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
Istituto Nazionale di Statistica
Via C. Balbo, 16
00184 Roma, Italy
manzari@istat.it,
Alessandra Reale
Istituto Nazionale di Statistica
Via C. Balbo, 16
00184 Roma, Italy
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 level with information for each person 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 into 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 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 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 actions addressed 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 components. 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 PC demographic variables and new research actions were undertaken. Among different research activities we consider the one concerning the joint development of a new software by ISTAT and academic researchers (Dipartimento di Informatica e Sistemistica dell' Università degli Studi di Roma "La Sapienza"). The new software performance will be 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 reports an overview of the new software (section 2), describes the evaluation study (section 3) and presents the early results of the CANCEIS application (section 4). The new software application will be performed in next future and its results will be presented afterwards, together with the overall conclusions.

2. The new software

The new software (Bruni et al., 2001), that is still being developed, allows the joint editing and correction of both qualitative and numeric variables. Edit rules can be defined in linear and non-linear inequalities format. The software checks for redundancies and logical inconsistencies of the edit rules.

The software performs donor imputation. Two imputation algorithms have been implemented. The first one identifies the absolute weighted minimum number of fields (variables) to change to assure that the modified record satisfied all edits; the second one identifies the weighted minimum number of fields to change, given a set of available donors. The two algorithms can be separately or jointly used. The user chooses the adopted strategy. Both imputation algorithms are based on optimization techniques that allow to overcome some computational limits related to the implicit edit generation and to the redundancy and consistency check (Winkler, 1999).

The adopted approach consists in defining the editing and imputation problems as combinatorial optimization problems. That is possible converting the set of edits in a *propositional Logic Formula*. In that way

the problems of consistency and non-redundancy can be converted into a sequence of propositional Satisfiability problems (Loveland, 1978). The computationally issues are solved by means of CLAS, a very efficient solver proposed by Bruni and Sassano (1999). The imputation problems are faced encoding the edit rules in linear inequalities, and therefore converting them into a sequence of set covering problems (Nemhauser and Wolsey, 1988). The relating computationally problems are solved by means of specialized algorithms.

3. The evaluation study

The evaluation of the performance is based on a comparison among *true* data (free of errors), *raw* data (contaminated by errors) and *clean* data (edited and imputed values) (Granquist, 1997). *True* data correspond to 1991 PC returns free of errors with respect to a defined set of *between persons* and *within person* edit rules (not reported because of space restrictions). *Raw* data are obtained by introducing artificial controlled errors in *true* data. *Clean* data are obtained by processing the *raw* data by 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 in *Age (in years)* and *Years married* in order to define the relating edits. Potential couples, which have non-unique relationships to Person 1, are identified prior to imputation in order to apply to them couple edits.

CANCEIS system runs on imputation groups having the same number of subunits (persons in the household). We analyse the performance on two different household dimensions: four-person households (45716 units from a single district) and six-person households (20306 units from a single region).

Artificial nonsampling errors are introduced into true data using the *item non-response* model (the true value is randomly replaced by a missing value) and the *interchange error* model (the true value is replaced by a wrong one randomly chosen in the admissible domain). For each variable, the perturbation percentage is chosen in such a way to closely resemble real life situation. We assume that 1991 editing procedure had the same sensitivity (probability to recognise modified values as erroneous = 1-β) and specificity (probability to recognise unmodified values as true = 1-α) of the CANCEIS editing procedure, and computed the frequencies of missing and modified values (x), to be introduced into true data, by adjusting the observed 1991 changing frequencies (y) according to the estimates of α and β. The estimates of α and β to set into the following equation $x(1-\beta)+(1-x)\alpha=y$, have been computed, for each variable, averaging the values obtained from three applications of CANCEIS system (running on four-person households) having only the *interchange error* perturbation model setting at different perturbation percentages (1%, 5% and 10%). Table 1 reports, for each variable, the adopted perturbation percentages.

