

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Statistical Data Editing
(Rome, Italy, 2-4 June 1999)

Topic (ii): Generalized software packages for statistical data editing, their evaluation

**A GENERALISED SYSTEM BASED ON A SIMULATION APPROACH TO TEST
THE QUALITY OF EDITING AND IMPUTATION PROCEDURES**

Submitted by the Italian National Statistical Institute¹

Invited paper

I. INTRODUCTION

1. The purpose of an editing and imputation (E&I) procedure is to detect and correct non sampling errors in the survey data in order to improve the quality of the statistical information. We define an error as any deviation from the true value in a given variable. The check is performed by using a set of rules (edits) identifying missing, invalid and suspicious values. The adjustment consists of the replacement of each considered "erroneous" value with another one considered as "true".

2. The quality of an E&I procedure can be evaluated according two points of view (Granquist, 1997; Stefanowicz, 1997):

- a) in terms of the distance between the estimates computed on the *true values* and the estimates computed on the *edited values*;
- b) in terms of the number of detected, undetected and introduced errors.

In the first approach, the study is focused on the effect of editing on reported data (*output oriented approach*), while in the latter the study is focused on the effect of editing on individual data items (*input oriented approach*). In both cases, we need three different data sets: the set of *true* data (without errors), the set of *raw* data (with errors) and, finally, the set of *clean* data, result of the application of the E&I procedure to raw data.

3. The problem is how to obtain these data. An ideal solution would be to carry out new interviews of (a subset of) respondents, with much greater care (professional interviewers, computer assisted interviewing, reconciliation of current answers with previous ones, etc.) to obtain true data corresponding to existing raw and clean data. But, obviously, this is a rather expensive (from costs, time, and organisational points of view) procedure, and therefore it can be seldom adopted. An alternative, and quite different solution, might be to define as "true" a given set of data (obtained in an artificial way (Nordbotten 1995), or as the result of the application of the current E&I system), and to insert in them errors, by simulating as faithfully as possible the way errors are generated in the real survey operations (Barcaroli, D'Aurizio, 1997). Finally, clean data are obtained, as usual, by simply applying the E&I system.

¹ Prepared by Antonia Manzari and Giorgio Della Rocca.

4. A generalised software based on a simulation approach has been developed to evaluate the quality of E&I procedures. The software is based on the artificial introduction of errors in a set of true data by a controlled generation process (Luzi, Della Rocca, 1998) and provides indicators to measure the quality of an E&I process in terms of its capability to detect as many errors as possible, and restore the true values, without introducing new errors in these operations. We refer to the automatic E&I procedure. When procedures are interactive, approaches like macroediting or selective editing can be adopted, the target is no longer to find and correct as many errors as possible, but only relevant errors, i.e. errors with significant impact on final estimates.

5. The proposed indicators are described in Section II. The software is described in Section III. Section IV concerns an application of the software in testing an E&I procedure using real data from the 1991 Italian Population Census. Finally, brief conclusions and future work are reported.

II. QUALITY INDICATORS OF THE E&I PROCESS

6. In this paper we evaluate the *quality of an E&I process* in terms of its *capability to recognise errors and adequately replace them with the true values*: the higher the proportion of corrected errors on the total, and the less the quantity of new errors introduced, the more accurate is the process. In other words, our evaluation is not estimates oriented: the accuracy of an E&I process is not measured by comparing the statistical estimates obtained from clean data with the corresponding values resulting from the true data. Moreover, in our paper, no evaluation is made with regard to efficiency factors, for example time and computer resources, even if their relevance is often crucial.

Quality of editing process

7. Our basic choice is to consider an editing process like an instrument to discriminate each value as erroneous or as true. To be sure about the correctness of every single choice of a given editing process, we can plant artificial errors in a set of *original data* by modifying some original values, process the new set of *raw data* against the editing procedure and compare the editing results with the real values. The process is accurate if it classifies as erroneous (i.e. to be imputed) a modified value and classifies as true (not to be imputed) a true value. These concepts suggest us to evaluate an editing process by scoring its performance separately with respect to the accuracy for modified data, and the accuracy for true data. For each class of values, modified and true, we can consider the probabilities that the editing process classifies it correctly or not.

8. Let us consider whichever raw value of a generic variable. We can regard the editing decision as the result of a screening procedure designed to detect deviations of the raw values from the true values. Accordingly to the basic concepts of diagnostic tests (Armitage and Berry, 1971), considering all the set of values assumed by the variable in the data set, we can define:

- a : the proportion of *false positives*, i.e. not modified (true) values considered as erroneous;
- b : the proportion of *false negatives*, i.e. modified values not recognised as erroneous.

The analogy of editing decisions with diagnostic tests allows us to resort to familiar concepts: the probability to recognise unmodified values as true is analogous to the *specificity* (1-a), while the probability to recognise modified values as erroneous is analogous to the *sensitivity* (1-b). The extent of (1-a) and (1-b) indicates the reliability of the editing process in assessing the correctness of true and modified values.

9. We can estimate these probabilities by applying the editing process to a set of known modified and true data. Suppose that, for a generic variable, the application gives the following frequencies:

		Editing classification	
		erroneous	true
Original data	modified	a	b
	not modified	c	d

Where:

a = number of modified data classified by the editing process as erroneous,
 b = number of modified data classified by the editing process as true,
 c = number of true (not modified) data classified by the editing process as erroneous,
 d = number of true (not modified) data classified by the editing process as true.

