WP. 38
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

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

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

# NEW PROCEDURES FOR EDITING AND IMPUTATION OF DEMOGRAPHIC VARIABLES

**Supporting Paper**

Submitted by ISTAT (Istituto Nazionale di Statistica), Italy [1]

**Abstract:** The overall editing and imputation process of the demographic variables from the 2001 Italian Population Census consists of several procedures that use different methods and/or approaches addressing specific E&I problems. This paper looks at the methodological aspects of three new procedures.

## I. INTRODUCTION

1. In handling the 2001 Population Census (PC) data ISTAT's purpose is to provide a complete and consistent set of data by *performing plausible imputations and preserving the maximum amount of collected* information. The strategy adopted to accomplish this task consists of dividing the editing and imputation (E&I) problem into simpler sub-problems and finding an appropriate solution for each of them. As a consequence, the overall E&I process has been composed of several procedures addressing specific E&I problems and implementing different E&I methods (Bianchi *et al*., 2004). The aim of this strategy is to improve the quality of final results because each problem is solved by a suitable tool. In this paper three new procedures used for handling of 2001 demographic census data are presented, each dealing with a specific E&I problem. As pointed out in the following, the three new procedures are connected to one another.

2. The first procedure was developed to face the problems occurring when connected subsets of variables are handled in sequential E&I steps. The division of the variables in subsets to be processed in sequential E&I steps is a common practice for surveys collecting a great number of variables and/or characterized by complex constraints (logical or arithmetic). It is generally known that if no constraint involves variables belonging to different subsets (unconnected subsets of variables), an optimal solution, in terms of accurate editing and imputation, could be attained regardless of the order of the processing steps. Otherwise, when some constraints involve variables belonging to different subsets (connected subsets of variables), during the processing of one step it is necessary to fix all the variables imputed by the previously processed step(s) in order to guarantee the consistency of the whole set of data. In this manner some loss of optimality could occur in the choice of the E&I solution, because not all the edit rules (edits) defined for a variable are used in the E&I step processing it. The loss of the optimality is generally related to the following problems:

[1] Prepared by Gianpiero Bianchi (gianbia@istat.it), Antonia Manzari (manzari@istat.it), Anna Pezone (pezone@istat.it), Alessandra Reale (reale@istat.it), Giorgio Saporito (saporito@istat.it)

- *poor imputation accuracy*: imputed values could be far from the true values because some relevant available information is not taken into account in the imputation phase;
- *editing failure*: the editing process can fail in the selection of the variable to impute, considering as erroneous a true value and/or considering as true an erroneous value, because some relevant available information is not taken into account in the editing phase;
- *loss of information due to improper deletion*: (this problem is a consequence of the editing failure) values of some variables handled in subsequent steps can be deleted to guarantee the consistency to the (fixed) values of variables handled in previous steps.

3. In the case of PC data, the variables concerning the persons usually resident in a dwelling have been divided in two subsets processed in sequential E&I steps. The *demographic variables* (*Year of birth*, *Sex*, *Relationship to the household reference person*, *Marital Status* and *Year of marriage*) have been handled in the first step, while all the remaining variables, named *individual variables* (*Nationality*, *Presence and dwelling*, *Degree and professional training*, *Professional status*, *Working activity*, *Place of study or work*), have been processed in the second step. The order of the processing steps has been suggested by the *relevance* and the *reliability* of the PC variables. In fact, the primary purpose of the census survey is to provide the population distribution of the family structure and the demographic variables, used to define it, are generally more reliable than the individual variables. As some demographic variables are connected with some individual variables, the above-mentioned problems could occur. In order to face them and to preserve the collected information, the E&I of the demographic variables has been performed taking into account the information provided by the individual variables, through an approach suggested by the Graph Theory (Picard, 1980).

4. The second procedure aims at locating the household reference person. One of the most important demographic variables is the *Relationship to the household reference person*. It is the basis variable for specifying all the constraints between values of variables belonging to different persons in the household (*between-person* edits) and most of the constraints between values of variables inside the person (*within-person* edits). Moreover, it is necessary to define the *family nucleus* and hence the *family typology* (target of the Population Census) that is a variable derived from all the demographic variables. For each household, the reference person needs to be located in order to allow all the remaining persons to define their relationship to it. It is common practice, in order to save processing time, technical and human resources, to locate the household reference person in a step preceding the E&I of the demographic variables. The procedure used to locate the household reference person for the 2001 Census data is based on optimization techniques and has been carried out adapting the error localization algorithm implemented in the DIESIS system (Bruni *et al.*, 2001) to the specific problem.