Table1. Perturbation percentages

Errors model	**Variable**				
	Relation	*Sex*	*Marital Status*	*Age*	*Years married*
Item non response	0.52	0.50	1.30	0.40	1.70
Interchange error	4.08	3.17	2.01	3.22	0.30

In order to analyse the performance of the two systems along the different errors incidences, the perturbation percentages have been then systematically varied by multiplying them by the following factors: 0.5, 1.5, 2. In that way we have results coming from four perturbation applications (in the following labelled by a number from 1 to 4, where the reference one, corresponding to the percentages in Table 1 is the second one). In each application new errors (*raw data*) are generated and processed against the two systems.

At the first CANCEIS application, the process of simulation, editing and evaluation was replicated more times in order to measure the variability of the evaluation indicators. As very low variability was observed, we decided to perform only a run for each application. That allows us to retain the four perturbed data sets and to process them against the new software (the comparative evaluation of the two performances will be based on evaluation indicators computed on the same data sets instead of average values of indicators). For each run the default values of system parameters are used (Janes, 2001).

For each variable, each perturbation application and each household dimension, the indicators evaluating the error detection performance of the editing process and the indicators for evaluating how well the imputation procedure preserves the marginal distribution and the true values are computed among those defined inside the EUREDIT project (Chambers, 2001). We also evaluate the capability of the two systems to preserve the distribution and the true values of a derived variable that synthesises the household typology (the derived variable is computed at the household level).

4. CANCEIS results

This section presents the early results concerning CANCEIS system.

For each perturbation application and the different household dimensions, Table 3 reports numbers and percentage of failed households (column 2 and 5). A household fails the edit rules if the combination of its data activates one or more of the rules. This makes the percentage of failed households rather high even if the adopted perturbation levels are low (see Table 1).

For each variable, each perturbation application and each given household dimension, Table 2 reports the values of some failure indicators. Columns 3 and 7 show the failure of the editing process in recognising *true values* (E_true=fraction of true data erroneously classified, it estimates α), columns 4 and 8 show the failure of the editing process in recognising *modified values* (E_mod = fraction of modified data erroneously classified, it estimates β), while columns 5 and 9 show the values of overall failure indicator of the editing process (E_tot = fraction of total data erroneously classified). For qualitative variables (*Relation to Person 1*, *Sex*, *Marital Status*) the imputation process is considered as a failure if the imputed value is different from the original one. For numeric variables (*Age* and *Years married*) the imputation process is considered as a failure if the imputed value lays out the interval of ±10% around the original value. Columns 6 and 10 show the values of inaccuracy indicator of the imputation process (I_imp = fraction of imputed values for which the imputation is a failure).

Table 2. Failure indicators - CANCEIS applications (percentage values)

Variable	Application	E_true	E_mod	E_tot	I_imp	E_true	E_mod	E_tot	I_imp
		Four-person households				Six-person households			
Relation to Person 1	1	0.02	38.99	0.94	5.16	0.09	44.71	1.13	15.34
	2	0.06	39.46	1.86	7.62	0.18	43.65	2.21	17.91
	3	0.13	39.72	2.85	10.60	0.34	42.51	3.25	20.80
	4	0.25	39.45	3.86	13.04	0.50	44.00	4.55	24.41
Sex	1	0.02	46.88	0.87	8.21	0.03	58.65	1.17	12.35
	2	0.04	48.91	1.86	8.80	0.06	60.25	2.30	13.54
	3	0.05	51.24	2.92	8.80	0.08	63.54	3.60	14.43
	4	0.07	54.35	4.09	9.50	0.10	64.23	4.80	14.91
Marital Status	1	0.03	3.17	0.08	2.56	0.06	9.79	0.23	5.80
	2	0.06	3.87	0.19	2.53	0.12	10.89	0.48	6.93
	3	0.09	5.53	0.36	2.58	0.23	12.49	0.83	7.72
	4	0.15	6.29	0.56	3.26	0.33	13.49	1.20	8.37
Age	1	0.18	36.61	0.84	52.89	0.19	40.90	0.92	51.62
	2	0.34	36.56	1.67	52.80	0.39	42.01	1.89	52.21
	3	0.47	37.65	2.52	54.32	0.55	42.68	2.84	54.62
	4	0.63	39.58	3.44	55.00	0.79	43.38	3.83	55.73
Years married	1	0.44	4.08	0.48	19.23	0.50	8.76	0.58	27.11
	2	0.88	3.90	0.94	19.81	1.02	9.56	1.19	27.76
	3	1.27	4.57	1.37	21.00	1.45	12.01	1.76	30.50
	4	1.74	5.63	1.89	22.27	1.84	13.37	2.29	30.95