10. Cases placed on the secondary diagonal represent the failures of the editing process. The ratio $c/(c+d)$ measures the failures of the editing process for true data and can be used to estimate a . The ratio $b/(a+b)$ measures the failure of the editing process for modified data and estimates b .

Ratios

$$I1 = d/(c+d)$$

and

$$I2 = a/(a+b)$$

estimate $(1-a)$ and $(1-b)$ respectively.

11. With regard to any single variable, for the whole set of data (modified and true) the *accuracy* of the editing process is measured by the fraction of total cases that are correctly classified:

$$I3 = (a+d) / (a+b+c+d).$$

The reader can easily verify that this is a linear combination of $I1$ and $I2$, whose weights are the fraction of the total cases that are true and the fraction of total cases that are modified:

$$I3 = I1 \frac{(c+d)}{(a+b+c+d)} + I2 \frac{(a+b)}{(a+b+c+d)}.$$

Therefore, its value can be strongly affected by the error level in data.

Quality of imputation process

12. We impose that the imputation process imputes only values previously classified by the editing process as erroneous. 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 can say that the imputation process fails.

13. Actually, in the case of quantitative variables, it is not necessary that the imputed value 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. In the case of qualitative variables, if the value classified as erroneous is a true one, the imputation process always fails.

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

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

where:

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

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

$$I5 = a_s/a.$$

The fraction of imputed true values for which the imputation is successful, i.e.

$$I4 = c_s/c$$

is not to be considered to evaluate the inner quality of the imputation process, because imputing equals to change the raw value and c_s/c is only an artificial result due to the definition of successful imputation for numeric variables. The same observation is valid for fraction of imputed total values, i.e.

$$I6 = (a_s + c_s)/(a + c)$$

because it under-estimates the inner quality of the imputation process.

The overall quality of edit and imputation process

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

$$I7 = I1 + [(1-I1) I4] = (c_s + d) / (c + d).$$

The accuracy of the E&I process for modified data is measured by estimating the probability to correct modified data. It is a joint probability of a combination of two results: to classify as erroneous a modified value *and* to restore the true value. It can be estimated by:

$$I8 = I2 I5 = a_s / (a + b).$$

For the whole set of data (modified and true) the accuracy of the E&I process is measured by the fraction of total cases whose true value is correctly restored: $I9 = (a_s + c_s + d) / (a + b + c + d)$. Even in this case, it is easy to verify:

$$I9 = I7 \frac{(c + d)}{(a + b + c + d)} + I8 \frac{(a + b)}{(a + b + c + d)}.$$

The indices defined above are summarised in Table 1.

Table 1. Accuracy indices

Class of data	Index	Calculation
	<i>Editing accuracy</i>	
true data	I1: fraction of true data that are correctly classified	$d/(c+d)$
modified data	I2: fraction of modified data that are correctly classified	$a/(a+b)$
total data	I3: fraction of total data that are correctly classified	$(a+d)/(a+b+c+d)$
	<i>Imputation accuracy</i>	
true data	I4: fraction of imputed true data whose true value is correctly restored	c_s/c
modified data	I5: fraction of imputed modified data whose true value is correctly restored	a_s/a
total data	I6: fraction of imputed total data whose true value is correctly restored	$(a_s+c_s)/(a+c)$
	<i>E&I accuracy</i>	
true data	I7: fraction of true data whose true value is correctly restored	$(c_s+d)/(c+d)$
modified data	I8: fraction of modified data whose true value is correctly restored	$a_s/(a+b)$
total data	I9: fraction of total data whose true value is correctly restored	$(a_s+c_s+d)/(a+b+c+d)$

17. Each index defined in Table 1 ranges from 0 (no accuracy) to 1 (maximum accuracy). The performance of an editing process is determined by a number of factors including the amount and kind of errors present in the data, the set of control rules (edits), and the characteristics of the error localisation algorithm.

18. Missing and invalid values are always detected by properly defined edits. It is a different matter when a really erroneous value is valid with respect to the domain of definition of the variable: its individuation depends strictly on the set of defined edits, which, in time, depend on the possible logical inconsistencies between subsets of domains of qualitative variables, or on the characteristics of multivariate distributions of quantitative variables. The more edits you can define, the higher the probability of a correct localisation of errors.

19. Even in case all possible inconsistency rules have been defined, the editing process will not detect an erroneous but valid value:

- i. if no inconsistency is verified between that value and the subsets of domains of the other variables;
- ii. in case inconsistencies are verified, if the error localisation algorithm fails. In this case there would be a double failure: with respect to the modified value and with respect to at least one true value of other different variables linked by the edits.

In case (ii), if the failure of the error localisation algorithm were systematic, we would observe low values of indicators I2 (and I8) for the current variable, together with low values of I1 (and I7) for at least one linked variable. In case (i) we would only observe low values of I2 (and I8) for the modified variable.

