Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology

Antonia Manzari & Alessandra Reale
Istituto Nazionale di Statistica – Roma, Italy
manzari@istat.it, reale@istat.it

1. Introduction

Population Census (PC) data are, together with administrative Register data, the primary source of information for demographic studies about population structural features. The surveyed statistical unit has a hierarchical structure: PC data are collected at the household (unit) level with information for each person (sub-unit) within the household. PC data, like data from any survey, can contain errors and missing values. Therefore, editing and imputation (E&I) procedures have to be performed if a complete and consistent dataset is required. A crucial problem in imputing hierarchical data is preserving the relationships between variables belonging to different persons within the household (*between persons* edit rules) in addition to the usual problem of preserving relationships between variables belonging to a given person (*within person* edit rules).

In 1991 PC data, the Italian National Statistics Institute (ISTAT) decided to apply the Fellegi-Holt (1976) approach implemented in the software SCIA (Riccini et al., 1995) for imputing the non responses and resolving the inconsistent responses. The huge set of edit rules, needed to handle *between persons* and *within person* relationships, does not allow one to handle in a single step the demographic variables (the implicit edit-generation could not be accomplished because of computational limits). Consequently, the process was divided into two sequential steps: in the first step the variables *Year of birth, Sex, Marital Status* and *Year of marriage* were handled together with all the other individual variables by means of the probabilistic approach; whereas in the second step the *Relation to Person 1* variable was handled by means of a deterministic approach. The error localization solutions were not optimal because not all implicit edits could be obtained. Moreover the Fellegi-Holt approach does not allow one to define edits that are critical in order to correct *between persons* relationships, that is the comparison of two ages, because linear inequalities expressing relationships between numeric variables cannot be specified as edit rules (joint editing and correction of both qualitative and numeric variables are not allowed).

Preparing for the 2001 PC, the Italian National Statistics Institute (ISTAT) planned research studies with an aim to improve the efficacy of the E&I process. Concerning demographic variables, it has been decided to tackle the problem of data completeness and consistency by means of an approach more suitable to handle hierarchical data.

Since 1994, the ISTAT Multipurpose Survey on Households (MSH) adopts an *ad hoc* procedure designed and implemented to edit and correct the relationships among household persons. The correction process is based on the identification of the main couple in the household and it requires that *sex* and *age* variables are free of errors. Moreover, interactive actions required in order to correct edit-failing households that cannot be automatically corrected are time and resources consuming. For these reasons the MSH procedure was deemed not suitable to handle demographic variables from Census data and new research studies were undertaken. Among the different research studies we consider the one concerning the joint development of a new software, the Data Imputation and Edit System - Italian Software (DIESIS), by ISTAT and academic researchers (Dipartimento di Informatica e Sistemistica dell'Università degli Studi di Roma "La Sapienza"). The new software performance has been evaluated and compared with the performance of the Canadian Nearest-neighbour Imputation Methodology (NIM) (Bankier, 1999) by a simulation study based on real data from the 1991 Italian PC. NIM has been selected for the comparative evaluation because nowadays it is deemed to be the best methodology to automatically handle hierarchical demographic data. The NIM version used for the test is the one implemented in the CANadian Census Edit and Imputation System (CANCEIS) (Bankier, 2000) supplied by Statistics Canada.

This paper describes the main characteristics of the two systems (section 2), describes the evaluation study (section 3) and presents the results (section 4).

2. Main systems characteristics

In this section a brief comparison of the main characteristics of the two systems is given. For more details refer to Janes (2001) for CANCEIS and to Bruni et al. (2001) for DIESIS.

Both systems treat invalid or inconsistent responses for qualitative and numeric variables simultaneously.

For both systems, edit rules must be defined by conjunctions of logical propositions or linear inequalities that can be extended to non-linear inequalities. At the moment DIESIS can accept also inequalities containing the product of two variables.

Both systems locate *redundancies* between edit rules (some edit rules are included in other edit rules). DIESIS checks also for the presence of complete and partial *inconsistencies* between edit rules. Complete inconsistency means that whatever combination of values causes edit failure, while partial inconsistency means that some rules imply that there are admissible values of at least a single variable which would automatically cause edit failure, irrespective of the values in the other variables. Moreover, if some deterministic imputation rules are used before or after the DIESIS E&I process, it is possible to locate the redundancies and inconsistencies between the whole set of rules (that is, between DIESIS edit rules and the deterministic imputation rules all together considered).

Both systems perform imputation on all failed units. They try to *impute the minimum number of variables given the available donors* (this approach will be called ***first donors then fields***). They achieve this by means of donor imputation from a single donor. The donor from which the responses are taken, is chosen among the nearest neighbours, that is, among donors that resemble the failed unit.

Both systems search for (and analyse) potential donors in stages and determine the number of stages by means of stop rules a priori defined. For each system, the main steps are described below.