5. The third procedure is concerned with the treatment of invalid or inconsistent responses for the demographic variables. The demographic variables have been processed by the DIESIS system using the *data driven* and *minimum change* approaches implemented through the "*first donors then fields"* and the "*first fields then donors"* algorithms (Manzari *et al.*, 2002b). The *first donors then fields* algorithm imputes the *minimum number of variables given the available donors*. Otherwise, the *first fields then donors* algorithm imputes the (absolute) *minimum number of variables.* The two algorithms have been jointly used in order to balance the plausibility of the imputation actions with the preservation of the collected information.

6. The approach used for treating two connected subsets of variables in sequential E&I steps is described in section II. The procedure to locate the household reference person is described in section III. The joint use of the *data driven* and *minimum change* approaches is described in section IV. Finally, some concluding remarks are provided.

## II. SUBSET OF ADMISSIBLE VALUES

7. The approach used for handling the two connected subsets of variables (demographic and individual) consists of three main phases:

a) location of the variable involved in the highest number of connections among the subsets (pivot variable);
b) definition of a new auxiliary variable, the *Subset of Admissible Values (SAV)* of the pivot variable, identifying the values of the pivot variable that are as much consistent as possible with the information provided by the other variables (Manzari *et al*, 2002a);
c) performing the E&I of the pivot variable using its *SAV*.

The three phases are described in the following.

## A. Location of the pivot variable

8. According with the graph theory, a questionnaire can be represented as a connected graph, where the vertices are the variables, while the answers define the edges. When the answer to a variable is required only for some values of another variable, the first variable is named *dependent* and the second one is named *filter*. As an example, the *Marital status* is a filter variable for the *Year of marriage* variable. The filter variables are represented by vertices that give rise to more than one edge. Each of these edges enters into a subsequent vertex representing a dependent variable. Two vertices are *adjacent* if they are connected by an edge. A synthetic representation of the questionnaire can be obtained by means of subsets of variables obtained by grouping adjacent vertices, where only the last one can be a filter. These subsets can be classified in thematic *groups*. A total of 17 groups is determined by the representation of the 2001 PC questionnaire. For instance, a group *i* is composed of the set of questions whose answers are requested only to persons that go to a place of study or work. These questions are: *Time* when the person leaves the house, *Mean of transport*, *Term of the daily journey* to go to the usual place of study or work. We observe that the groups of demographic variables are connected with the groups of individual variables mainly through the *Year of birth* variable. In fact, the *Year of birth* is connected to the variables located in the 17 groups by means of consistency edits, that is, by edits checking if a combination of values in a record is plausible. Therefore, the *Year of birth* has been considered the pivot variable.

## B. Definition of the SAV of the *Year of birth* variable

9. Our aim is to perform the E&I of the *Year of birth* variable, handled in the first step together with the other demographic variables, taking into account the information provided by the groups and therefore the consistencies between the *Year of birth* and the variables in the groups. For this purpose we define the new auxiliary variable: *SAV* of *Year of birth*. The (overall) *SAV* of *Year of birth* is defined for each person *j* in the household $SAV^{(j)}$. It is obtained combining the *SAV*s of *Year of birth* defined for each group *i* for the person *j* ($SAV^{(j)}_i$).

10. The domain *D* of the *Year of birth* variable (*D*=[1888-2001]) is partitioned into 27 sub-domains $S_k$ ($k$=1, 2,…, 27). The breakpoints of each sub-domain are derived from the set of values of the *Year of birth* variable used in the within-person edits. Two persons having different values only for the *Year of birth* variable belong to the same sub-domain if and only if both of them fail the set of edits or pass the set of edits.

11. For a person *j*, a group *i* and a sub-domain $S_k$ a dummy variable $x^{(j)}_{ik}$ is defined. The $x^{(j)}_{ik}$ variable is set to 1 when the values of *Year of birth* in the sub-domain $S_k$ are consistent with the values of the variables in the group *i* collected for the person *j*. In this case, we say that the values in $S_k$ are *admissible* values for the *Year of birth* variable with respect to the collected values of the variables in the group *i*. The $x^{(j)}_{ik}$ variable is set to 0 when the values in $S_k$ are *not admissible* for the *Year of birth* with respect to the collected values of the variables in the group *i*.