III. EDITING SYSTEM STANDARD EVALUATION (ESSE)

20. ESSE is a generalised software, based on a simulation approach, developed by ISTAT to evaluate the quality of E&I procedures. A prototype version of the software has been entirely developed in the SAS programming language and is currently running on WINDOWS and UNIX platforms. The software is generalised so that:

- it can be applied to any statistical survey without modification;
- it can handle both qualitative and numeric variables;
- it implements the main stochastic or systematic error generation mechanisms related to the most frequent types of error (measurement errors, misplacement errors, keying errors, etc.) arising during any statistical survey process.

21. Two main modules compose ESSE software:

(i) *Error simulation*: introduces artificial controlled errors in *true data* providing *raw data* to be processed against an E&I system;

(ii) *Evaluation*: provides reports containing indices to assess the quality of the E&I process.

In the *Error simulation* module (Luzi and Della Rocca, 1998), the generation of errors is performed record by record using a set of predefined error models. Generally, error models can be applied to given variables, with the exception of the *keying error* model that is to be applied to each single byte in the record. To perform a given perturbation strategy, the user has to indicate the expected error incidence, i.e. the probability with which values are modified on the basis of the selected error model.

22. Once the set of error generation models (with the associated error incidence) has been defined for each variable, the sequence of error models to apply is randomly determined. Starting from the first model in the selected application sequence, the software verifies its applicability: an error model is applied if the current value has not been previously modified by another model, and if the random values generated from a uniform distribution is less than the defined perturbation probability.

23. The output of the *Error simulation* process is a raw data file containing both original and modified values. For each value the software produces also an indicator variable whose code is equal to 0, if the value is unmodified, or equal to the code of the applied error model, if the value has been modified. Moreover, for each selected variable, the software provides the perturbation frequencies for each model. This information allows the user to control the perturbation process for each variable and model. The *Evaluation* module produces the indicators defined in Table 1.

IV. APPLICATION

IV.1. Description

24. A demonstrative application was performed to show the features of the ESSE software. We tested the quality of an E&I system developed using the prototype implementing the New Imputation Methodology (Bankier et al., 1994), kindly supplied by Statistics Canada, in removing a specific kind of error. The simulation study has been based on real data from the 1991 Italian Population Census, within the context of research actions addressed to improve the efficacy of the E&I process for the next 2001 Italian Population Census.

25. The NIM was used in the 1996 Canadian Census to carry out edit and imputation for demographic variables. NIM allows, once given the available donors, minimum change imputation of numeric and qualitative variables simultaneously. NIM identifies as potential donors those passed edit records which are as similar as possible to the failed edit record (a distance function is defined). Then for each nearest donor,

the smallest subsets of the non-matching variables, which, if imputed, allow the record to pass the edits, are identified. One of these possible imputation actions is randomly selected. Additional details are given in Bankier et al (1996).

26. The test has been carried out on 1000 (this number is due to a prototype limitation) four-person households, considered free of errors with respect to a defined set of “between persons” and “within person” control rules (summarised in the Annex). Demographic variables selected for the test were: *Relation to Person 1, Sex, Marital Status, Year of birth, and Year of marriage*. In order to define the edits we had to transform the last two variables in *Age in years* and *Duration of marriage in years*. The original variables were again calculated, from the clean data, at the end of the imputation process.

27. The distance between a failed record \mathbf{V}_f and a passed record \mathbf{V}_p is defined as follows:

$$D(\mathbf{V}_f, \mathbf{V}_p) = \sum D_i(\mathbf{V}_{fi}, \mathbf{V}_{pi})$$

where $D_i(\mathbf{V}_{fi}, \mathbf{V}_{pi})$ is the distance function for variable i . For qualitative variables (*Relation to Person 1, Sex, Marital Status*) $D_i(\mathbf{V}_{fi}, \mathbf{V}_{pi})$ is defined as 0 when $\mathbf{V}_{fi} = \mathbf{V}_{pi}$, and 1 when they are different. For quantitative variables (*Age in years, Duration of marriage in years*) $D_i(\mathbf{V}_{fi}, \mathbf{V}_{pi})$ is defined as 0 when $\mathbf{V}_{fi} = \mathbf{V}_{pi}$, is defined as 1

- if $|\mathbf{V}_{fi} - \mathbf{V}_{pi}| \geq 10$
- if only one between \mathbf{V}_{fi} and \mathbf{V}_{pi} has a blank response (for *Duration of marriage* a blank response is in the domain).

If \mathbf{V}_{fi} and \mathbf{V}_{pi} have both a non blank response and is $|\mathbf{V}_{fi} - \mathbf{V}_{pi}| < 10$, then $D_i(\mathbf{V}_{fi}, \mathbf{V}_{pi}) = |\mathbf{V}_{fi} - \mathbf{V}_{pi}| / 10$.

An imputation action is selected among those that minimise the following

$$D(\mathbf{V}_f, \mathbf{V}_p, \mathbf{V}_a) = 0.9 D(\mathbf{V}_f, \mathbf{V}_a) + 0.1 D(\mathbf{V}_a, \mathbf{V}_p)$$

28. Artificial random errors were planted in the original data by ESSE software using the *interchange error* model: the original value was replaced by a wrong one randomly chosen in the admissible range. In this way we only simulated a particular type of error occurring in the phase of the compilation of the questionnaire. We did not use other error models occurring both in the phase of the compilation of the questionnaire and in the data entry phase, for example the *item non response* model or the *keying error* model, to not introduce missing or invalid values which would have increased the values of the editing accuracy indices without giving information on the performance of the error localisation algorithm.

