difference) to which erroneous data deviate from true ones. Hence, modelling measurement errors generally requires more care than modelling non-response. One approach, that could be defined non-parametric (Barcaroli *et al.*, 1997), consists of introducing errors according to some known mechanism such as *unity measure errors* in the compilation phase or *digits permutation* at the data entry stage. Specialized software have been used in ISTAT to simulate easily some of the most common non-sampling errors in the different survey phases (Luzi *et al.*, 1998; Manzari *et al.*, 2000; Della Rocca *et al.*, 2000). This approach has the advantage of being independent of explicit parametric model assumptions for the error mechanism. Another commonly adopted approach consists of drawing samples from a probability distribution $P(Y|X; \xi)$ of the corrupted values conditional on the true ones (ξ denotes a set of parameters). In this context error typology reflects on the distributional characteristics of $P(\cdot)$. For instance, in case of quantitative data, “random errors” are represented by probability distributions $P(Y|X; \xi)$ such that $E(Y|X; \xi) = 0$, and *independent errors* are associated with distributions that can be factorized into as many univariate distributions as the number of involved variables. In many cases (Hansen, 1953) measurement errors are modelled through an additive mechanism $Y_i=X_i+e_i$ where typically it is assumed that the errors have zero mean and are independent of each other. Often independence between e_i and X_i is also assumed. This mechanism can be easily simulated by generating errors from a normal distribution with zero mean and diagonal variance-covariance matrix. Straightforward extensions are obtained by considering non-diagonal variance-covariance matrices and non-zero mean vectors.

31. Another important issue in modelling and simulating non-sampling errors is the assumption of the “intermittent” nature of errors. In the additive model, this means that the e_i 's are generated by semi-continuous distributions or, in other words, the distribution of the corrupted data is a mixture whose components are associated with the different error patterns. In this context, the objective of an editing procedure can be viewed as that of assigning each observation to one pattern, that is, to “localize” items in error. In many editing procedures, localization of items in error is performed with the aid of a set of logical or arithmetical rules (*edits*). For each record failing one or more edits, the procedure tries to identify the items responsible for edit violation. Typically this is accomplished with the aid of some further ad hoc assumptions such as the *minimum change principle* (MCP) according to which the minimum number of items has to be changed in order the record pass all the edits. Based on the MCP, a lot of methods and software have been developed and are being widely used in the context of Official Statistics (Fellegi *et al.*, 1976; Kovar *et al.*, 1988). They all assume that errors in data are comparatively rare, so that better accuracy is obtained by changing data as little as possible. Thus, the MCP implies particular assumptions on the error mechanism. For instance, in presence of three numerical variables (X_1, X_2, X_3) related by the equality $X_1+X_2 = X_3$, the MCP implies that not more than one variable can be affected by error. Note that in this model errors are not independent of each other. In the simulative approach to the evaluation of an editing procedure, it is interesting to assess the performance of the procedure when the MCP is met by the error mechanism (that is when errors in data are rare) and, on the other hand, to test the robustness of the method in different situations where the assumptions underlying the MCP are no longer valid. In the just mentioned example, the two experimental situations can be

simulated as follows:

a) *no more than one error per record*

- a1) for each record i an error indicator z_i is drawn from a bernoullian distribution $P(z_i)$ with parameter π [$P(z_i) = \pi^{z_i}(1-\pi)^{1-z_i}$]. If $z_i = 1$, then the observation i is to be perturbed, otherwise it is left unchanged;
- a2) for each unit to be perturbed, a value j^* is selected from the set of indices $\{1,2,3\}$ according to the probabilities $\{p_1, p_2, p_3\}$ previously defined;
- a3) the variable X_{j^*} is perturbed by adding to the true value a random error ε_j drawn, for example, from a normal density $N(0, \sigma_j^2)$;

b) *independent errors*

for every unit i an error ε_{ij} is added to each variable X_j ($j=1,\dots,3$) with semi-continuous distribution, i.e. ε_{ij} is zero with probability $(1-\pi_j)$ and $N(0, \sigma_j^2)$ with probability π_j .