As regards CANCEIS, it determines potential donors with the best imputation actions (imputation actions are changes to the failed household so that the new adjusted household may pass the edit rules). To do this CANCEIS computes two distance values: the first distance represents how close the failed household is to the potential donor household:

$$D_{fp} = \Sigma_i \, w_i \, D_i \, (V_{fi}, V_{pi})$$

D_{fp} represents the distance between a **f**ailed and a **p**assed household. The summation is over each variable for a household, so the w_i is the weight of the i^{th} variable, and the D_i is a distance score between the values of the i^{th} variable of the failed household (V_{fi}) and the i^{th} variable of the passed household (V_{pi}). For qualitative variables, CANCEIS uses 0/1 distance score while for numeric variables, it uses distance score that can take on values in the range [0,1]. Note that the D_{fp} distance is computed over all sub-units. Households with the lowest D_{fp} are retained on a *list of potential donors*. For each household in the list, CANCEIS generates feasible imputation actions. For each of them, CANCEIS computes a second distance function comparing the new adjusted household simultaneously to the failed household and the potential donor:

$$D_{fpa} = \alpha \, D_{fa} + (1\text{-}\alpha) \, D_{ap}$$

D_{fpa} is a weighted average of the distance between the adjusted household and the failed household $D_{fa} = \Sigma_i \, w_i \, D_i \, (V_{fi}, V_{ai})$ (this is a measure of the amount of imputation done), and the distance between the adjusted household and the potential donor household $D_{ap} = \Sigma_i \, w_i \, D_i \, (V_{ai}, V_{pi})$ (this is a measure of the "plausibility" of the imputation action). The imputation actions with the smallest D_{fpa} are retained on a list of *near minimum change imputation action*s. CANCEIS randomly selects the imputation action to use from those in the list by giving a better chance to those imputation actions with the smaller D_{fpa}.

As regards DIESIS, it first finds a subset of potential donor households with the smallest distance from the failed household. The distance function used can be, like CANCEIS, a weighted sum of the distance scores for each variable over all persons (this approach will be called ***all sub-units donor***) or over the subset of persons involved in failed edit rules (this approach will be called ***involved sub-units donor***). In both approaches all variables are considered (and not only those which enter failed edits). For example, consider six-person households and five variables for each person. In a given failed household, assume that only three persons are involved in failed edit rules. With the first option (*all sub-units donor*) the system searches for donors from the sets of six persons (i.e. from the six-person households) and the distance is a summation of 30 distance scores (5 by 6). With the second option (*involved sub-units donor*) the system searches for donors from the sets of three persons from the six-person households and the distance is a summation of 15 distance scores (5 by 3).

Then DIESIS selects the imputation action to use, by minimising the weighted number of changes. It does this by solving a problem of minimum change with the following constraints:

1) the adjusted record must pass the whole set of edit rules (all the defined edit rules and not only those that originally failed);
2) the imputed values must come from a single donor.

The aim is to select from the nearest neighbours the one that allows the adjusted household to preserve the largest number of values from the failed household (minimum change), respecting the original frequency distributions. In other words, DIESIS selects the imputation action to uses by minimising the following function:

$$\Sigma_k \Sigma_i \, c_i \, y_i$$

where y_i is a dummy variable which value is 0, if the value of the i^{th} variable of the k^{th} person in the failed household is equal to the value of the i^{th} variable of the k^{th} person in the adjusted household, or 1, if the value of the i^{th} variable of the k^{th} person in the failed household is not equal to the value of the i^{th} variable of the k^{th} person in the adjusted household. The c_i is the weight of the i^{th} variable and can assume whatever numeric value. The weight can be dynamic, that is, a different weight can be assigned to each value of the variable in the failed household. In other words, the weight can depend on the value taken by the variable in the failed household. The internal summation is over each variable of a person, the external summation can be over all the persons (this approach will be called ***all sub-units imputation action***), like CANCEIS does, or only over the subset of persons involved in failed edit rules (this approach will be called ***involved sub-units imputation action***). Continuing the previous example: with *all sub-units imputation action* the summation of c_iy_i is over six sub-units and has 30 elements (5 by 6); while with *involved sub-units imputation action* the summation of c_iy_i is over three sub-units and has 15 elements (5 by 3).

Moreover DIESIS performs also the *absolute minimum weighted change* (this approach will be called ***first fields then donors***) for qualitative and numeric variables simultaneously overcoming the computational limits, related to the implicit edit generation (Winkler, 1999), of the systems implementing the Fellegi-Holt approach (Fellegi and Holt, 1976). The solution is obtained defining the error localisation problem as combinatorial optimisation problems that can be solved by using branch-and-cut procedure (Bruni et al., 2001). At this moment the localisation module is already completed while the imputation module is still being developed. The two imputation algorithms ("*first donors then fields*" and "*first fields then donors*") will be able to be separately or jointly used. By jointly use, we mean that in an "*impute the minimum number of variables given the available donors"* strategy, the user will be able to choose the "*absolute minimum weighted change*" when, for a given failed household, the number of changes proposed by the first algorithm is exceedingly high, compared to the number of changes proposed by the second algorithm.

Note that, both CANCEIS and DIESIS, whatever is the algorithm, always select the imputation action to use among the set of *feasible* imputation actions, that is the set of imputation actions that will allow the adjusted unit to pass all the edit rules and not only the edit rules that originally failed.

Both systems run on the Windows platform in PC environment

3. The evaluation study

Our purpose is to compare the performance of the two E&I systems in terms of their accuracy. This means that we want to measure the closeness between the "true" values and the outcomes obtained by the systems.

We distinguish the editing process from the imputation process. We consider editing as the process of detecting erroneous values. Its purpose is to detect the maximum number of erroneous values (in order to impute them). Its outcome is the classification of each observed value as *correct* or *incorrect*. Imputation is the process by which *incorrect* values are replaced by new correct values. Its purpose is to restore the true value and its outcome is the new assigned value.

We carry out the evaluation of the performance of the two E&I systems by comparing three different sets of data: the *original*, *perturbed* and *corrected* data (Granquist, 1997).

Original data correspond to the data that would have been observed under a perfect data production process, free of missing data and inconsistencies according a defined set of edit rules. Original data used in our study were obtained by applying an E&I procedure to 1991 Italian PC data, that are therefore free of missing values and inconsistencies with respect to a defined set of *between persons* and *within person* edit rules (reported in the Appendix).