29. The perturbation probability (percentage of errors) per variable was systematically varied in order to keep track of how the performance of the E&I system changes as the error incidence in raw data varies. The selected perturbation probabilities per variable were set to: 1%, 5%, 10% and 15%. As the software simulates only univariate errors, i.e. it works under the hypothesis that errors are independent, the expected value of the maximum frequency of modified records (persons) in raw data had to be approximately the sum of the percentage of errors for the five variables, i.e.: 5%, 25%, 50% and 75%.

30. The imputation process was considered as successful:

- for numeric variables (*Year of birth* and *Year of marriage*): if the imputed value lay in an interval of ± 3 years around the original value;
- for qualitative variables (*Relation to Person 1, Sex, Marital Status*): if the imputed value was equal to the original value.

To reduce effects of the perturbed data, which are typical for that data set only, the simulation, editing and evaluation process was replicated three times, just to obtain more reliable results. Each time new errors (*raw*

data) were generated and processed against the E&I system. For each variable and for each level of perturbation probabilities, the average value of the accuracy indices were computed.

IV.2 Results

31. Before analysing the results, we have to consider that:

- the approach used by the NIM is data driven and therefore a large suitable set of donors is needed to obtain a good performance of the E&I system (while our test used only 1000 households);
- the imputation methodology implemented in the prototype program is somewhat different to what implemented in the NIM production system (it is a simplified version).

So it is reasonable to think that using a production version of the NIM, we would achieve even better results.

32. In Table 2, for each variable and each perturbation probability, the average values of the accuracy indices are given. The effect of the percentage of errors planted in data is clear: the value of each index tends to decrease as the error level increases.

Editing accuracy

33. High values of the fraction of true data that are correctly classified, i.e. I1 index, indicate that the E&I system does not introduce new errors in data. This appealing feature depends on the restrictions imposed by the set of the defined edits and on the characteristics of the error localisation algorithm. If the set of inconsistencies were too restrictive, then the E&I system could classify as erroneous some plausible but uncommon true values. If the localisation algorithm fails, the E&I system classifies as erroneous some true values. In our application we can observe an equally excellent performance for all the variables except *Year of marriage*: its lower performance results from some imputations of true values due to the edit that forces spouses to have the same year of marriage. Considering the whole situation we can state that the error localisation algorithm implemented in the E&I system works correctly.

34. High values of the fraction of modified data that are correctly classified, i.e. I2 index, indicate that the E&I system is able to detect a large proportion of errors in data. As we did not simulate missing or invalid values, also in this case its performance depends on the set of defined edits and the characteristics of the error localisation algorithm. The high values of I1 for all the variables (good performance of the error localisation algorithm) suggest that low values of I2 for *Year of birth*, *Relation to Person 1* and *Sex* strictly depend on the set of defined inconsistencies for these variables. The power of the set of edits to detect errors in these variables results to be too low, so new inconsistencies should be defined (the more close are the inconsistencies between subset of domains the higher is the probability to detect erroneous values).

35. The fraction of total data that are correctly classified, i.e. I3 index, is determined by the values of I1 and I2, but also by the percentage of errors present in data. It is easy to note, in our application, that as the I2 value is always lower than I1 value, the reduction of I3 value (respect to I1 value) is stronger as the error level grows.

Imputation accuracy

36. The performance of an imputation process is determined by the characteristics of the imputation algorithm and by the set of defined edits but also by the characteristics of the multivariate distribution of data in the original set. The last one is an external factor whose evaluation is beyond the purpose of this paper. For any given imputation algorithm, we will notice high value of the fraction of imputed modified data whose true value is correctly restored, i.e. I5 index, if the range of admissible values to impute is highly restricted by both the characteristics of the domain of the variable and the constraints imposed by the edits in

which the variable is involved. This is true also in the case of a set of edits with low power to detect the errors (low values of I2 index), because I5 measures only the accuracy for the subset of modified values correctly classified by the editing process as erroneous. In our application high values of I5 for *Sex* result from the obliged imputation of the true value, due to the edit that forces spouses to be opposite in *Sex* (there are only two admissible values: male and female). On the contrary, low values of I5 for *Year of birth* are due to a lack of restrictions on the admissible value to impute to pass the edits given the erroneous value has been localised (for example when a household fails the edit which states 14 years as minimum age difference between a parent and a child). We can say that, for both variables, the defined inconsistencies are not sufficient to detect all the errors, but when an error is recognised, in *Sex* the defined inconsistencies force to impute the true value, whilst in *Year of birth* the inconsistencies can be passed by imputing values even very different from the true one.