32. In the above example, the probability for a variable of being in error as well as the “error intensity” are independent of the variables to be analysed. In analogy to non-response, we could refer to such a situation as a *Error Completely at Random (ECAR)*: in other words, when the error mechanism is ECAR, the subset of units in error are a simple random sample of the whole set of data and the error distribution is independent of the true values. However, as for non-response, it is plausible that units with different characteristics have different probabilities of being in error and that the error magnitude depends on the value of some attribute. For instance, in business surveys, small enterprises are more likely to provide data with lower precision, and in social surveys the income variables tend to have negative bias in units with larger income. In these cases errors are not completely at random, and analogously to the non-response case, we could attempt to introduce the concept of *Error at Random (EAR)* mechanism for referring to situation of independence between errors and true values conditional on some set of not corrupted variables correlated with the contaminated variables.

33. Simulating simple EAR mechanisms is easy. For instance, considering case a) of the previous example, one could generate the z_i 's through a logistic or probit model using a set of k reliable covariates $\mathbf{v} = (v_1, \dots, v_k)$ and, conditional on $z_i = 1$, draw errors from a Gaussian density $N(\mathbf{m}, \sigma_v^2)$ with parameters (\mathbf{m}, σ_v) depending on the covariates \mathbf{v} .

34. A last concern is the extent to which perturbed data are coherent with respect to a prefixed set of logical or arithmetical rules that, plausibly, will be used as edits in most editing strategies. Calibrating the parameters of the error model allows setting the relative frequency of errors determining “out of domain” units with respect to errors that do not affect the eligibility of data.

VI. QUALITY INDICATORS

35. In this section we focus on the evaluation of the accuracy of E&I using some micro indicators based on the *number* of detected, undetected, introduced and corrected errors (Stefanowicz, 1997; Manzari *et al.*, 2000). In particular, the ability of the editing process in detecting the maximum number of errors without introducing new ones (editing as *error detection*) and the ability of the imputation process in restoring the individual true values are evaluated using some “*hit rate*” measures: the higher the proportion of corrected errors on the total, and the less the quantity of new errors introduced, the more accurate is the E&I process. As shown in the following, this approach allows to evaluate the accuracy of the editing and the imputation processes separately considered, as well as of the E&I process as a whole.

36. Other sets of indicators based on the *differences* between values, distributions or aggregates have been proposed (Granquist, 1997; Chambers, 2000), could be used to evaluate the accuracy of E&I. In this case different types of *distance* measures are used to evaluate the ability of the editing process in reducing the total amount of error (editing as *error reduction*) and the ability of the imputation process in preserving individual values, distributions or aggregates.

37. In the following, accuracy indicators are first provided for the editing process and the imputation process separately, and then for the E&I process as a whole. Each indicator ranges from 0 (no accuracy) to 1 (maximum accuracy). This type of indicators has been used in studies aiming at comparing the performance of different E&I methods (Manzari *et al.* 2002; EUREDIT project <http://www.cs.york.uk/euredit/>).

A. Quality of Editing

38. Our basic choice is to consider an editing procedure like an instrument to classify each raw value into one of two states: (1) unacceptable and (2) acceptable. We put in the unacceptable class the missing values, the invalid values and the valid values considered suspicious by the editing methods. The unacceptable values will be imputed, while the acceptable values will not be imputed. Because the information about the actual erroneous/true status of the values is available, we can verify the correctness of every single choice of the editing process. We consider the editing process accurate if it classifies an erroneous value as unacceptable and a true value as acceptable. These concepts suggest evaluating the accuracy of an editing process by scoring its ability in obtain a correct classification separately for the erroneous and the true data and for the whole set of data. In other words, for each class of values, erroneous and true, and for the whole set of values, we consider the probability that the editing process correctly classifies it.

39. Let us consider whichever raw value of a generic variable. We regard the editing decision as the result of a screening procedure designed to detect deviations of the raw values from the original values. Accordingly to the basic concepts of diagnostic tests (Armitage *et al.*, 1971), we can define:

a : the probability of the Type-I error, that is the probability of incorrectly classifying true values as unacceptable;

b : the probability of the Type-II error, that is the probability of incorrectly classifying erroneous values as acceptable.