Perturbed data correspond to the result of the statistical production process. Our perturbed data were obtained by a simulation approach based on the controlled artificial introduction of erroneous values into original data.

Corrected data were edited and imputed values obtained by processing the *perturbed* data by means of the investigated E&I procedures.

Demographic variables used for the test are *Relation to Person 1*, *Sex*, *Marital Status*, *Year of birth*, and *Year of marriage*. The last two variables are transformed into the variables *Age (in years)* and *Years married* in order to define edits involving these variables. Potential couples, which have non-unique relationships to Person 1, are identified prior to imputation in order to apply some couple edits to them. A *Couple* variable is defined for each person. Pair of persons that could form a couple have the same value for the *Couple* variable.

CANCEIS system processes data by imputation groups having the same number of sub-units (persons in the household). We analyse the performance on two different household dimensions: four-person households (45,716 units from a single district) and six-person households (20,306 units from a single region).

We perturbed original data by replacing valid responses with *non-response* (the original value was randomly replaced by a missing value) or with *other valid responses* (the original value was replaced by a wrong one randomly chosen in the admissible domain). Note that the latter perturbation model may not cause households to fail the edit rules because the modified values may not enter any failing edit rules.

For each variable, the perturbation percentages were chosen from the observed 1991 frequencies of missing values and of values considered erroneous, in order to closely resemble real life situation (the invalid values were treated like missing values).

For the *other valid response* perturbation model, the adopted perturbation percentages (x) were obtained by adjusting the observed 1991 imputation frequencies (y) according to an estimate of the probabilities of the following editing errors: to classify as *incorrect* a true values (α) and to classify as *correct* an erroneous values (β). In order to get estimates for the probabilities α and β, we assume that 1991 editing procedure had the same probabilities of detecting errors than the CANCEIS editing procedure. So, we performed three runs of CANCEIS system (processing four-person households) having only the *other valid response* perturbation model setting at different perturbation percentages (1%, 5% and 10%). Then, we used the average of the three frequencies of not modified data erroneously imputed as estimate of α and the average of the three frequencies of modified data not imputed as estimate of β. Their values are reported in Table 1:

Table 1. Estimates of the probabilities α and β (percentage values)

Estimate of	Variable				
	Relation	*Sex*	*Marital Status*	*Age*	*Years married*
α	0.37	0.06	0.01	0.49	1.34
β	44.92	60.77	10.82	43.24	10.81

Finally, the adopted perturbation percentages (x) for the *other valid response* perturbation model were computed setting the estimates of α and β into the equation $x(1-\beta)+(1-x)\alpha=y$ and solving it by x.

Table 2 reports, for each perturbation model and for each variable, the adopted perturbation percentages x.

Table 2. Perturbation percentages

Perturbation model	Variable				
	Relation	*Sex*	*Marital Status*	*Age*	*Years married*
Non response	0.52	0.50	1.30	0.40	1.70
Other valid response	4.08	3.17	2.01	3.22	0.30

In order to analyse the performance of the two systems along the different error incidences, the perturbation percentages in Table 2 have been systematically varied by multiplying them by the following factors: 0.5, 1, 1.5 and 2. In that way we have results coming from four applications at different levels of perturbation percentages (in the discussion below, these levels are numbered from 1 to 4, where a factor of 1 applied to the perturbation percentages of Table 2 corresponds to the perturbation level 2). In each run, new perturbed data were generated and processed by the two systems.

Initially, with the CANCEIS runs, the process of perturbation, E&I and evaluation was replicated several times in order to measure the variability of the evaluation indicators. As very low variability was observed, we decided to perform only one run for each level of perturbation. That allowed us to retain the eight perturbed data sets and to process them against DIESIS (the comparative evaluation of the two performances is based on evaluation indicators computed on the same data sets instead of average values of indicators).

For each variable, each perturbation level and each household dimension, we compute indicators evaluating the error detection performance of the editing process and indicators for evaluating how well the imputation procedure preserves the individual values and the marginal distribution. Some indicators are defined by the EUREDIT project (Charlton et al., 2001).

We assess the accuracy of an editing method by computing:

- the percentage of not modified data erroneously imputed (*E_true*);
- the percentage of modified data not imputed (*E_mod*).

We assess the accuracy of an imputation method by comparing *imputed* against *original* in individual values as well as in the marginal distributions.

We evaluate the preservation of individual original values by means of :

- the percentage of imputed values for which imputation is a failure (*I_imp*).
  For qualitative variables (*Relation to Person 1*, *Sex*, *Marital Status*) the imputation process is considered as a failure if the imputed value does not equal the original one. For numeric variables (*Age* and *Years married*) the imputation process is considered as a failure if the imputed value differs by more than ±10% from the original one.
- the average absolute deviation between imputed and original values (only for numeric variables):

$$\mathbf{d} = \frac{\sum_{1}^{\mathbf{n_{imp}}} \left| \mathbf{y_i'} - \mathbf{y_i^*} \right|}{\mathbf{n_{imp}}}$$

It is an indicator of the distance between imputed (y')and original values (y*).

We evaluate the preservation of the marginal distribution of the original values by means of:

- the simple relative dissimilarity index (Leti, 1983) between the relative distribution of corrected values and the relative distribution of original values

$$\Phi = \left(\sum_{1}^{\mathbf{r}} \left| \mathbf{f(i)} - \mathbf{g(i)} \right| / 2 \right) * 100$$

where f(i) and g(i) are the relative frequencies of the i-th value in the distributions, ϕ is an indicator of the distance between the two relative distributions and varies between 0 (the relative distributions are homogeneous) and 100 (maximum dissimilarity between the relative distributions). For numeric variables (*Age* and *Years married*) the dissimilarity index was computed after categorising the distribution of values by 5-years classes and taking the cumulative frequencies. We chose this index because it is suitable to compare relative distributions coming from large samples having different sizes. Even if it does not measure very well a large relative difference in the proportions for a small class, it gives a good overview of what is happening.