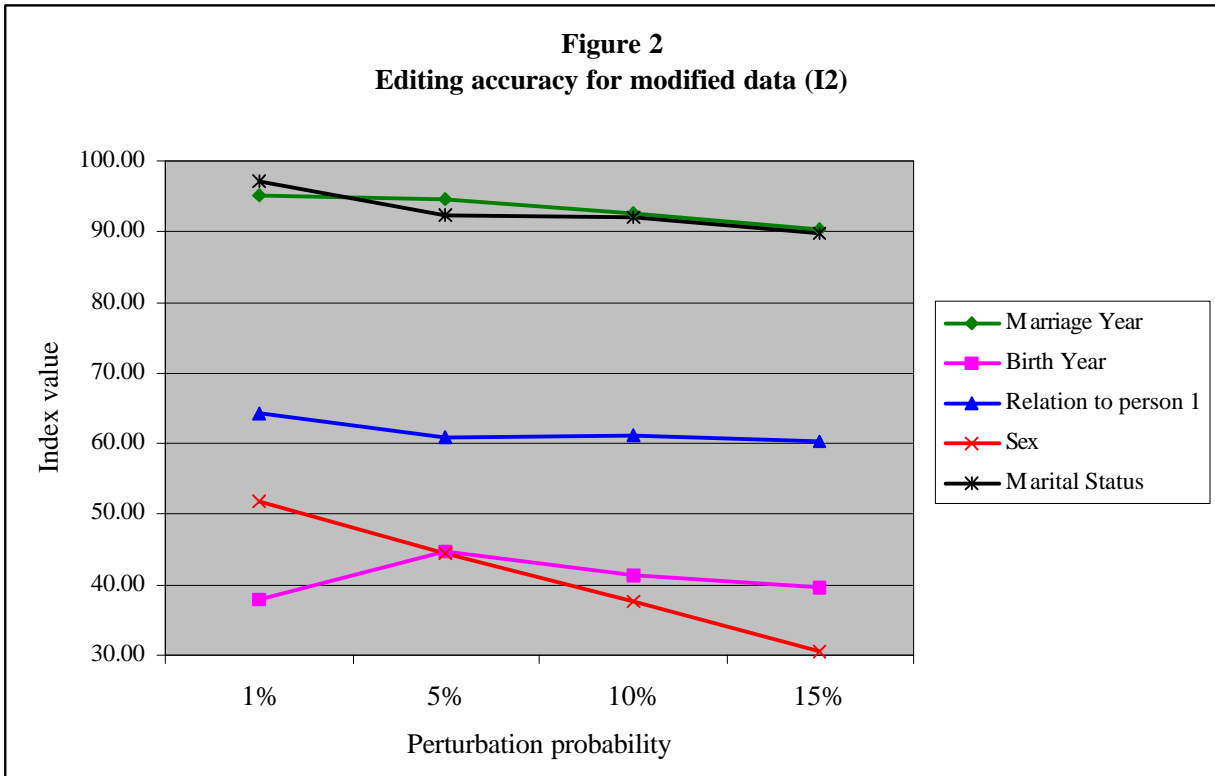
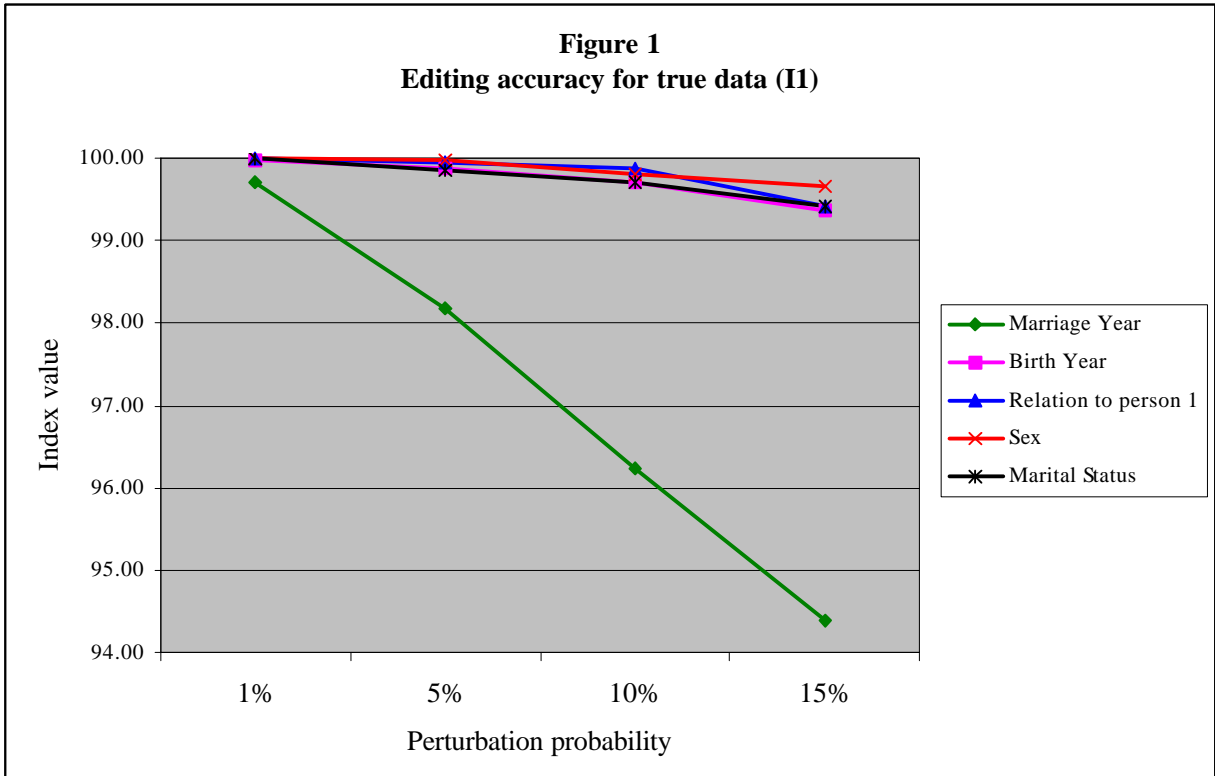
37. Concerning the fraction of imputed true data whose true value is correctly restored, i.e. I4 index, we note that its consideration is useful because it counterbalances the failure of the editing process for true data (only for numeric variables) in the overall editing and imputation evaluation. In our application low values for *Year of birth* are due to the already mentioned lack of restrictions on the admissible values, while higher values for *Year of marriage* result from a richer set of restrictions on the admissible values.