12. The $SAV^{(j)}_i$ is obtained by the concatenation of the dummy variables $x^{(j)}_{ik}$ for $k$=1, 2,..., 27 that is:

$$SAV^{(j)}_i = x^{(j)}_{i1} \,||\, x^{(j)}_{i2} \,||\ldots||\, x^{(j)}_{ik} \,||\ldots||\, x^{(j)}_{i27}$$

An example of a generic $SAV^{(j)}_i$ is given in the following:

1888 2001

| $S_1$ | $S_2$ | $S_3$ | $S_4$ | $S_5$ | $S_6$ | $S_7$ | $S_8$ | $S_9$ | $S_{10}$ | $S_{11}$ | $S_{12}$ | $S_{13}$ | $S_{14}$ | $S_{15}$ | $S_{16}$ | $S_{17}$ | $S_{18}$ | $S_{19}$ | $S_{20}$ | $S_{21}$ | $S_{22}$ | $S_{23}$ | $S_{24}$ | $S_{25}$ | $S_{26}$ | $S_{27}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

13. The $SAV^{(j)}$ aims at identifying the sub-domains that are consistent with the highest number of variables in the questionnaire. Therefore, it is obtained by the intersection of the $SAV^{(j)}_i$ taking into account also the information provided by the *Year of birth*.

14. When the collected value of *Years of birth* is valid (inside the domain) it is analysed with respect to the collected values of the variables in the groups. For each group two cases can occur:

a) the collected value of *Years of birth* is consistent with the collected values of the variables in the group;
b) the collected value of *Years of birth* is inconsistent with the collected values of the variables in the group.

15. A new dummy variable $y^{(j)}_i$ is defined to indicate the consistency between the value of *Years of birth* and the values of the variables in the group $i$. In case a) the $y^{(j)}_i$ variable is set to 0 (consistency), in case b) the $y^{(j)}_i$ variable is set to 1 (inconsistency). For each person the sum of the $y^{(j)}_i$ variables is computed over all the groups.

16. If the value of the sum is less than a pre-specified threshold, then the $SAV^{(j)}_i$ is obtained by the intersection of the $SAV^{(j)}_i$ of the groups having $y^{(j)}_i = 0$. In this case the collected value of *Year of birth* is retained and the $SAV^{(j)}$ is computed using the information provided by the groups consistent with it.

17. If the value of the sum is greater or equal to the pre-specified threshold, the $SAV^{(j)}_i$ is obtained by the intersection of the $SAV^{(j)}_i$ of the groups having $y^{(j)}_i = 1$. In this case the collected value of *Year of birth* is discarded (blanking out) and the $SAV^{(j)}$ is computed using the information provided by the groups inconsistent with it.

18. When the collected value of *Year of birth* is invalid (out of the domain or missing) the $y^{(j)}_i$ variable is set to 1 for all groups and the $SAV^{(j)}$ is obtained by the intersection of the $SAV^{(j)}_i$ of all groups.

### C. E&I of the *Year of birth* variable using its *SAV*

19. The *SAV* has been used as stratum variable to identify the persons that can be used as donors when handling the demographic variables. A person in a passed edit household (recipient person) is a suitable donor for a person in a failed edit household if and only if his *Year of birth* is inside the *SAV* of the recipient person. This constraint is applied for the E&I of whatever demographic variable. In other words, a demographic variable is imputed taking the value from a person in a passed edit household, having the *Year of birth* inside the *SAV* of the person to impute.

20. The individual variables are handled in the second E&I step, therefore their values are conditioned by the values of the variables handled in the first step. The use of the SAV allows to impute a *Year of birth* consistent with the highest number of individual variables so that, the loss of information due to improper deletion is strongly reduced.

## III. LOCALIZATION OF THE HOUSEHOLD REFERENCE PERSON

21. An important phase in the imputation process of the demographic variables is the validation of the household reference person (Person 1). The Person 1 is central to the household. The variable *Relationship to the household reference person* (*relpar*) indicates the relationship of each member with the Person 1. For the Person 1 the *relpar* variable is equal to 1. All the constraints between values of

variables belonging to different persons in the household (between-person edits) and a lot of constraints between values of variables inside the person (within-person edits) use this variable.

22. In order to reduce the number of between-person edits and to save processing time and technical resources, the Person 1 is placed in the first position on the questionnaire. The validation process of the Person 1 consists of finding him inside the household and placing him in the first position.