The previous indicators have been computed for each selected demographic variable.

As Census data are the primary source of information about the *household typology*, that is a variable derived from all the demographic variables, it is of interest to evaluate the capability of E&I systems to preserve the original value and the distribution of this summary variable. The *household typology* variable is computed at the household level from the demographic variables that are given at the individual level.

We define seven categories of *household typology* based on the *family nucleus* definition. A *family nucleus* is a married or not married couple or a one-parent family with at least a child. The categories are as follows (co-habitants stands for other persons that have not a familiar relationship to the person1):

- ✔ one-person family with co-habitants (one person without children and with co-habitants)
- ✔ couple with children and co-habitants
- ✔ couple with children without co-habitants
- ✔ couple without children with co-habitants
- ✔ one-parent family with co-habitants (one person with children and with co-habitants)
- ✔ one-parent family without co-habitants

✔ extended families (two or more *family nuclei*)

We evaluate the preservation of the household typology variable by computing:

- the percentage of off-diagonal entries for the square tables obtained by cross-classifying the original and corrected typology categories;
- the simple relative dissimilarity index between original and corrected distribution of the typology variable.

4. Results

As regards CANCEIS, the results refer to runs performed setting the default values for all system parameters (Janes, 2000) except for the value of Number of 1° Stage Donors parameter that was set to 2000 and for the Ordering method that was set to 3 (the Iterative Method of Position Selection). The 0.9 value was set to the α parameter in the D_{fpa} distance function.

As regards DIESIS system, the results refer to runs performed by selecting the *first donors then fields* imputation algorithm with the *involved sub-units donor* and *involved sub-units imputation action* options. In selecting the imputation action the value of 1 was set to the weight (c_i) of each variable.

The used distance functions were the same for both systems.

Table 3 reports frequencies of failed and passed households.

Table 3. Frequency of failed and passed households

Perturbation level	Four-person household		Six-person household	
	Failed	Passed	Failed	Passed
1	10288 *22.5%*	35428 *77.5%*	6174 *30.4%*	14132 *69.6%*
2	18496 *40.5%*	27220 *59.5%*	10426 *51.3%*	9880 *48.7%*
3	24662 *54.0%*	21054 *46.0%*	13322 *65.6%*	6984 *34.4%*
4	29450 *64.4%*	16266 *35.6%*	15393 *75.8%*	4913 *24.2%*

A household fails the edit rules if the combination of its data corresponds to one or more of the edit rules. This causes the percentage of failed households to be rather high even if the adopted perturbation levels are low (see Table 2). This results in a low number of passed households, that can be considered as donors, especially for the six-person households at the four perturbation level.

Tables 4a and 4b report the values of the following failure indicators:

E_true = percentage of not modified values erroneously imputed;

E_mod = percentage of modified values not imputed;

I_imp = percentage of imputed values for which imputation is a failure.

Table 4a. Indicators of preservation of individual values. Four-person household

Variable	Perturbation level	CANCEIS			DIESIS		
		E_true	E_mod	I_imp	E_true	E_mod	I_imp
Relation to Person 1	1	0.01	38.81	4.87	0.01	38.28	4.38
	2	0.03	39.17	6.14	0.04	38.78	5.04
	3	0.08	39.43	9.46	0.10	39.08	7.88
	4	0.16	39.36	11.42	0.15	38.73	8.92
Sex	1	0.02	46.88	8.41	0.01	46.30	7.50
	2	0.04	48.56	8.20	0.04	47.63	8.35
	3	0.05	50.68	8.65	0.05	49.77	8.73
	4	0.06	53.81	9.69	0.07	52.47	9.60
Marital Status	1	0.02	3.00	1.85	0.02	2.90	2.09
	2	0.05	3.45	2.03	0.05	3.27	2.21
	3	0.07	4.93	2.07	0.07	4.04	2.33
	4	0.11	5.44	2.57	0.12	4.54	2.75
Age	1	0.16	35.88	50.67	0.09	32.90	53.49
	2	0.31	35.43	50.57	0.18	32.77	53.61
	3	0.43	36.99	52.99	0.27	33.75	53.19
	4	0.63	38.54	53.95	0.37	34.85	53.04
Years married	1	0.25	3.55	14.56	0.05	4.45	8.15
	2	0.50	3.49	16.53	0.12	4.78	9.18
	3	0.76	3.80	17.36	0.16	4.91	9.53
	4	1.03	4.86	18.27	0.22	6.10	11.22

Table 4b. Indicators of preservation of individual values. Six-person household

Variable	Perturbation level	CANCEIS			DIESIS		
		E_true	E_mod	I_imp	E_true	E_mod	I_imp
Relation to Person 1	1	0.07	44.64	14.42	0.07	44.54	14.21
	2	0.15	43.58	15.45	0.14	43.08	15.57
	3	0.27	42.29	18.80	0.21	42.14	16.45
	4	0.41	43.80	21.18	0.30	43.42	18.20
Sex	1	0.03	58.52	11.13	0.02	57.72	9.45
	2	0.05	59.52	11.96	0.04	58.48	8.71
	3	0.08	62.89	13.74	0.08	61.20	11.29
	4	0.11	64.10	13.72	0.11	61.79	11.30
Marital Status	1	0.05	9.32	4.99	0.05	10.70	6.06
	2	0.10	10.62	6.02	0.13	10.65	7.16
	3	0.19	11.68	6.73	0.21	11.06	7.46
	4	0.28	12.68	7.38	0.29	11.58	7.95
Age	1	0.20	40.26	49.48	0.15	36.95	50.68
	2	0.39	41.63	50.88	0.29	37.75	50.29
	3	0.56	41.82	51.02	0.43	37.81	50.27
	4	0.79	43.12	52.54	0.61	38.82	50.57
Years married	1	0.44	7.92	24.71	0.12	8.00	22.22
	2	0.93	8.03	23.90	0.22	8.83	19.90
	3	1.34	10.75	26.77	0.34	10.39	22.67
	4	1.71	11.86	26.51	0.44	11.21	22.61