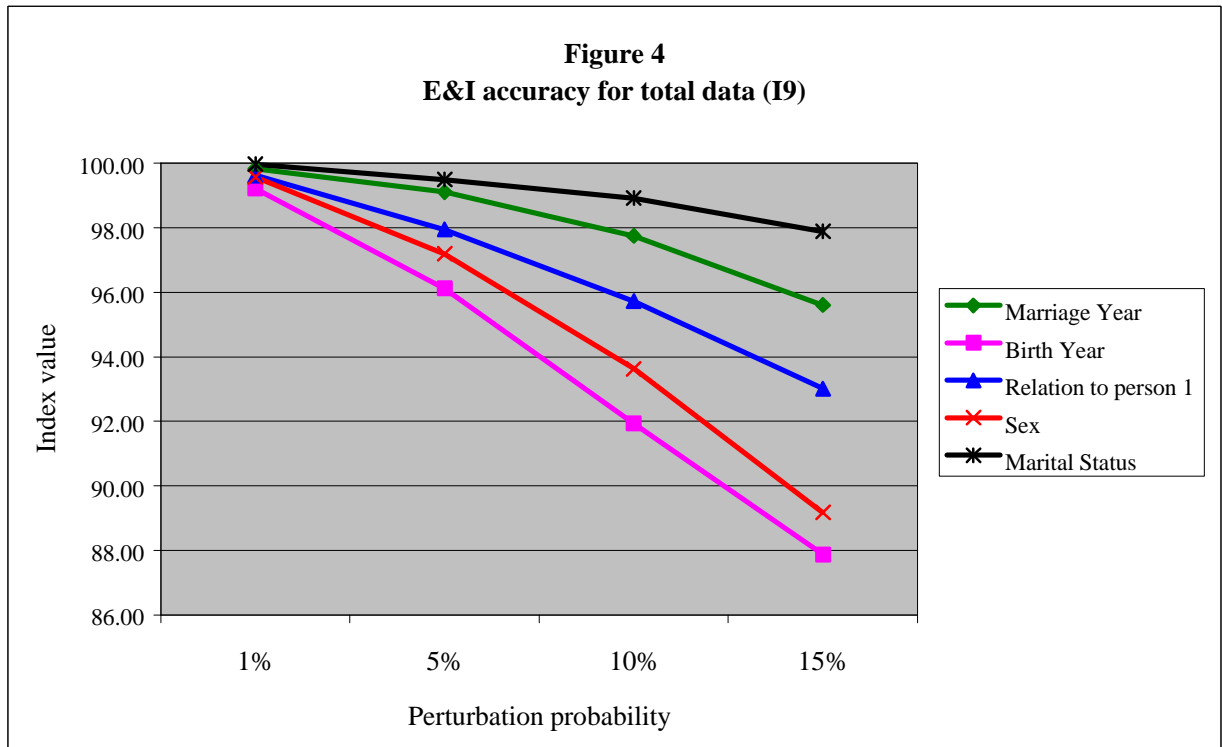
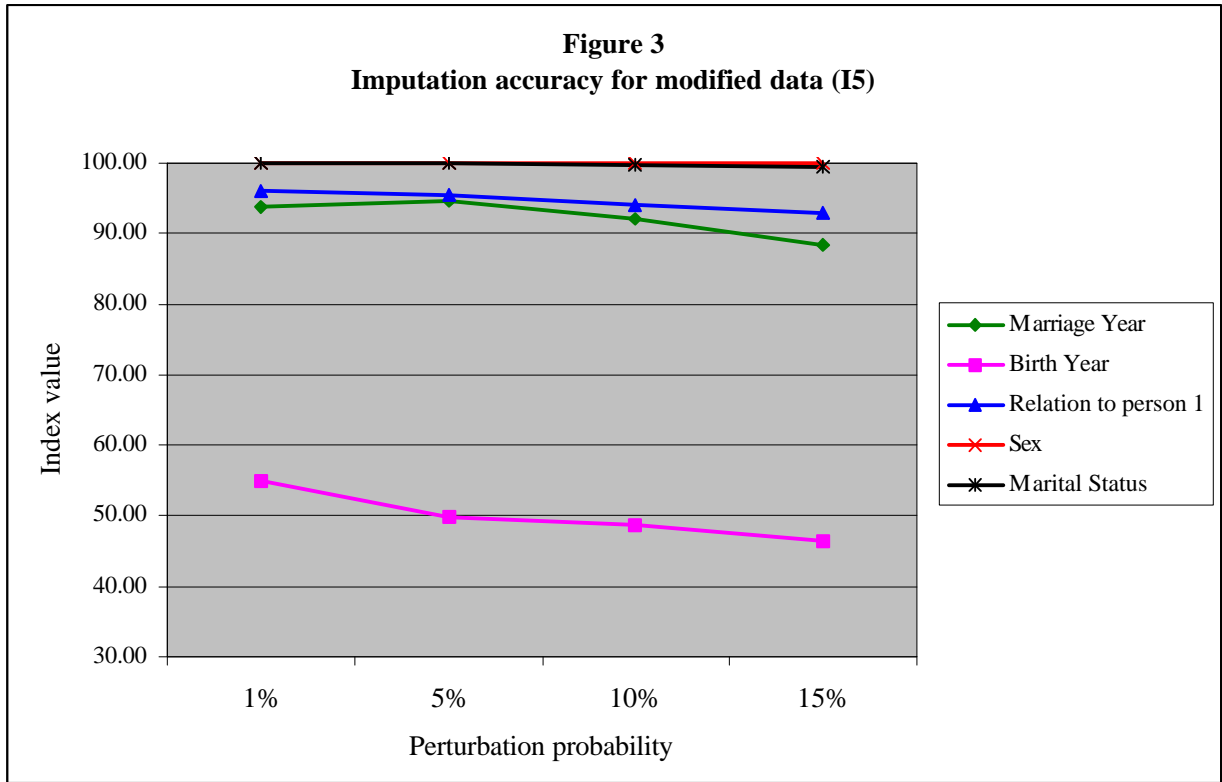
Overall editing and imputation accuracy