23. A household is a set of *n* individual records $H=\{r_1,r_2,\ldots,r_n\}$. An individual record consists of a set of values, one for each variable, $r=\{v_1,v_2,\ldots,v_p\}$. The editing phase can identify three possible erroneous situations about the Person 1:

a. one person has declared to be the Person 1 but his *Year of birth* is missing or is not consistent with such a role (17 years old or younger);
b. more than one person has declared to be the Person 1;
c. no one in the household has declared to be the Person 1.

24. The procedure assigns the Person 1 role to the person which allows the minimum change of the values of the demographic variables to restore the household consistency with the edits. To attain this purpose we have used the *first donors then fields* (*FDTF*) algorithm implemented in the DIESIS system. According to this algorithm, the minimum number of variables to impute is obtained by minimizing the total cost needed to allow the adjusted household to pass all the edits. The function to minimize is the following one:

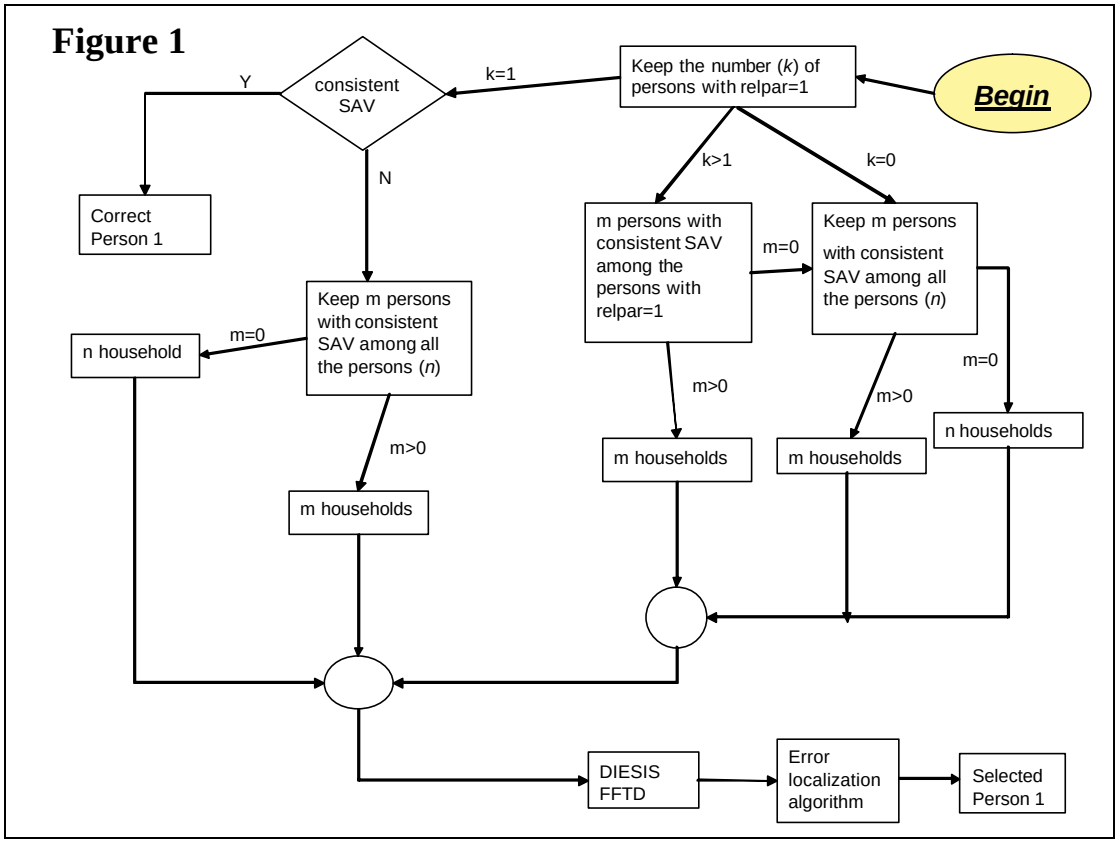
$$W(I) = \sum_{j \in I} w_j$$

where $I \subseteq H$ is the set of the demographic variables to impute and $w_j$ is the weight assigned to the change of the corresponding demographic variable.