40. The analogy of editing decisions with diagnostic tests allows us to resort to familiar concepts: the probability to recognize true values as acceptable is analogous to the specificity ($1-\mathbf{a}$), while the probability to recognize erroneous values as unacceptable is analogous to the sensitivity ($1-\mathbf{b}$). The extent of ($1-\mathbf{a}$) and ($1-\mathbf{b}$) indicate the ability of the editing process in correctly classifying true and erroneous values and can be used as measures of the accuracy.

41. We can estimate these probabilities by applying the editing process to a set of known erroneous and true data. Each raw value can be cross-classified into one of four distinct classes: (a) erroneous and unacceptable, (b) erroneous and acceptable, (c) true and unacceptable, and (d) true and acceptable. Suppose that, for a generic variable, the application gives the following frequencies:

		Editing classification	
		<i>unacceptable</i>	<i>acceptable</i>
Actual status	<i>erroneous</i>	<i>a</i>	<i>b</i>
	<i>true</i>	<i>c</i>	<i>d</i>

Where:

a = number of erroneous data classified by the editing process as unacceptable,

b = number of erroneous data classified by the editing process as acceptable,

c = number of true data classified by the editing process as unacceptable,

d = number of true data classified by the editing process as acceptable.

42. Cases placed on the secondary diagonal represent the failures of the editing process, therefore, the ratio $c/(c+d)$ measures the failure of the editing process for true data and estimates \mathbf{a} while the ratio $b/(a+b)$ measures the failure of the editing process for erroneous data and estimates \mathbf{b} .

The specificity ($1-\mathbf{a}$) is estimated by the proportion of true data classified as acceptable:

$$E_{\text{tru}} = d/(c+d)$$

43. Large values (close to 1) of the E_{tru} index indicate high ability of the editing process in preserving the true values. On the contrary, small values (close to 0) indicate that a large proportion of true values have been classified as unacceptable, in other words, the editing process introduces new errors in data.

44. The sensitivity ($1-\beta$) is estimated by the proportion of erroneous data classified as unacceptable:

$$E_{\text{err}} = a/(a+b)$$

45. Large values of the E_{err} index indicate that the editing process is able to detect a large proportion of errors in data. On the contrary, small values indicate that only few errors have been detected by the editing process.

46. Note that the performance of the editing process, and therefore the values of the accuracy indices, is determined by a number of factors. For instance, in case of error localization algorithms making use of edits, a too restrictive set of edits can cause the incorrect classification of some uncommon true values, giving rise to small value of the E_{tru} index. Otherwise, a too loose set of edits can cause a poor ability in localizing errors, determining small values of the E_{err} index. What's more, even in case of well defined edits, the editing process could not detect an erroneous value in a given variable when, having defined some inconsistencies between the erroneous value and the subsets of domains of other variables, the error localization algorithm fails considering true values of other variables as suspicious. In this case there would be a double failure: with respect to the erroneous value of the current variable and with respect to the true values of other variables linked to the current one by the edits. If the failure of the error localization algorithm were systematic, we would observe a low value of the E_{err} index for the current variable together with low values of E_{tru} for at least one linked variable.

47. With regard to the total set of data (true and erroneous), the accuracy of the editing process (for any single variable) can be measured by the fraction of total cases that are correctly classified:

$$E_{\text{tot}} = (a+d)/(a+b+c+d).$$

48. The reader can easily verify that the E_{tot} index is a linear combination of E_{tru} and E_{err} , whose weights are the fraction of the total cases that are true and the fraction of total cases that are erroneous:

$$E_{\text{tot}} = E_{\text{tru}} \frac{(c+d)}{(a+b+c+d)} + E_{\text{err}} \frac{(a+b)}{(a+b+c+d)}.$$

Therefore, the E_{tot} value is strongly affected by the error proportion in data.

B. Quality of Imputation

49. We assume that the imputation process imputes only values previously classified by the editing process as unacceptable (missing values, invalid values and valid values considered suspicious). The new assigned value can be equal to the original one or different. In the first case the imputation process can be deemed as successful, in the latter case we say that the imputation process fails.