The derived variable that synthesises the household typology is based on the *family nucleus* definition. A *family nucleus* is a couple (married or not married) or a one-parent family (with at least a child). Seven categories have been defined for the typology variable: one-person family with co-habitants, couple with children and co-habitants, couple without children with co-habitants, one-parent family with co-habitants, couple with children without co-habitants, one-parent family without co-habitants, extended families (two or more *family nuclei*).

Table3. Error rate and dissimilarity index for typology variable (on the imputed households)

Application	Failed households	Error rate * 100	ϕ* 100	Failed households	Error rate * 100	ϕ* 100
	Four-person household			Six-person household		
1	10288	4.76	3.51	6174	5.39	2.12
2	18496	8.43	6.82	10426	8.34	3.75
3	24662	12.39	10.12	13322	11.90	5.58
4	29450	16.63	13.76	15393	16.04	7.28

The extent to which the marginal distribution of the typology variable after the imputation procedure is homogeneous to the marginal distribution of the typology variable in original data is assessed by computing, on the total number of imputed households, the simple relative dissimilarity index: $\phi=(\Sigma|f(i)-g(i)|)/2$ (Leti, 1983) 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 1 (maximum dissimilarity between the relative distributions). The values of ϕ, multiplied by 100, are reported in columns 3 and 6 of Table 3. To have notion on the number of changes in the typology we also report the fraction of off-diagonal entries (error rate) for the square tables obtained cross-classifying the original and after-imputation typology categories (columns 4 and 7 of Table 3).

5. Conclusions

Regarding the error detection performance of the editing process, we observe an equally good performance in not introducing new errors in data for all the variables, while the power of the set of edits to detect errors is not as good (it is quite good only for *Marital Status* and *Years married* variables). Regarding the predictive accuracy of the imputation process, among qualitative variables the better performance is for the *Marital Status* variable while, between the numeric ones, the *Years married* variable performs slightly better than *Age*. Comparing the figures for different errors levels and for different household dimensions, we observe that the higher the error level and the dimension, the less accurate are both the editing and imputation processes.

Regarding the distributional accuracy of the typology variable, we observe that the higher the error level and the dimension, the more dissimilar are the marginal distributions and the higher is the error rate.

The same set of indicators will be computed on results coming from the new system applications in order to evaluate the comparative performance of the two systems.

Moreover, additional runs will be carried out in order to investigate the effect of some system parameters on the performance. As concerns the CANCEIS system, we have changed the value of the weight parameter α in the D_{fpa} score from 0.9 to 0.5 and the value of the Number of 1st Stage Donors from 500 to 2000. In the first case the adjusted record was requested to be as similar to the failed record as to the donor record (relaxing the minimum change requirement). In

the second case the number of potential donors that are collected in the first stage of the donor searching was increased together with the chance of finding a good donor.

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, to be presented at the Statistics Canada Symposiuum 2001 "Achieving Data Quality in a Statistical Agency: a Methodological Perspective" *XVIIIth International Symposiuum on Methodological Issues.*

Bruni R. and Sassano A. (1999) CLAS: a Complete Learning Algorithm for Satisfiability. Dipartimento di Informatica e Sistemistica, Università di Roma "La Sapienza", *Technical Report 6-99*.

Chambers R. (2001) Evaluation Criteria for Statistical Editing and Imputation, *Technical report, EUREDIT project, Work Package 3*

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 1.0 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)

RESUME

ISTAT is developing a new software for the editing and imputation of hierarchical demographic data. The performance of this new software will be evaluated against the Canadian Nearest-neighbour Imputation Methodology, by means of an evaluation study based on real data from the 1991 Italian Population Census, perturbed by introducing various amounts of artificial errors and missing values. The evaluation is performed by computing, for each variable, some accuracy indicators defined inside EUREDIT project. This paper reports an overview of the new software, describes the characteristics of the evaluation study, and presents the early results concerning CANCEIS system.