25. We define *potential Person* 1 those persons having the *SAV* consistent with the Person 1 role, that is the ones having a *SAV* consistent with an age of 18 years or older, and select the one which minimizes the *W(I)* function. In particular:

- in cases a) and c) the potential Person 1 are all the persons in the household having a *SAV* consistent with the Person 1 role;
- in case b) the potential Person 1 are the persons that have declared to be the Person 1 if their *SAV* is consistent with the Person 1 role, otherwise the potential Person 1 are all the persons in the household having a *SAV* consistent with the Person 1 role.

The algorithm is illustrated in Figure 1:




26. The algorithm illustrated in Figure 1 works as in the following:
- In cases a) and c) *m* persons $\{r_1,r_2,\ldots,r_m\}$ with a consistent *SAV* are kept inside the household. Then *m* households $\{H_1,H_2,\ldots,H_m\}$ are created moving each potential Person 1 into the first position. More specifically $H_1$ is obtained moving the record $r_1$ into the first position, $H_2$ is obtained moving the record $r_2$ into the first position, and so on. For each of the *m* households $W(I_i)$ is calculated so to obtain the set of values $\{W(I_1),W(I_2),..,W(I_m)\}$. The Person 1 will be the person in the first position of the household $H_i$ that satisfies the condition:

$$W(I^*)=\min\{W(I_1),W(I_2),..,W(I_m)\}.$$

- In case b) the first step is to keep *m* persons having a consistent SAV and declaring themselves to be the Person 1. Then, as in case a) and c), *m* households are created and the Person 1 will be the person in the first position of the household that minimizes the function W(H). If nobody with a consistent SAV has declared himself to be the Person 1, the selection is made considering all the household members with a consistent SAV as potential Person 1.
- If no person in the household has a consistent SAV *n* households are created, one for each person, regardless of the consistent SAV requirement.

## IV. JOINT USE OF DATA DRIVEN AND MINIMUM CHANGE APPROACHES

27. The *data driven* and the *minimum change* approaches have been implemented in the DIESIS system respectively through the *first donors then fields* algorithm and the *first fields then donors* algorithm.

28. The *first donors then fields* algorithm first identifies a subset of potential donors and then determines the minimum number of variables to impute on the basis of these donors. The potential donors are the passed edit households as similar as possible to the failed edit household. The similarity between each failed edit household *e* and each passed edit household *d* is calculated by a function $f(e, d)\in[0,1]$ defined as the weighted sum of the distances (for quantitative variables) or similarities (for qualitative

variables) for each household variable over all the persons. The set of potential donors contains only the nearest *k* passed edit households (where *k* is a pre-specified value) provided that their distance is below a pre-specified threshold. The algorithm selects, from the potential donors, the minimum (weighted) set of values to impute so that the new adjusted household will pass all the edits (minimum change given the potential donors).

29. It must be stressed that by using this algorithm the imputed values for a household come from a single donor household. In some cases more than the (absolute) minimum change could be imputed. However this algorithm generally performs more plausible imputation actions than minimum change approach. In order to have a good E&I performance, this approach does require the availability of a large number of potential donors that resemble the failed edit unit. When this requirement is not met, due to low frequency of donors and/or donors too dissimilar from the failed edit household (critical situations), a poor E&I performance could occur, because a large number of unnecessary variables could be imputed.

30. The *first fields then donors* algorithm first determines the minimum number of variables to impute and then performs the imputation taking the values to impute from the set of potential donors so that the new adjusted household will pass all the edits. The potential donors are identified as previously described. This algorithm imputes the variables of one person in turn. If possible, the variables inside the person are imputed simultaneously. For each failed edit household, the algorithm first determines the minimum (weighted) number of variables to impute and identifies the potential donors. Then, for each person having some variables to impute (recipient person), the imputed values are taken from the donor person as similar as possible to the recipient one. It must be stressed that by using this algorithm the imputed values for a household could come from two or more donor households. The two algorithms have been jointly used to impute for non-response and resolve inconsistent responses for the demographic variables.

31. The *first donors then fields* algorithm has been selected as default one, with the option to turn to the *first fields then donors* algorithm when, for a given failed edit household, the number of changes proposed by the first algorithm is exceedingly high in comparison with the number of changes proposed by the second algorithm. When this is not the case the first algorithm has been preferred because it better guarantees that the combination of imputed and not imputed responses for the adjusted household is plausible and preserves the population distributions.

32. In particular, the *first donors then fields* algorithm has been mainly used to process the households having common structure, that are usually those having smaller household size. For these households it is generally possible to find enough potential donors. Otherwise, treating households having uncommon structure, usually those with largest size, few donors are generally available, and often they are not very similar to the failed edit household. In the latter case the *data driven* imputation action could require many changes to obtain an adjusted household passing the edits, therefore the *minimum change* approach has been preferred in order to preserve the collected information.