The first index (*E_true*) represents the editing failure on true values. High values indicate that the E&I system causes a loss of information by means of the subsequent imputation (new errors are introduced in data). The second index (*E_mod*) represent the editing failure on modified values. High values indicate that the E&I system is not able to localise errors in data (variable to impute). The third index (*I_imp*) represents the imputation failure on imputed values. High values indicates that the E&I system does not restore individual values in imputed data.

We observe that, in general, all indexes values are higher when the perturbation level and household dimension increase.

The comparison between the two systems shows a similar performance of *E_true* for the qualitative variables and a better performance of DIESIS for numeric variables.

As regards *E_mod* index for qualitative variables, we observe a general better performance of DIESIS. In case of numeric variables, figures show a better performance of DIESIS for *Age* and a general better performance of CANCEIS for *Years married.*

As regards *I_imp*, results are different by household dimension. It is however of interest to observe that the index value from DIESIS is less variable than the index value from CANCEIS at the increase of the perturbation level.

Tables 5a and 5b report for *Age* and *Years married* (numeric variables) the mean values of original, perturbed and imputed data together with the average absolute deviation between imputed and original values. Note that each figures in Tables 5a and 5b is computed on the subset of values imputed by the considered system, and the subsets of values imputed by DIESIS can be different from the subset of values imputed by CANCEIS.

Table 5a. Indicators of preservation of individual values. Four-person household (subset of imputed values)

Variable	Perturb. level	CANCEIS				DIESIS			
		Mean original	Mean perturbed	Mean imputed	Average absolute deviation	Mean original	Mean perturbed	Mean imputed	Average absolute deviation
Age	1	32.9	60.7	33.6	4.5	32.5	61.8	32.9	4.7
	2	32.8	59.8	33.6	4.8	32.7	61.0	33.3	4.8
	3	32.6	60.4	33.7	5.0	32.5	61.5	33.3	4.8
	4	32.9	60.2	34.2	5.3	32.8	61.5	33.5	4.9
Years married	1	22.3	32.7	22.1	1.1	22.1	46.4	21.7	0.8
	2	22.0	32.0	21.9	1.1	21.8	45.6	21.5	0.8
	3	21.8	32.4	21.7	1.2	21.6	46.3	21.3	0.9
	4	21.7	32.5	21.6	1.3	21.5	46.3	21.0	1.0

Table 5b. Indicators of preservation of individual values. Six-person household (subset of imputed values)

Variable	Perturb. level	CANCEIS				DIESIS			
		Mean original	Mean perturbed	Mean imputed	Average absolute deviation	Mean original	Mean perturbed	Mean imputed	Average absolute deviation
Age	1	36.7	59.4	37.3	4.9	35.9	59.3	36.9	5.8
	2	37.8	59.8	38.8	5.5	36.9	59.9	38.2	6.2
	3	37.6	58.6	38.5	5.8	37.2	59.1	38.4	6.2
	4	37.6	58.1	38.8	6.0	37.2	59.1	38.4	6.5
Years married	1	26.8	32.5	26.6	1.9	27.2	47.6	25.4	2.6
	2	26.9	32.3	26.7	2.0	26.7	45.0	25.4	2.1
	3	26.9	33.4	26.7	2.2	26.8	46.4	25.6	2.5
	4	27.3	33.4	26.9	2.2	27.4	46.8	25.6	2.5

Please note that we perturbed the *Age* variable with values taken from a random uniform distribution in the range [0, 110], while the *Years married* variable was perturbed with values taken from a random uniform distribution in the range [0, 91].

It is of interest to note the capability of the two system in restoring the original mean values in spite of the bias caused by the perturbation.

The figures show that for six-person households the distance between imputed and original values is lower with CANCEIS, while DIESIS works better in case of four-person households.

Tables 6a and 6b report the dissimilarity index between original and corrected marginal distribution.

Table 6a. Dissimilarity index between original and corrected distribution. Four-person household

Variable	Perturbation level	CANCEIS		DIESIS	
		Imputed values	Total values	Imputed values	Total values
Relation to Person1	1	1.9	0.9	1.8	0.9
	2	3.8	1.8	2.8	1.7
	3	6.4	2.8	5.0	2.7
	4	8.4	3.8	6.2	3.6
Sex	1	0.4	0.0	0.2	0.0
	2	0.3	0.0	0.3	0.0
	3	0.8	0.0	0.2	0.0
	4	0.1	0.2	0.7	0.2
Marital status	1	1.3	0.0	1.2	0.0
	2	1.3	0.1	1.2	0.1
	3	1.4	0.2	1.3	0.1
	4	1.8	0.3	1.8	0.3
Age	1	1.0	0.1	0.7	0.1
	2	1.1	0.2	0.7	0.1
	3	1.4	0.3	1.0	0.2
	4	1.6	0.4	0.9	0.3
Years married	1	0.3	0.0	0.6	0.0
	2	0.4	0.0	0.4	0.0
	3	0.3	0.0	0.5	0.1
	4	0.3	0.1	0.9	0.1