50. In the real case it is generally not necessary that the imputed value of a quantitative variable precisely equals the original value to consider the imputation as successful: it could be sufficient that the new value lies in an interval whose centre is the original value. Otherwise, in the case of qualitative variables, we consider the imputation as successful only when the imputed value equals the original one. This means that, in the case of qualitative variables, whenever the editing process classifies as unacceptable a true value we say that the imputation process fails.

51. We refer to the general case of both qualitative and numeric variables. The previous figures a and c can be decomposed in

$$a = a_s + a_f \quad \text{and} \quad c = c_s + c_f$$

giving rise to the following frequencies:

		Editing & Imputation classification		
		<i>unacceptable</i>		<i>acceptable</i>
		<i>imputed with success</i>	<i>imputed without success</i>	<i>not imputed</i>
Actual status	<i>erroneous</i>	a_s	a_f	b
	<i>true</i>	c_s	c_f	d

Where b and d are previous defined while:

- a_s = number of erroneous data classified by the editing process as unacceptable and successfully imputed,
- a_f = number of erroneous data classified by the editing process as unacceptable and imputed without success,
- c_s = number of true data classified by the editing process as unacceptable and successfully imputed,
- c_f = number of true data classified by the editing process as unacceptable and imputed without success.

52. The quality of the imputation process can be evaluated by the fraction of imputed data for which the imputation process is successful. For imputed erroneous values we can compute:

$$I_{err} = a_s/a.$$

Large values (close to 1) of the I_{err} index indicate high ability of the imputation process in restoring the original values.

For imputed true values we can compute:

$$I_{true} = c_s/c.$$

53. The fraction of imputed true values for which the imputation is successful (I_{true} index), is not able to evaluate the inner quality of the imputation process because the imputation consists in changing the raw values and the I_{true} index only measures an artificial result due to the definition of successful imputation for numeric variables (it always equals to zero for qualitative variables). In the overall E&I evaluation, we consider the I_{true} index as measure of the (artificial) counterbalance of the imputation process to the failure of the editing process for numeric true data.

54. In analogy to editing process, also the performance of the imputation process is determined by a number of factors. For instance, if the imputation method takes into account edits and a close set of edits is defined, imposing constraints on the admissible value to impute can help in determining the original value giving rise to a large value of the I_{err} index (and also of the I_{tru} index). The same consideration holds when the range of admissible values to impute is strongly restricted by the characteristics of the domain of the variables to impute. This is true also with a loose set of edits (small value of the E_{err} index), because I_{err} measures only the imputation accuracy for the subset of erroneous values correctly

classified by the editing process as unacceptable. As marginal case, when an erroneous value is recognised in a variable having only two admissible values (e.g the *sex* variable), the imputation of the true value is forced giving rise to the maximum value (1) of the I_{err} index.

55. For the total set of imputed values (erroneous and true) we can compute:

$$I_{tot} = (a_s + c_s) / (a + c)$$

Note that if the imputed values ($a+c$) were only missing or invalid values, the I_{tot} would be suited to estimate the inner quality of the imputation process. Otherwise, as in real situation, when the imputed values ($a+c$) consist of missing values, invalid values and valid values considered suspicious by the editing process, the I_{tot} index under-estimates the inner quality of the imputation process because it also pays for the failure of the editing process.

C. Quality of Editing and Imputation

56. We can now consider the E&I process as a whole. The accuracy of the E&I process with regard to true data can be measured by estimating the probability of not introducing new errors in data. It is the total probability of two mutually exclusive events: to classify a true value as acceptable OR to impute a value close to the original one in case of incorrect editing classification of the true value. It can be estimated by:

$$E\&I_{tru} = E_{tru} + [(1 - E_{tru}) (I_{tru})] = (c_s + d) / (c + d).$$

57. The accuracy of the E&I process for erroneous data can be measured by estimating the probability to correct erroneous data. It is the joint probability of a combination of two dependent events: to classify an erroneous value as unacceptable AND to restore the original value (to impute the erroneous value with success since it has been correctly classified). It can be estimated by:

$$E\&I_{err} = (E_{err}) (I_{err}) = a_s / (a + b).$$

58. For the total set of data (erroneous and true values) the accuracy of the E&I process can be measured by the fraction of total cases whose original value is correctly restored:

$$E\&I_{tot} = (a_s + c_s + d) / (a + b + c + d).$$

Even in this case, it is easy to verify that the $E\&I_{tot}$ value is affected by the error proportion in data:

$$E\&I_{tot} = (E\&I_{tru}) \frac{(c + d)}{(a + b + c + d)} + (E\&I_{err}) \frac{(a + b)}{(a + b + c + d)}.$$