38. The fraction of true data whose true value is correctly restored, i.e. I7 index, and the fraction of modified data whose true value is correctly restored, i.e. I8 index, summarise, respectively for true and modified data, the quality of the overall editing and imputation process. The fraction of total data whose true value is correctly restored, i.e. I9 index, measures the overall accuracy for the whole set of data. The same consideration made for I3 values are valid for I9 values.

Table 2. Average value of the accuracy indices

Variable	Perturbation probability	I1	I2	I3	I4	I5	I6	I7	I8	I9
<i>Year of marriage</i>	1%	99.72	95.06	99.67	77.50	93.88	90.06	99.93	89.28	99.83
	5%	98.17	94.62	98.00	78.96	94.64	90.53	99.62	89.56	99.11
	10%	96.24	92.68	95.88	76.90	92.02	87.98	99.13	85.29	97.75
	15%	94.38	90.48	93.81	70.38	88.37	83.66	98.33	79.95	95.60
<i>Year of birth</i>	1%	99.98	38.04	99.39	0.00	54.82	52.74	99.98	20.95	99.22
	5%	99.87	44.64	97.19	18.10	49.94	48.40	99.89	22.25	96.12
	10%	99.70	41.34	93.99	13.49	48.73	46.72	99.75	20.14	91.94
	15%	99.36	39.53	90.83	12.12	46.34	43.27	99.44	18.30	87.88
<i>Relation to Person I</i>	1%	99.99	64.27	99.64	0.00	95.99	94.70	99.99	61.51	99.62
	5%	99.94	60.79	98.08	0.00	95.40	93.46	99.94	58.03	97.94
	10%	99.89	61.12	96.07	0.00	94.19	92.65	99.89	57.55	95.72
	15%	99.42	60.32	93.63	0.00	92.98	88.18	99.42	56.10	93.01
<i>Sex</i>	1%	99.99	51.89	99.58	0.00	100.00	98.67	99.99	51.89	99.58
	5%	99.97	44.55	97.19	0.00	100.00	98.89	99.97	44.55	97.19
	10%	99.82	37.79	93.63	0.00	100.00	95.76	99.82	37.79	93.63
	15%	99.67	30.61	89.18	0.00	100.00	94.23	99.67	30.61	89.18
<i>Marital Status</i>	1%	99.99	97.28	99.97	0.00	100.00	99.12	99.99	97.28	99.97
	5%	99.86	92.41	99.49	0.00	100.00	97.00	99.86	92.41	99.49
	10%	99.70	92.16	98.95	0.00	99.63	96.71	99.70	91.82	98.92
	15%	99.42	89.82	97.97	0.00	99.38	95.91	99.42	89.26	97.89





V. CONCLUDING REMARKS AND DEVELOPMENTS

39. The simulation approach to evaluate the performance of E&I systems can be useful to carry out the following analyses :

- (i) *Evaluation of the quality of an E&I system at different error levels*: information coming from accuracy indices is particularly useful in phase of definition and tuning of the set of control rules and gives the user useful insight into the limits of a given set of rules in detecting and correcting the errors; this work allows the user to optimise the definition of the set of control rules;
- (ii) *Comparison of the quality of two or more E&I systems at different error levels*: once a set of control rules has been selected, the comparison among the performance of different E&I systems supports the user in the choice of the best one (with respect to a selected criterion).

ESSE software is a standard tool to perform both the analyses.

40. Regarding future developments, the authors are planning to implement additional feature in the software in order to:

- (i) permit the automated iteration of the whole test process (*Error simulation/E&I/Evaluation*) for whatever number of times, in order to estimate mean values and variances of the defined indicators;
- (ii) carry out the evaluation of the quality of an E&I procedure also in terms of comparison between statistics computed from the *true values* and those computed by the *edited values* (estimates oriented evaluation).