Table 6b. Dissimilarity index between original and corrected distribution. Six-person household

Variable	Perturbation level	CANCEIS		DIESIS	
		Imputed values	Total values	Imputed values	Total values
Relation to Person1	1	4.6	0.8	5.0	0.8
	2	5.8	1.6	6.3	1.6
	3	6.9	2.4	6.7	2.3
	4	10.3	3.5	8.0	3.3
Sex	1	0.5	0.1	1.1	0.1
	2	0.4	0.0	0.2	0.0
	3	0.2	0.2	0.8	0.2
	4	0.2	0.1	0.5	0.1
Marital status	1	3.5	0.1	2.3	0.1
	2	3.5	0.3	3.1	0.3
	3	4.6	0.5	4.0	0.5
	4	5.0	0.8	4.7	0.7
Age	1	0.7	0.1	1.3	0.1
	2	1.2	0.2	1.6	0.2
	3	1.0	0.3	1.4	0.3
	4	1.3	0.5	1.4	0.4
Years married	1	0.3	0.0	1.7	0.0
	2	0.5	0.1	1.3	0.1
	3	0.6	0.1	1.4	0.1
	4	0.7	0.2	2.5	0.2

As regard the subsets of imputed values, the figures in Tables 6a and 6b show a general better performance of CANCEIS for numeric variables while for qualitative variables the results are different by variables and perturbation levels. On total values both systems show an equally good performance for all variables.

Tables 7-8 report the results concerning the preservation of the summary variable *household typology*.

Table 7a. Percentage of off-diagonal entries for the square tables of original vs. corrected typology categories of the typology variable. Four-person household

Perturbation level	CANCEIS		DIESIS	
	Imputed households	Total households	Imputed households	Total households
1	4.7	3.7	3.6	3.1
2	7.9	7.3	7.1	6.2
3	11.9	11.3	11.0	9.4
4	16.1	15.4	14.7	12.5

Table 7b. Percentage of off-diagonal entries for the square tables of original vs. corrected typology categories of the typology variable. Six-person household

Perturbation level	CANCEIS		DIESIS	
	Imputed households	Total households	Imputed households	Total households
1	5.3	3.7	3.6	1.9
2	8.0	6.8	6.7	3.4
3	11.4	10.5	10.1	5.2
4	15.5	14.7	13.7	6.7

As regards the percentage of off-diagonal entries for the square tables of original vs. corrected categories of the typology variable, the figures show a better performance of DIESIS system. The gap is larger for the set of total values than for the subset of imputed values and for six-person households than for four-person households.

Table 8a. Dissimilarity index between original and corrected distribution of typology variable. Four-person household

Perturbation level	CANCEIS		DIESIS	
	Imputed households	Total households	Imputed households	Total households
1	3.5	3.2	3.3	3.1
2	6.6	6.3	6.1	6.2
3	9.9	9.7	9.4	9.4
4	13.4	13.0	12.5	12.5

Table 8b. Dissimilarity index between original and corrected distribution of typology variable. Six-person household

Perturbation level	CANCEIS		DIESIS	
	Imputed households	Total households	Imputed households	Total households
1	2.3	1.9	2.1	1.9
2	3.8	3.4	3.8	3.4
3	5.2	5.2	5.2	5.2
4	7.0	7.0	6.6	6.7

As regards the dissimilarity index between original and corrected distribution of typology variable, we observe a slightly better performance of DIESIS for four-person households. For six-person households, we observe an equal performance for both systems.

Finally, Table 9 reports the number of imputed values.

Table 9. Number of imputed values

Perturbation level	Four-person household		Six-person household	
	CANCEIS	DIESIS	CANCEIS	DIESIS
1	12048	11682	7818	7428
2	24058	23413	15404	14689
3	35519	34676	22488	21536
4	46703	45635	29520	28489

We observe lower numbers of imputed values by DIESIS software.

It is our opinion that the differences observed in the presented results are due to the different approaches used by the two systems in selecting donors and imputation action. The approaches are only outlined in section 2, but a more detailed specification is probably necessary for a better understanding of the difference between them. Moreover, for a given approach, different results can arise from different specifications of the parameters.

We are supported in our opinion by the results obtained from an additional CANCEIS run, on six-person households at the perturbation level four, that was executed setting the 0.9999 value to the α parameter in the D_{fpa} distance function. We did not set the 1 value to the α parameter because the present CANCEIS version asks for it a value less than 1. However 0.9999 is very close to 1 and in that case the adjusted unit was requested to be very similar only to the failed unit, because negligible weight was being put on having the imputation action resemble the donor. Doing that we forced CANCEIS to mimic the DIESIS application, that is, selection of the imputation action based on minimising the distance between the adjusted unit and the failed unit, in other words, based on a function of the imputed variables. As result, a lower number of values was imputed by CANCEIS (28618). This means that most of the difference between the two systems, in terms of number of values imputed, was due to the difference in the variables used in the function to minimise in selecting the imputation action.

Note that, even if the minimised functions used only the contributions from the imputed variables, another source of difference between the results could be in the values used for the distance scores and the weights. Remember that in our test, for numeric variables, CANCEIS used a distance function in the range [0,1] both in selecting donors and in selecting the imputation action while DIESIS used a distance function in the range [0,1] only in selecting donors and used weights 1 in selecting the imputation action.

Other source of the difference between the results could be in the number of sub-units used in searching the donors and in selecting the imputation action (*all sub-units* option versus *involved sub-units* option). To verify this we could execute additional runs of DIESIS setting the *all sub-unit donor* option and the *all sub-units imputation action* option, that is, forcing DIESIS to mimic CANCEIS.