VII. A TOOL FOR EVALUATING EDITING AND IMPUTATION IN A SIMULATION CONTEXT: THE ESSE SOFTWARE

59. ESSE is an experimental software developed at ISTAT for evaluating the quality of E&I procedures in a simulation context (Luzi *et al.*, 1988; Manzari *et al.*, 2000; Della Rocca *et al.*, 2000). The software consists of two main modules: an *error simulation module* for the artificial contamination of a set of true data by a controlled generation of errors under simple error mechanisms; an *evaluation module* providing indicators measuring the accuracy of E&I processes in terms of their capability to detect as many errors as possible, to restore the true values, to avoid the introduction of new errors in data. The set of quality indicators available in ESSE have been described in section VI.

60. Concerning the simulation module, let X be the $n \times p$ data matrix containing the values of p variables observed (or simulated) on n units (*true data*). In ESSE the simulation of *raw data* consists in

the modification of some of the $p \cdot n$ values based on some error models. Actually two types of mechanisms are available in ESSE: *MCAR* (Missing Completely At Random) and *MAR* (Missing At Random) models. The two approaches have been introduced in the area of *non-responses* (Rubin, 1987; Schafer, 1997), but the same stochastic mechanisms can be assumed for any type of non-sampling errors contaminating survey data.

61. In ESSE the user is allowed to control the *impact* of each error model on each variable, i.e. the percentage of values to be modified by using each error model. In the following, the available models under the two mechanisms are synthetically described. The aim is to reproduce as faithfully as possible the most common mechanisms generating errors in the real survey operations.

A. MCAR error models

62. In this context, the user can introduce several typologies of errors, which can be further grouped with respect to the nature of the variable to be treated: qualitative variables, quantitative variables or both.

a) Errors for both qualitative and quantitative variables

Item non-response: the original value of the variable X_j ($j=1, \dots, p$) is replaced by a missing value. The *item non-response* model generates missing values on the basis of a MCAR mechanism.

Misplacement errors: the values of two adjacent variables X_j, X_k ($j, k=1, \dots, p; j \neq k$) of the same type (categorical or continuous) are swapped. This model mimes a particular kind of errors due to either data capturing or data entry.

Interchange of value: if the length of a variable X_j is one digit, its value is replaced with a new one chosen in the variable domain; if the variable length is greater than one digit, a random pair of digits is chosen and swapped. Also this model mimes an error due to either data capturing or data entry.

Interchange errors: the original value of a variable X_j is replaced by a new one chosen in the variable domain. This model principally mimes response errors.

Routing errors: this model mimes the common situation in surveys where skip (or routing) instructions are used in questionnaires to indicate that the answer to a variable X_j (dependent item) is required only for some values of another variable X_k (filter item).

The condition determining the need of filling in a value for the dependent variable X_j has to be defined on the filter variable. For each unit whose condition is true the value of X_j is set to blank (categorical variables) or to zero (continuous variables). For each unit whose condition is false, the true value of X_j is replaced with a new value randomly chosen in the domain of X_j .

b) Error models for quantitative variables only

Loss or addition of zeroes: if the value of X_j ends in zero, this value is re-written adding a zero to or dropping a zero from the original value. This model principally mimes errors due to either data capturing or data entry.

Under-report error: the original value of X_j is replaced by a wrong nonnegative value lower than the true one. This model principally mimes response errors.

Outliers: the original value of X_j is replaced by a value belonging to the range $[\max(X_j), 1.05 \times \max(X_j)]$, where \max is the maximum observed value of X_j .

c) Overall error

Keying error: a digit of the X_j value is replaced by a value lying in the interval [0-9]. This error model is defined as *overall* because it can affect randomly all the variables under analysis. The model mimes errors due to either data capturing or data entry.