ANNEX

Between persons and within person control rules defined in the application.

In the following: *relation* stands for *Relation to person1*; *marsta* for *Marital Status*; *age* for *Age in years*; and *duration* for *Duration of marriage in years*.

The indices *i,j,k* point out the position of the person and could take on values 2 to 4.

Between persons control rules

relation(i) = spouse and *relation(j)* = spouse

relation(i) = common-law spouse and *relation(j)* = common-law spouse

relation(i) = common-law spouse and *relation(j)* = spouse

relation(i) = spouse and *sex(1)* = *sex(i)*

relation(i) = common-law spouse and *sex(1)* = *sex(i)*

relation(i) = spouse and *duration(1)* ≠ *duration(i)*

relation(i) = spouse and *marsta(1)* ≠ *married*

relation(i) = spouse and *marsta(i)* ≠ *married*

relation(i) = parent and *relation(j)* = parent and *relation(k)* = parent

relation(i) = father/mother-in-law and *relation(j)* = father/mother-in-law and *relation(k)* = father/mother-in-law

relation(i) = parent and *relation(j)* = parent and *marsta(i)* ≠ *married*

relation(i) = parent and *relation(j)* = parent and *marsta(j)* ≠ *married*

relation(i) = parent and *relation(j)* = parent and *duration(i)* ≠ *duration(j)*

relation(i) = parent and *relation(j)* = parent and *sex(i)* = *sex(j)*

relation(i) = father/mother-in-law and *relation(j)* = father/mother-in-law and *marsta(i)* ≠ *married*

relation(i) = father/mother-in-law and *relation(j)* = father/mother-in-law and *marsta(j)* ≠ *married*

relation(i) = father/mother-in-law and *relation(j)* = father/mother-in-law and *duration(i)* ≠ *duration(j)*

relation(i) = father/mother-in-law and *relation(j)* = father/mother-in-law and *sex(i)* = *sex(j)*

relation(i) = parent and *relation(j)* = parent and *age(i)-age(1)* <14 and *age(j)-age(1)* <14

relation(i) = spouse and *relation(j)* = child and *age(1)-age(j)* <14 and *age(i)-age(j)* <14

relation(i) = spouse and *relation(j)* = father/mother-in-law and *relation(k)* = father/mother-in-law and *age(j)-age(i)* <14 and *age(k)-age(i)* <14

relation(i) = brother/sister and *age(1)-age(i)* ≥30

relation(i) = brother/sister and *age(i)-age(1)* ≥30

relation(i) = grandchild and *age(1)-age(i)* <30

relation(i) = father/mother-in-law and *marsta(1)* = single

relation(i) = common-law spouse and *relation(j)* = father/mother-in-law

Within person control rules

marsta(i) ≠ single and *age(i)* ≤13

marsta(i) ≠ single and *duration(i)* = blank

marsta(i) = single and *duration(i)* ≠ blank

marsta(i) ≠ single and *age(i)-duration(i)* <14

relation(i) = spouse or son/daughter-in-law and *duration(i)* = blank

relation(i) = common-law spouse and *age(i)* ≤13

relation(i) = spouse or son/daughter-in-law and *age(i)* ≤13

relation(i) = parent or father/mother-in-law and *age(i)* ≤35

marsta(i) = divorced and *age(i)* ≤17

marsta(1) ≠ single and *duration(1)* = blank

$marsta(1) = \text{single}$ and $duration(1) \neq \text{blank}$
 $marsta(1) \neq \text{single}$ and $age(1) - duration(1) < 14$
 $marsta(1) = \text{divorced}$ and $age(1) \leq 17$
 $relation(i) = \text{common-law spouse}$ and $marsta(i) = \text{married}$
 $relation(i) = \text{son/daughter-in-law}$ and $marsta(i) = \text{single}$
 $relation(i) = \text{son/daughter-in-law}$ and $marsta(i) = \text{divorced}$
 $relation(i) = \text{spouse}$ and $marsta(i) \neq \text{married}$

REFERENCES

Armitage P., Berry G. (1971) *Statistical methods in medical research*. Oxford, London, Edinburgh, Boston, Melbourne: Blackwell Scientific Publications.

Bankier M., Filion J.-M., Luc M. and Nadeau C. (1994) Imputing Numeric and Qualitative Variables Simultaneously, *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 242-247.

Bankier M., Luc M., Nadeau C. and Newcombe P. (1996) Imputing Numeric and Qualitative Variables Simultaneously, *Statistics Canada Technical Report*

Barcaroli G., D'Aurizio L. (1997) Evaluating editing procedures: the simulation approach, Working paper, Conference of European Statisticians, Work Session on Statistical Data Editing, Prague 1997

Granquist L. (1997) An overview of methods of evaluating data editing procedures, In *Statistical Data Editing, Methods and Techniques, Vol. 2*. Statistical Standard and Studies No 48, UN/ECE, 112-123.

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

Nordbotten S. (1995) Editing statistical records by neural networks, Working paper N. 40, Conference of European Statisticians, Work Session on Statistical Data Editing, Athens 1995

Stefanowicz B. (1997) Selected issues of data editing, In *Statistical Data Editing, Methods and Techniques, Vol. 2*. Statistical Standard and Studies No 48, UN/ECE, 109-111.