In conclusion, it is our opinion that additional investigations need to be performed in order to explain the sources of the difference in the observed results.

5. Conclusions

In this paper we present the results of a study to evaluate the performance (in terms of *accuracy*) of a new system, DIESIS, against the CANCEIS system, in presence of different percentages of random errors. The adopted evaluation procedure (error simulation and accuracy indicators) provides a rigorous statistical evaluation of the comparative performance of the two systems. The results allow us to state that the two systems do not show great difference in the quality and provide support for further development of DIESIS software in order to obtain a powerful generalised E&I system for the treatment of qualitative and numeric variable simultaneously. Further work will be addressed to tuning and testing DIESIS with data from social surveys and also with data from business surveys.

Acknowledgement

Special thanks to Mike Bankier for his precious comments and suggestions.

REFERENCE

Bankier M. (1999) Experienced with the New Imputation Methodology used in the 1996 Canadian Census with extension for future Censuses, *Proceedings of the Workshop on Data Editing*, UN/ECE, Italy (Rome).

Bankier M. (2000) Canadian Census Minimum change Donor imputation methodology, *Proceedings of the Workshop on Data Editing*, UN/ECE, United Kingdom (Cardiff).

R. Bruni, A. Reale, R. Torelli (2001) Optimization Techniques for Edit Validation and Data Imputation, presented at the Statistics Canada Symposiuum 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective" *XVIIIth International Symposiuum on Methodological Issues.*

Charlton J., Chambers R. Nordbotten S. (2001) New developments in edit and imputation practices – needs and researches, *Proceedings of the 53rd Session of the International Statistical Institute,* Korea(Seoul).

Fellegi I. P. e Holt D. (1976) A systematic approach to edit e imputation, *Journal of the American Statistical Association,* vol.71, pp. 17-35.

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, pp. 112-123.

Janes D. (2001) CANCEIS version 1.2 Users' Guide

Leti G. (1983) Statistica descrittiva, Il Mulino, Bologna.

Loveland D.W. (1978). Automated Theorem Proving: a Logical Basis. North Holland.

Nemhauser G. L. and Wolsey L. A. (1988) Integer and Combinatorial Optimization. J. Wiley, New York.

Riccini E., Silvestri F., Barcaroli G., Ceccarelli C., Luzi O., Manzari A. (1995) La metodologia di editing e imputazione per variabili qualitative implementata in SCIA, *Documento interno ISTAT.*

Winkler W. E. (1999) State of Statistical Data Editing and current Research Problems, *Proceedings of the Workshop on Data Editing*, UN/ECE, Italy (Rome)

APPENDIX

Between persons and *within person* conflict edit rules defined in the evaluation study.

In the following *relation* stands for *Relation to person1; mstatus* for *Marital Status; age* for *Age in years;* and *ymarried* for *Years married.*
The indices *i,j,k* point out the position of the person and could take on values 2 to 6.

1 *relation*(i) = spouse and *relation*(j) = spouse
2 *relation*(i) = common-law spouse and *relation*(j) = common-law spouse
3 *relation*(i) = common-law spouse and *relation*(j) = spouse
4 *relation*(i) = spouse and *sex*(1) = *sex*(i)
5 *relation*(i) = common-law spouse and *sex*(1) = *sex*(i)
6 *relation*(i) = spouse and *ymarried*(1) ≠ *ymarried* (i)
7 *relation*(i) = spouse and *mstatus*(1) ≠ *mstatus* (i)
8 *relation*(i) = common-law spouse and *mstatus*(1) = spouse
9 *relation*(i) = parent and *relation*(j) = parent and *relation*(k) = parent

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

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

12 *relation*(i) = parent and *relation*(j) = parent and *mstatus*(i) = *married* and *mstatus*(j) ≠ *married*

13 *relation*(i) = parent and *relation*(j) = parent and *mstatus*(i) ≠ *married* and *mstatus*(j) = *married*

14 *relation*(i) = parent and *relation*(j) = parent and *mstatus*(i) = *married* and *mstatus*(j) = *married* and *ymarried*(i) ≠ *ymarried*(j)

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

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

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

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

19 *relation*(i) = son/daughter and *relation*(j) = spouse or common-law spouse and *age*(1)-*age*(i) <12 and *age*(j)-*age*(i) <12

20 *relation*(i) = son/daughter and there is no spouse or common-law spouse and *age*(1)-*age*(i) <12

21 *sex*(1) = male and *relation*(i) = son/daughter and *age*(1)-*age*(i) >70

22 *sex*(1) = female and *relation*(i) = son/daughter and *age*(1)-*age*(i) >55

23 *sex*(i) = male and *relation*(i) = spouse or common-law spouse and *relation*(j) = son/daughter and *age*(i)-*age*(j) >70

24 *sex*(i) = female and *relation*(i) = spouse or common-law spouse and *relation*(j) = son/daughter and *age*(i)-*age*(j) >55

25 *relation*(i) = parent and *age*(i)-*age*(1) <12

26 *relation*(i) = parent and *sex*(i) = male and *age*(i)-*age*(1) >70

27 *relation*(i) = parent and *sex*(i) = female and *age*(i)-*age*(1) >55

28 *relation*(i) = father/mother-in-law and *relation*(j) = spouse or common-law spouse and *age*(i)-*age*(j) <12

29 *relation*(i) = father/mother-in-law and *sex*(i) = male and *relation*(j) = spouse or common-law spouse and *age*(i)-*age*(1) >70