In Figure 2 an example of how error models and the corresponding rates are defined in ESSE is shown.

B. MAR error models

63. A separate module in ESSE generates *item non-response* according to a MAR mechanism. This mechanism is mimed by modelling the expected probability of occurrence of a missing entry by means of a logistic model (Agresti, 1990).

64. Let X be the complete data set. Our goal is to introduce a given amount of missing values in one of the available variables, say X_l . Let the probability of an item of X_l to be missing, be dependent on observed data of X_1, \dots, X_k ($k \neq l$) variables (independent items). According to this model, the probability to be missing for the X_l variable in the i^{th} unit, given the observed values x_{1i}, \dots, x_{ki} , is:

$$P(X_{il} = \text{missing} \mid x_{1i}, \dots, x_{ki}) = \frac{e^{b_0 + \sum_{j=1}^k b_j x_{ij}}}{1 + e^{b_0 + \sum_{j=1}^k b_j x_{ij}}}$$

65. Once the parameters values have been properly determined, i.e. when the required expected amount \bar{P} of item non responses for X_l is obtained, missing values are introduced among original data on the basis of the model so far introduced. In Figure 3 an example of how missing values can be simulated in ESSE based on a MAR mechanism is shown.

VIII. CONCLUDING REMARKS

66. In official statistics, the impact of non-sampling errors on survey results is usually not negligible, on the contrary it often accounts for the greatest part of the total error. On the other hand, the E&I activities made on entered survey data in order to deal with errors contaminating them, may to some extent affect survey results.

Figure 2: Generating error models in the software ESSE

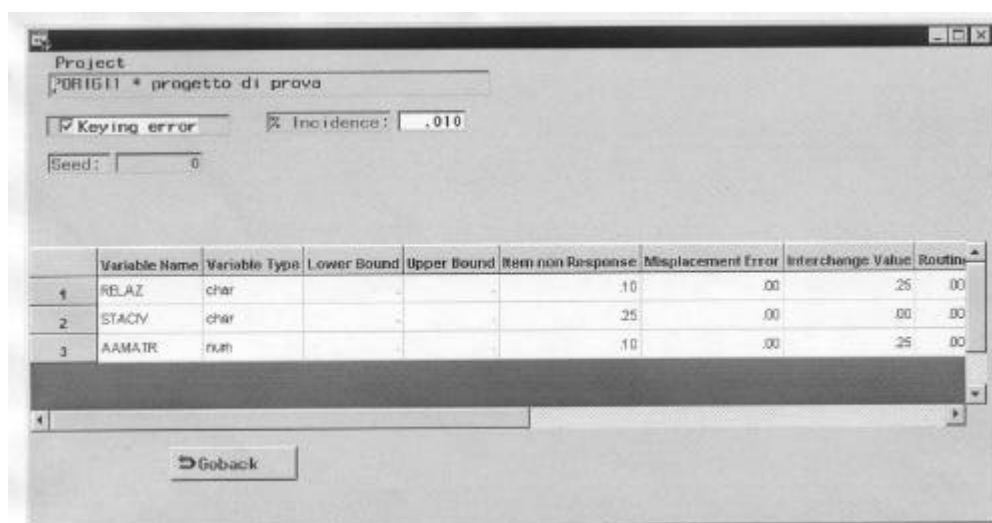
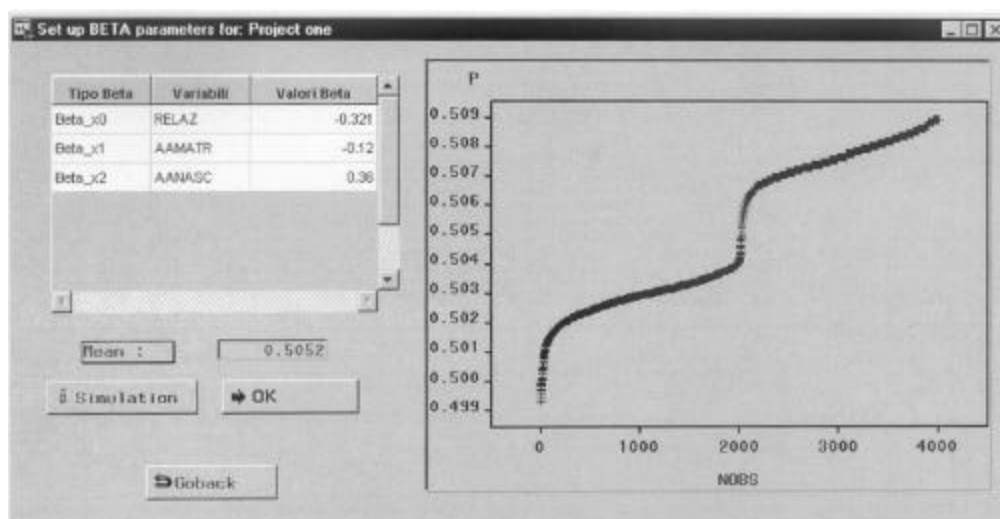