33. Given a failed edit household *H(e)*, the selection of the algorithm used for the imputation is performed in two steps. In the first step the system computes the weighted sum of changes to the failed household for obtaining the *first fields then donors* adjusted household *H(l):*

$$W(H(l)) = \sum_{j \in H} w_j x_j(e,l) \tag{1}$$

and the weighted sum of changes to the failed household for obtaining the *first donors then fields* adjusted household *H(d)*:

$$W(H(d)) = \sum_{j \in H} w_j x_j(e,d) \tag{2}$$

In 1) and 2) $w_j$ is the weight assigned to the change of the corresponding demographic variable, $x_j(e,l)$ and $x_j(e,d)$ are dummy variables assuming value 1 if the corresponding field is changed, 0 otherwise.

34. In the second step the system compares the values of the weighted sums of changes by the following function:

$$c = \frac{W(H(d))}{1+W(H(l))}$$

The *data driven* approach is selected if $c$ is less than a pre-specified threshold $\alpha_n$ depending on the household size ($n$), otherwise the *minimum change* approach is selected.

## V. CONCLUDING REMARKS

35. The three procedures outlined in the previous sections are deeply connected with one another. The SAV of *Year of birth* is the common element. It is computed in the first step of the E&I process and used for the localization of the Person 1 as well as for the handling of the demographic variables.

36. The three procedures are elements of the overall E&I process of the 2001 demographic Census data. The overall E&I process has been submitted to an accurate evaluation analysis based on the comparison of some final micro data with the corresponding raw micro data, the comparison of the final data distributions with the raw data distributions and with the distributions coming from administrative sources and on the computation of simple demographic indicators. The results obtained confirm that the combination of different procedures addressing to specific E&I problems is a good strategy to solve complex E&I problems.

37. Generally speaking, the first step of an E&I procedure should be the identification of the "illness" affecting the data. As an accurate diagnosis allows assigning the suitable treatment, so an accurate classification of the error situations can point out the associations with specific family structures and, consequently, suggest the appropriate solution. Handling PC data we have often observed such kind of associations and, in order to select a solution that takes into account the family structure, we have preferred the data driven approach to the minimum change approach. The data driven approach, in fact, selects the imputation action from donor households most resembling the failed household, that is, considering the family structure of the failed household.

38. The overall strategy aims at performing plausible imputations preserving the maximum amount of collected information. The achievement of this objective does not necessarily imply the use of the minimum change approach. Other authors have observed that "it is not always appropriate to impute the minimum number of variables…. this is particularly evident in the case of systematic errors…" and "where plausibility is preferred over minimum change"(Bankier *et al*, 2002). Our experience let us agree with these statements. Moreover, we observe that self-completion questionnaires, as the PC ones, are mostly affected by a high partial non-response rate. When this occurs the minimum change could cause a loss of collected information due to the blanking out of the values of some variables instead of the imputation of a higher number of variables. To avoid falling into this trap we have chosen in some cases not to consider as a "change" the imputation of a missing value. This allowed preserving the maximum amount of collected information.

**References**

Bankier M., Mason P. and Poirier P. (2002) Imputation of demographic variables from the 2001 Canadian Census of Population, *Proceedings of the UN/ECE Work Session on Statistical Data Editing*, Finland (Helsinki).

Bianchi G., Pezone A., Reale A., Saporito G. (2004) Metodi e Procedure per il Controllo e la Correzione delle Variabili Demografiche Familiari del Censimento della Popolazione 2001, *Internal document* (*in italian),* ISTAT

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

Manzari A., Pezone A., Reale A. (2002a) Evaluation of a new approach for edit and imputation of social and demographical data with hierarchical structure, *Atti della XLI Riunione Scientifica SIS*, Milano, 5-7 Giugno 2002, Sessioni spontanee, pp 689-692.

Manzari A. and Reale A. (2002b) Towards a new system for edit and imputation of the 2001 Italian Population Census data: A comparison with the Canadian Nearest-neighbour Imputation Methodology, In *Proc. 53rd Session of The International Statistical Institute, August 22-29, 2001,* pp. 634-655. Sydney: International Statistical Institute.

Picard C. F. (1980) *Graphs and questionnaires*, North-Holland, Netherlands.

---