30 *relation*(i) = father/mother-in-law and *sex*(i) = female and *relation*(j) = spouse or common-law spouse and *age*(i)-*age*(1) >55

31 *relation*(i) = parent and *relation*(j) = brother/sister and *age*(i)-*age*(j) <12

32 *relation*(i) = parent and *sex*(i) = male and *relation*(j) = brother/sister and *age*(i)-*age*(j) >70

33 *relation*(i) = parent and *sex*(i) = female and *relation*(j) = brother/sister and *age*(i)-*age*(j) >55

34 *relation*(i) = grandchild and *age*(1)-*age*(i) <28

35 *relation*(i) = son/daughter and *relation*(j) = parent and *age*(j)-*age*(i) <28

36 *relation*(i) = grandchild and *relation*(j) = parent and *age*(j)-*age*(i) <42

37 *relation*(i) = son/daughter and *relation*(j) = son/daughter and |*age*(i)-*age*(j)| >48

38 *relation*(i) = brother/sister and *relation*(j) = brother/sister and |*age*(i)-*age*(j)| >48

39 *relation*(i) = brother/sister and |*age*(i)-*age*(1)| >48

40 *relation*(i) = spouse or common-law spouse and |*age*(i)-*age*(1)| >35

41 *relation*(i) = parent and *relation*(j) = parent and |*age*(i)-*age*(j)| >35

42 *relation*(i) = father/mother-in-law and *relation*(j) = father/mother-in-law and |*age*(i)-*age*(j)| >35

43 *relation*(1) ≠ person 1

44 *relation*(i) = person 1

45 *age*(1) ≤14

46 *mstatus*(1) ≠ single and *ymarried*(1) = blank

47 *mstatus*(i) ≠ single and *ymarried*(i) = blank

48 *mstatus*(1) = single and *ymarried*(1) ≠ blank

49 *mstatus*(i) = single and *ymarried*(i) ≠ blank
50 *ymarried*(1) ≠ blank and *age*(1)-*ymarried*(1)<14
51 *ymarried*(i) ≠ blank and *age*(i)-*ymarried*(i)<14
52 *mstatus*(i) ≠ single and *age*(i) <14
53 *mstatus*(1) = divorced and *age*(1) <17
54 *mstatus*(i) = divorced and *age*(i) <17
55 *relation*(i) = spouse or common-law spouse or son/daughter-in-law and *age*(i) <14
56 *relation*(i) = parent or father/mother-in-law and *age*(i) <26
57 *relation*(i) = spouse and *mstatus*(i) ≠ *married*
58 *relation*(i) = spouse and *ymarried*(i) = blank
59 *relation*(i) = common-law spouse and *mstatus*(i) = married
60 *couple*(i)=*couple*(j) and *relation*(i) = son/daughter and *relation*(j)= son/daughter-in-law and *sex*(i) = *sex*(j)
61 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister and *relation*(j)= brother/sister -in-law and *sex*(i) = *sex*(j)
62 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister -in-law and *relation*(j)= brother/sister -in-law and *sex*(i) = *sex*(j)
63 *couple*(i)=*couple*(j) and *relation*(i) = son/daughter and *relation*(j)= son/daughter-in-law and *age*(i) <14
64 *couple*(i)=*couple*(j) and *relation*(i) = son/daughter and *relation*(j)= son/daughter-in-law and *age*(j) <14
65 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister and *relation*(j)= brother/sister -in-law and *age*(i) <14
66 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister and *relation*(j)= brother/sister -in-law and *age*(j) <14
67 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister -in-law and *relation*(j)= brother/sister -in-law and *age*(i) <14
68 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister -in-law and *relation*(j)= brother/sister -in-law and *age*(j) <14
69 *couple*(i)=*couple*(j) and *relation*(i) = son/daughter and *relation*(j)= son/daughter-in-law and *mstatus*(i) = married and *mstatus*(j) ≠ married
70 *couple*(i)=*couple*(j) and *relation*(i) = son/daughter and *relation*(j)= son/daughter-in-law and *mstatus*(i) ≠ married and *mstatus*(j) =married
71 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister and *relation*(j)= brother/sister -in-law and *mstatus*(i) = married and *mstatus*(j) ≠ married
72 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister and *relation*(j)= brother/sister -in-law and *mstatus*(i) ≠ married and *mstatus*(j) =married
73 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister -in-law and *relation*(j)= brother/sister -in-law and *mstatus*(i) = married and *mstatus*(j) ≠ married
74 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister -in-law and *relation*(j)= brother/sister -in-law *mstatus*(i) ≠ married and *mstatus*(j) =married
75 *couple*(i)=*couple*(j) and *relation*(i) = son/daughter and *relation*(j)= son/daughter-in-law and *mstatus*(i) = married and *mstatus*(j) = married and *ymarried*(i) ≠ *ymarried*(j)
76 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister and *relation*(j)= brother/sister -in-law and *mstatus*(i) = married and *mstatus*(j) = married and *ymarried*(i) ≠ *ymarried*(j)
77 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister -in-law and *relation*(j)= brother/sister -in-law and *mstatus*(i) = married and *mstatus*(j) = married and *ymarried*(i) ≠ *ymarried*(j)
78 *couple*(i)=*couple*(j) and *relation*(i) = son/daughter and *relation*(j)= son/daughter-in-law and |*age*(i)-*age*(j)| >35

79 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister and *relation*(j)= brother/sister -in-law and |*age*(i)-*age*(j)| >35

80 *couple*(i)=*couple*(j) and *relation*(i) = brother/sister -in-law and *relation*(j)= brother/sister -in-law and |*age*(i)-*age*(j)| >35