Figure 3: Generating MAR item non responses by using the software ESSE



67. In this paper we focused on the problem of evaluating the capability of an E&I procedure or method of correctly identifying non-sampling errors possibly affecting a given set of statistical data and recovering the true values. This evaluation can be performed only when true data are available for the phenomenon under analysis. In this situation, the evaluation can be performed by using appropriate indicators comparing the final data (obtained by using the E&I procedure under evaluation) and the corresponding true ones. In the paper the specific situation of evaluating E&I in a simulation context was analysed: some problems concerning how to artificially obtain raw and/or true sets of data were discussed, and some specific indicators suitable for assessing the capability of E&I activities of correctly identifying errors and recovering true values illustrated.

68. From the discussion, the need emerges for more studies and further developments from both a theoretical and a practical point of view in the specific area of official statistics. In particular, several aspects should be analysed taking into account the typical characteristics, requirements and constraints of survey processes performed at Statistical Agencies:

- data models suitable for reproducing statistical data in the different survey contexts;

- error models suitable for reproducing the most typical errors and error patterns in the different survey contexts;
- approaches and measures for assessing the quality of E&I taking into account the impact (at different data levels) of the stochastic elements possibly induced by E&I itself.

69. From a practical point of view, we recognize the need to develop appropriate tools implementing algorithms for data simulations and quality indicators computation in order to encourage and facilitate subject-matter experts in performing such type of evaluation tests on their own data processing activities. In this regard, at ISTAT a first experiment has been carried out in developing a prototype software, called ESSE, in which some error models and a set of quality indicators are available for evaluation purposes.

REFERENCES

- Agresti A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons, Inc.
- Armitage P., Berry G. (1971) *Statistical methods in medical research*. Oxford, London, Edinburgh, Boston, Melbourne: Blackwell Scientific Publications.
- Azzalini, A., Capitanio A. (1999). Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B-Statistical Methodology*. Vol. 61, pp. 579-602.
- Azzalini, A., Capitanio A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. *Journal of the Royal Statistical Society, Series B-Statistical Methodology*. Vol. 65, pp. 367-389.
- Barcaroli G., D'Aurizio L. (1997). Evaluating editing procedures: the simulation approach, Working paper, *Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing*, Prague.
- Barndorff-Nielsen O., Blæsild P. (1983). Hyperbolic distributions. In: *Encyclopedia of Statistical Sciences* (ed. N.L.Johnson, S.Kotz & C.B.Read), vol. 3, 700ñ707. Wiley, New York.
- Chambers, R. (2001). Evaluation Criteria for Statistical Editing and Imputation, *EUREDIT Deliverable D3.3*.
- Di Zio M., Luzi O., Manzari A. (2002). Evaluating editing and imputation processes: the Italian experience, *Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing*, Helsinki, May 27-29.
- Della Rocca G., Di Zio M., Manzari A., Luzi O. (2000). "E.S.S.E. Editing System Standard Evaluation", *Proceedings of the SEUGI 18*, Dublin, June 20-23.
- Devroye L. (1986). *Non-uniform Random Variate Generation*. Springer-Verlag, New York.
- Efron B., Tibshirani R.J. (1993). *An Introduction to the Bootstrap*. New York and London: Chapman & Hall.
- Efron B. (1994). Missing data, Imputation and the Bootstrap. *Journal of the American Statistical Association*, vol. 89, No. 426, pp. 463-475.
- Garcia E., Peirats V. (1994). Evaluation of Data Editing Procedures: Results of a Simulation Approach. In *Statistical Data Editing, Vol. 1. Methods and Techniques, Conference of European Statisticians, Statistical Standard and Studies*, No 44, pp.52-68.
- Genton M. G. (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*, Edited Volume, Chapman & Hall / CRC, Boca Raton, FL, 416 pp.
- Ghosh-Dastidar B., Schafer J. L. (2003). Multiple Edit/Multiple Imputation for Multivariate Continuous Data. *Journal of the American Statistical Association*, vol. 98, No. 464, pp.807-817
- Granquist L. (1997). An overview of methods of evaluating data editing procedures. In *Statistical Data Editing, Methods and Techniques, Vol. 2. Conference of European Statisticians, Statistical Standard and Studies* No 48, UN/ECE, pp. 112-123.

- Hansen M.H., Hurwitz W.N., Bershad M. (1961). Measurement Errors in Census and Surveys. *Bulletin of the International Statistical Institute* 38(2), pp.359-374.
- Hosmer D. W., Lemeshow S. (1989). *Applied Logistic Regression*. John Wiley & Sons, Inc.
- Johnson N. L., Kotz S. (1972). *Distributions in statistics: continuous multivariate distributions*. Wiley, New York
- Kovar, J.G., MacMillan, J., Whitridge, P. (1988). *Overview and Strategy for the Generalized Edit and Imputation System*. Statistics Canada, Methodology Branch Working Paper No. BSMD-88-007E/F, Ottawa.
- Lee H., Rancourt E., Särndal C.-E. (2002). Variance Estimation from Survey Data under Single Value Imputation. In *Survey Nonresponse*, Groves R.M., Dillman D.A., Eltinge J.L., Little R.J.A. (eds), New-York:John Wiley&Sons, Inc., pp. 315-328.
- Lessler J.T., Kalsbeek W.D. (1995). *Non-sampling Errors in Surveys*, John Wiley.
- Little R.J.A., Rubin, D.B. (2003). *Statistical Analysis with Missing Data*. Wiley & Sons, New York.
- Luzi O., Della Rocca G. (1998). A Generalised Error Simulation System to Test the Performance of Editing Procedures, *Proceedings of the SEUGI 16*, Prague, June 9-12.
- Manzari A., Della Rocca G. (2000). A generalised system based on a simulation approach to test the quality of editing and imputation procedures, *ISTAT Essays* n. 6/2000, 83-103.
- Manzari A., Reale A. (2002). Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology, *Volume of the International Association of Survey Statisticians, Invited Papers, The 53rd Session of The International Statistical Institute, August 2001, Seoul, South Korea*, pp.634-655.
- Nordbotten S. (2000). Evaluating Efficiency of Statistical data Editing: General Framework. *United Nations Statistical Commission and Economic Commission for Europe, United Nations*, Geneva.
- Poulsen M.E. (1997). Evaluating Data Editing Process Using Survey Data and Register Data. *Statistical Data Editing Methods and Techniques Vol. II*, Conference of European Statisticians, United Nations, 1997.
- Rancourt E. (2002). Using Variance components to measure and evaluate the quality of editing practices, *Conference of European Statisticians, UN/ECE Work Session on Statistical Data Editing*, Helsinki, May 27-29.
- Rubinstein R. Y. (1981). *Simulation and the Monte Carlo method*. New York: Wiley.
- Schafer J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.
- Stefanowicz B. (1997). Selected issues of data editing, In *Statistical Data Editing, Methods and Techniques, Vol. 2. Conference of European Statisticians*, Statistical Standard and Studies No 48, UN/ECE, pp. 109-111.
- Tang G., Little R.J.A., Raghunathan T.E. (2003). Analysis of multivariate missing data with nonignorable nonresponse, *Biometrika*, 90, 4, pp. 747-764.
